DISS. ETH NO. 28191

Deep Learning of Entity-Guided Representations in Digital Pathology

A dissertation submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zürich)

presented by

Pushpak Pati
Master of Science
in Electrical Engineering and Information Technology
ETH, Zürich, Switzerland
born on 17 July 1991
citizen of India

accepted on the recommendation of

Prof. Dr. Orçun Göksel (ETH Zürich), examiner
Prof. Dr. Anne Martel (University of Toronto), co-examiner
Prof. Dr. Geert Litjens (Radboud University), co-examiner
Dr. Maria Gabrani (IBM Research), co-examiner

2022

Dedicated to my loving family.

Abstract

Pathological examination is the gold standard for cancer diagnosis, prognosis, and therapeutic response predictions. Advancements in scanning technologies and an increased focus on precision medicine have paved the way for developing digital-pathology-based assessments. Digital pathology has enabled the digitization of microscopy slides into high-resolution whole-slide images and opened up opportunities for *computational pathology* (CP). CP aspires to alleviate the cumbersome and time-consuming routine workflow of pathologists by introducing computer-aided assistive tools. To this end, CP leverages computational techniques for automated exploration and extraction of meaningful information from histopathology images. The demand for CP has recently gained even more attention due to the growing incidence rate of diagnostic cases per year.

The basis of a typical CP system is artificial intelligence, in particular, *deep learning* (DL) due to its recent large-scale success. Capability of DL to automatically extract and utilize informative representations from complex histopathology images in a data-driven manner have popularized its adoption in CP. Several DL methods have been developed to address various histopathology tasks, such as nuclei detection and characterization, tumor delineation, tissue grading and staging, and survival estimation. However, the clinical adoption of DL methods is inhibited by several challenges, including: (1) infeasibility of acquiring large high-quality annotated histopathology datasets for training models; (2) requirement of prohibitive computational resources for processing large whole-slide images; and (3) a lack of transparency and interpretability of DL decisions. Further, most DL models in CP are built based on convolutional neural networks (CNNs), which treat an image as a composition of multisets of pixels, to perform analyses in a pixel-paradigm. However, operating in pixel-paradigm induces several crucial bottlenecks, such as: (i) not being able to easily utilize tissue composition and well-established prior pathological knowledge, due to a disregard for histological entities, e. g., nuclei, cells, glands; (ii) an inability to simultaneously capture both local cell microenvironment and global tissue microenvironment; (iii) intensive computational requirements for operating on large whole-slide images; and (iv) non-straight-forward model interpretations due to the trained models not making diagnostic decision explicitly based on well-defined histological entities.

This thesis aims to address the aforementioned challenges and limitations concerning DL methods in CP. The motivation herein is that the analysis of tissues should rely on the phenotype and topological distribution of their constituting histological entities. Therefore, the analytical paradigm is proposed to be shifted from conventional pixels to *entities*. A histopathology image is first transformed into an entity-guided representation, specifically an *entity-graph*. The nodes and edges of the graph denote comprehensible histological entities and entity-to-entity interactions, respectively. The local entity-level phenotypical properties are embedded in the nodes and the global tissue-microenvironment is captured by the graph topology. Subsequently, the advancements of DL techniques on graph-structured data, in particular Graph Neural Networks (GNNs), are leveraged to efficiently construct a relation-aware entity-graph-representation for addressing downstream histopathology tasks. Operating in the entity-paradigm enables the incorporation of task-relevant entity-level prior knowledge for comprehensive tissue modeling. Entity-graphs being more flexible and memory efficient, compared to pixel-based counterparts, can scale to images of arbitrary shapes and sizes. Further, interpreting an entity-graph-based model can highlight salient entities and interactions for model decisions, which the pathologists can directly comprehend.

Relevance and superiority of learning on entity-guided tissue representations are established for a variety of histopathology tasks across several tissue types. The proposed entity-graphs encode different entity types, i. e., nuclei, tissue regions, and both; and include different graph topologies, i. e., uni-level, multi-level, hierarchical. Further, various entity-guided GNNs are proposed herein to tackle the challenges of: (1) learning from weak supervision and limited annotations; (2) processing histopathology images of arbitrary sizes; and (3) interpretability and explainability of model decisions in pathologist-friendly terminologies. Specifically, the proposed methodologies are applied for the following histopathology tasks: (a) supervised subtyping breast carcinoma tumor regions, (b) weakly-supervised simultaneous classification and semantic segmentation of prostate cancer needle biopsies, and (c) generating qualitative and quantitative interpretations of breast subtyping model decisions. The proposed methods achieve state-of-the-art performance for these tasks, and have been validated by domain-expert pathologists. The generalization ability of the proposed methods is also substantiated by classifying and segmenting prostate cancer biopsies from multiple data sources. In addition, a flexible open-source python library, HistoCartography, has been developed to facilitate effective graph analytics in digital histopathology.

Zusammenfassung

Pathologische Untersuchungen sind der Goldstandard für die Krebsdiagnostik, Prognostik sowie die Vorhersage von Behandlungserfolgen. Dank Fortschritten in Scanning-Technologien und einem verstärkten Fokus auf Präzisionsmedizin ist es möglich geworden, Untersuchungsverfahren basierend auf digitaler Pathologie zu entwickeln. Die digitale Pathologie hat die Digitalisierung von Mikroskopie-Objektträgern in hochauflösende, sogenannte Whole Slide Images ermöglicht und Chancen für die computergestützte Pathologie (CP) eröffnet. Die CP zielt darauf ab, mühsame und zeitaufwändige Arbeitsabläufe von Pathologen durch die Einführung computergestützter Hilfsmittel zu erleichtern. Zu diesem Zweck nutzt die CP computergestützte Techniken zur automatisierten Suche und Extraktion von aussagekräftigen Informationen aus histopathologischen Bildern. Aufgrund der steigenden Zahl von jährlichen diagnostischen Fällen hat der Bedarf an CP in jüngster Zeit noch weiter zugenommen.

Typische CP-Systeme stützen sich im Kern auf künstliche Intelligenz, insbesondere Deep Learning (DL), welches in letzter Zeit weitreichende Erfolge ermöglicht hat. Durch die Fähigkeit automatisch und datengesteuert informative Repräsentationen aus komplexen histopathologischen Bildern zu extrahieren und zu nutzen, hat die Anwendung von DL in der CP zunehmend an Popularität gewonnen. Zahlreiche DL-Methoden wurden bereits für verschiedene histopathologische Aufgaben entwickelt, z.B. zur Erkennung und Charakterisierung von Nuklei, Abgrenzung von Tumoren, Einstufung und Einteilung von Gewebe sowie zur Einschätzung der Überlebenswahrscheinlichkeit. Die klinische Anwendung von DL-Methoden wird jedoch durch mehrere Probleme erschwert, darunter: (1) die Tatsache, dass es nicht möglich ist, große, qualitativ hochwertige annotierte Histopathologiedatensätze für das Training von Methoden zu erwerben; (2) der hohe Bedarf an Rechenressourcen für die Verarbeitung großer Whole Slide Images; und (3) die mangelnde Transparenz und Interpretierbarkeit von DL-Entscheidungen. Darüber hinaus basieren die meisten DL-Modelle in der CP auf Convolutional Neural Networks (CNNs), die ein Bild als eine Zusammensetzung mehrerer Pixel betrachten und Analysen im Pixel-Paradigma durchführen. Das Arbeiten im Pixel-Paradigma führt jedoch zu mehreren entscheidenden Engpässen, wie z.B. (1) die fehlende Verwertung von bekanntem pathologischen Vorwissen und Informationen über die Gewebezusammensetzung, da histologische Instanzen, d.h., Nuklei, Zellen, Drüsen usw. außer Acht gelassen werden; (2) die mangelnde Fähigkeit, sowohl die lokale Zell-Mikro-Umgebung als auch die globale Gewebe-Mikro-Umgebung gleichzeitig zu erfassen; (3) intensive Rechenanforderungen für die Vearbeitung großer Whole Slide Images; und (4) Schwierigkeiten beim Verständnis von Modellinterpretationen aufgrund der unterschiedlichen Diagnoseverfahren von Pathologen und Modellen.

Diese Arbeit hat zum Ziel, die oben genannten Herausforderungen und Einschränkungen in Bezug auf DL-Methoden in der CP anzugehen. Das grundlegende Konzept besteht darin, das Vorwissen zu nutzen, dass die Analyse eines Gewebes auf dem Phänotyp und der topologischen Verteilung der einzelnen histologischen Instanzen beruht. Daher schlägt diese Arbeit vor, das analytische Paradigma von konventionellen Pixeln auf Instanzen zu verlagern. Zunächst wird ein histopathologisches Bild in eine instanzbasierte Repräsentation transformiert, d.h. in einen Instanzgraphen. Die Knoten bzw. Kanten des Graphen bilden nachvollziehbare histologische Instanzen bzw. Instanz-zu-Instanz-Interaktionen ab. Die lokalen phänotypischen Eigenschaften auf der Instanzebene sind in den Knoten eingebettet, während die globale Gewebe-Mikro-Umgebung durch die Graphentopologie erfasst wird. Anschließend nutzt die Arbeit die Fortschritte von DL-Techniken auf graphenbasierten Daten, insbesondere Graph Neural Networks (GNNs), zur effizienten Konstruktion einer relationsbewussten Instanzsgraphen-Darstellung, welche für anschliessende histopathologische Aufgaben genutzt werden kann. Das Arbeiten im Instanz-Paradigma ermöglicht die Berücksichtigung von aufgabenrelevantem Vorwissen auf Instanzebene für eine umfassende Gewebemodellierung. Die Instanzgraphen sind flexibler und speichereffizienter als deren pixelbasierte Gegenstücke und können auf Bilder beliebiger Form und Größe skaliert werden. Darüber hinaus können bei der Interpretation eines auf Instanzgraphen basierenden Modells Instanzen und Interaktionen hervorgehoben werden, die wichtig für die Modellentscheidung sind. Pathologen können einen direkten Bezug zu diesen Instanzen und Interaktionen herstellen und entsprechende Schlussfolgerungen ziehen.

Die Relevanz und Überlegenheit des Lernens auf der Grundlage von instanzbasierten Gewebedarstellungen wird für eine Vielzahl von Histopathologieaufgaben in verschiedenen Gewebetypen nachgewiesen. Die vorgeschlagenen Instanzsgraphen kodieren verschiedene Instanzstypen, d.h. Nuklei und/oder Geweberegionen, und umfassen verschiedene Graphentopologien, d.h. einstufig, mehrstufig und hierarchisch. Darüber hinaus werden verschiedene instanzbasierte GNNs vorgeschlagen, um die folgenden Herausforderungen zu bewältigen: (1) das Lernen aus schwacher Überwachung und begrenzten Annotationen; (2) die Verarbeitung von Histopathologiebildern beliebiger Größe; und (3) die Interpretierbarkeit und Erklärbarkeit von Modellentscheidungen in pathologisch verständlicher Terminologie. Im Einzelnen werden die vorgeschlagenen Methoden angewandt für (1) die überwachte Subtypisierung von Brustkrebs-Tumorregionen, (2) die schwach überwachte gleichzeitige Klassifizierung und semantische Segmentierung von Prostatakrebs-Nadelbiopsien und (3) die Generierung qualitativer und quantitativer Interpretationen von Modellentscheidungen in der Brust-Subtypisierung. Die vorgeschlagenen Methoden erreichen für die in Betracht gezogenen histopathologischen Aufgaben eine Leistung auf dem neuesten Stand der Technik und wurden von spezialisierten Pathologen mit fundierten Fachkenntnissen validiert. Die bessere Generalisierbarkeit der vorgeschlagenen Methoden wird auch für die Klassifizierung und Segmentierung von Prostatakrebs-Biopsien aus mehreren Datenquellen nachgewiesen. Darüber hinaus wird eine generische Open-Source-Python-Bibliothek, HistoCartography, entwickelt, um eine effektive Graph-Analyse in der digitalen Histopathologie zu ermöglichen.

Acknowledgements

The Ph.D. journey has been one of the most fabulous chapters in my life and filled with great experiences. During this exciting expedition, I have met many amazing colleagues and friends, learned from them, grown up a lot, and built countless memories together. Here, I would like to take the opportunity to express my gratitude to some of these great people. This thesis would not have been possible without your sincere efforts of guidance, assistance, support, and encouragement.

First and foremost, I want to express my deepest gratitude to my advisor Orçun Göksel. He has always inspired me through his immense enthusiasm for research and has provided relentless support in the past four years. I want to thank Orçun for giving me freedom to pursue various research ideas, providing me detailed guidance along the way, while at the same time ensuring that I stay focused.

Next, my deepest gratitude goes to my co-advisor, Maria Gabrani. She made my journey possible by her belief in my research potential, her patience, and her scientific and moral support throughout my time at IBM Research Zurich. She was not only a supervisor, but also a mentor and a friend, whose door was always open to me.

I am indebted to the members of my dissertation committee—Orçun Göksel, Anne Martel, Geert Litjens, and Maria Gabrani—for providing valuable feedback and discussions during the exam.

I am thankful to my collaborators and co-authors at IBM Research, and beyond – each one of you contributed uniquely towards this dissertation. Thanks to my colleagues and mentors at IBM Research—Antonio Foncubierta Rodríguez, Aditya Kashyap, Anna Fomitcheva Khartchenko, Govind V. Kaigala, and Marianna Rapsomaniki—for long-lasting collaborations and research guidance. I am also thankful for the opportunity to work with amazing Master students.

A special thanks to my colleague—Guillaume Jaume—for being a great research partner. I couldn't have wished for better. I deeply cherish your positive characters and our endless scientific discussions, co-developments, banters, and pool matches.

I have been lucky to be surrounded by wonderful friends who have lifted my spirit countless times during this chapter in my life. I would like to specifically thank—Kevin Thandiackal, Sonali Andani, Anca-Nicoleta Ciubotaru, Alexandra Kim, An-phi Nguyen, and Alex Sobczyk—my friends from IBM Research. Krishna Chaitanya, Prabhakaran Santhanam, Sai Rama Rao Aitharaju, Bhaskara Rao Chintada, Nirav Karani, Prateek Purwar, Arun Balajee Vasudevan, Vaishakh Patil, and Samarth Shukla—my friends from ETH, Ashutosh Pattnaik, Srinit Das, Poojitha Alluri, Hrudayjyoti Biswal, and Apoorva Gupta—my friends from old times, and Varaha Karthik, Vinayak Chaitanya, Peshal Agarwal, Asil Örgen, Akash Joshi, Shyam Sundhar, and Sunil Kumar—my flatmates. I thank you all for nice chats, food sessions, and all the great memories. I am heartily

grateful to Jaydev Mishra and Sunita Soumyadarshana for bestowing me the warmth of family in the foreign land.

Along this journey, I have met many incredible mentors, who have constantly guided and motivated me—Dwarikanath Mahapatra, Behzad Bozorgtabar, Nuri Murat Arar, Pankaj Gupta, and Umesh Chandra Pati.

Finally, my thanks go to my family, the real heroes in my life—Ramanath Pati, Maitreyee Pati, Subhakumari Pati, Mamata Pati, Jagruti Patel, Om Prakash Patel, Diptirani Choudhury, Azad Patel—for their sacrifices, unconditional love, endless support, and inspiration. I specially want to thank my wife Jagruti for constantly believing in me, and instilling my life with purpose and motivation. I am extremely fortunate to have her by my side.

Contents

1 Introduction **1**
- 1.1 Histopathology diagnostic workflow . . . 1
  - 1.1.1 Conventional histopathological diagnosis . . . 1
  - 1.1.2 Digital pathology and computer-aided diagnosis . . . 2
- 1.2 Motivation . . . 2
- 1.3 Thesis Goals . . . 7
- 1.4 Thesis Outline and Contributions . . . 8

2 Background **11**
- 2.1 A Review of Deep Learning in Computational Pathology . . . 11
  - 2.1.1 Deep learning methods based on training supervision . . . 11
  - 2.1.2 Learning from limited labeled data . . . 14
  - 2.1.3 Processing large histopathology images . . . 16
  - 2.1.4 Interpretability of deep learning methods . . . 17
  - 2.1.5 Deep learning of entity-guided representations . . . 18
- 2.2 Technical Preliminaries on Graph Representation Learning . . . 19
  - 2.2.1 Graphs: definitions and notations . . . 20
  - 2.2.2 Machine learning on graphs . . . 20
  - 2.2.3 Graph Neural Networks . . . 22
  - 2.2.4 Interpretability of Graph Neural Networks . . . 26

3 High-Quality Immunohistochemical Stains through Computational Assay Parameter Optimization **33**
- 3.1 Introduction . . . 33
- 3.2 Methodology . . . 37
  - 3.2.1 Diagnostically relevant signal segmentation . . . 37
  - 3.2.2 Staining quality metric learning . . . 38
  - 3.2.3 Image and quality representation . . . 40
  - 3.2.4 Sensitivity to process parameters . . . 42
- 3.3 Materials . . . 42
- 3.4 Results . . . 43
  - 3.4.1 Extraction of first quality indicator . . . 43
  - 3.4.2 Extraction of second quality indicator . . . 46
  - 3.4.3 Staining quality assessment . . . 46
  - 3.4.4 Sensitivity to process parameters . . . 47
  - 3.4.5 Comparison with a clinical staining protocol . . . 48
- 3.5 Conclusion . . . 49

4 Hierarchical Graph Representations in Digital Pathology 53
4.1 Introduction . . . 53
4.2 Related work . . . 55
4.3 Methodology . . . 57
  4.3.1 Notations . . . 57
  4.3.2 Pre-processing . . . 57
  4.3.3 Graph representation . . . 58
  4.3.4 Graph learning . . . 60
4.4 Datasets . . . 62
4.5 Results . . . 64
  4.5.1 Baselines . . . 64
  4.5.2 Implementation . . . 66
  4.5.3 Ablation studies . . . 66
  4.5.4 Classification results on BRACS dataset . . . 68
  4.5.5 Classification results on BACH dataset . . . 73
  4.5.6 Qualitative analysis . . . 74
4.6 Conclusion . . . 76

5 Towards Explainable Graph Representations in Digital Pathology 77
5.1 Introduction . . . 77
5.2 Methodology . . . 78
  5.2.1 Cell-graph representation and learning . . . 78
  5.2.2 Cell-graph explainer . . . 79
5.3 Dataset . . . 81
5.4 Results . . . 81
  5.4.1 Implementation . . . 81
  5.4.2 Quantitative and qualitative analyses . . . 81
5.5 Conclusion . . . 83

6 Quantifying Explainers of Graph Neural Networks in Digital Pathology 85
6.1 Introduction . . . 85
6.2 Related work . . . 87
6.3 Methodology . . . 89
  6.3.1 Entity-graph notations . . . 89
  6.3.2 Entity-graph construction . . . 89
  6.3.3 Entity-graph learning . . . 89
  6.3.4 Post-hoc graph explainer . . . 91
  6.3.5 Quantitative metrics for graph explainability . . . 92
6.4 Dataset . . . 93
6.5 Results . . . 94
  6.5.1 Implementation . . . 94
  6.5.2 Qualitative assessment . . . 94
  6.5.3 Quantitative assessment . . . 95
6.6 Conclusion . . . 97
6.7 Appendices . . . 97
  6.7.1 BRACS dataset . . . 97

- 6.7.2 Concepts and Attributes . . . 97
- 6.7.3 Quantitative assessment . . . 98
- 6.7.4 Qualitative assessment . . . 100

7 Reducing Annotation Effort in Digital Pathology 105

- 7.1 Introduction . . . 105
- 7.2 Related work . . . 107
- 7.3 Background . . . 110
  - 7.3.1 Preliminaries . . . 110
  - 7.3.2 Categorical cross-entropy . . . 111
  - 7.3.3 Deep metric learning . . . 111
- 7.4 Methodology . . . 112
  - 7.4.1 Co-representation learning framework for classification . . . 112
  - 7.4.2 Improving deep metric learning . . . 113
  - 7.4.3 Co-representation learning training strategy . . . 118
- 7.5 Results . . . 118
  - 7.5.1 Implementation . . . 118
  - 7.5.2 Ablation framework . . . 119
  - 7.5.3 CRCHistoPhenotypes . . . 120
  - 7.5.4 AMIDA13 . . . 124
  - 7.5.5 Classification with CoReL . . . 126
  - 7.5.6 ICPR12 . . . 126
  - 7.5.7 CoNSeP . . . 127
  - 7.5.8 Kather Multiclass . . . 127
  - 7.5.9 Qualitative analysis . . . 128
  - 7.5.10 Runtime analysis . . . 129
- 7.6 Discussion . . . 129
- 7.7 Conclusion . . . 131
- 7.8 Appendices . . . 133

8 Weakly Supervised Segmentation and Classification of Prostate Cancer Slides 137

- 8.1 Introduction . . . 138
- 8.2 Related work . . . 140
  - 8.2.1 Weakly-supervised histopathology image classification . . . 140
  - 8.2.2 Weakly-Supervised histopathology image segmentation . . . 140
  - 8.2.3 Domain shift, generalization, and uncertainty in computational pathology . . . 141
- 8.3 Methodology . . . 142
  - 8.3.1 Notation and preliminaries . . . 142
  - 8.3.2 Preprocessing and tissue-graph construction . . . 144
  - 8.3.3 Contextualized node embeddings . . . 145
  - 8.3.4 WSI classification . . . 145
  - 8.3.5 Weakly supervised semantic segmentation . . . 146
  - 8.3.6 Extension to Bayesian models . . . 148
- 8.4 Datasets . . . 149

8.5 Results . . . 150
  8.5.1 Implementation and metrics . . . 150
  8.5.2 Baselines . . . 152
  8.5.3 WSS performance analysis . . . 153
  8.5.4 Generalization: performance, uncertainty, and calibration . . . 154
  8.5.5 Qualitative analysis . . . 157
8.6 Conclusion . . . 159

9 HistoCartography: A Toolkit for Graph Analytics in Digital Pathology 165
9.1 Introduction . . . 165
9.2 Related work . . . 166
  9.2.1 Graphs in computational pathology . . . 166
  9.2.2 Extant libraries in computational pathology . . . 167
9.3 HISTOCARTOGRAPHY: Graph analytics tool for pathology . . . 167
  9.3.1 Preprocessing module . . . 167
  9.3.2 Graph machine learning module . . . 169
  9.3.3 Explainability module . . . 170
9.4 Benchmarking HISTOCARTOGRAPHY . . . 170
  9.4.1 Computational time . . . 170
  9.4.2 Performance benchmark . . . 172
9.5 Conclusion . . . 173
9.6 Appendices . . . 173
  9.6.1 HISTOCARTOGRAPHY ecosystem . . . 173
  9.6.2 HISTOCARTOGRAPHY syntax . . . 173
  9.6.3 Supported handcrafted features . . . 173
  9.6.4 Future of HISTOCARTOGRAPHY . . . 174

10 Conclusions 177
10.1 Summary and limitations . . . 177
10.2 Future work . . . 179

List of Figures

1.1 Overview of non-digital, digital, and computer-aided diagnostic workflows . 3
1.2 H&E stained breast tumor regions and respective cell-graphs to illustrate the distribution of epithelial nuclei 4
1.3 H&E stained breast tumor regions and respective cell-graphs to illustrate the distributions of individual nuclei 5
1.4 Explanation heatmaps produced by post-hoc feature attribution techniques applied to a CNN 5
1.5 Field-of-view of fix-sized concentric patches extracted from a tissue region at different resolutions. 6
1.6 Illustration of pixel- and entity-based analytical workflows in computational pathology. 8

2.1 An overview of major deep graph learning tasks 21
2.2 Topological distribution of words, pixels, and objects, in text-, image-, and network-graphs, respectively 22
2.3 An overview of a Message Passing Neural Network 23
2.4 An overview of GIN and PNA layers. 26

3.1 Immunostaining process and variability in staining quality due to process parameter variations in IHC 35
3.2 Overview of immuno-staining quality assessment methodology 37
3.3 Diagnostically relevant signal segmentation in an IHC image 38
3.4 Variability in immunostaining quality expectations with HER2 antibody for different disease types 39
3.5 Overview of the disease-specific staining quality metric (SQM) learning . . 40
3.6 Misclassified samples by the disease-type classifier 46
3.7 ROC curves and AUC scores for individual SQMs and their aggregation. . 47
3.8 Disease-type specific 95% confidence ellipses fitted to quality values in the process parameter space 47
3.9 Manifold fitting and sensitivity analysis of quality values acquired from disease-type specific SQMs. 48
3.10 Staining qualities of an IHC stained PT+ TMA over parametric configurations 49
3.11 SKBR3 cell blocks stained with clinical protocol and optimized staining protocol. 50
3.12 IHC images of PT+ tissues stained with the proposed optimized staining protocol. 50

4.1 Overview of the hierarchical entity-graph based tissue analysis methodology 57
4.2 Overview of HACT graph construction for a TRoI 58

4.3 Overview of the HACT-Net architecture 60
4.4 Samples of class-wise tumor regions-of-interest in BRACS dataset. 62
4.5 Overview of the variability for DCIS category in BRACS dataset. 63
4.6 Mean and standard deviation of per-class precision and recall for 7-class classification.. 70
4.7 Mean and standard deviation of row-normalized 7-class confusion matrix. 71
4.8 Binary decision tree used by pathologists for breast cancer diagnosis.. . . . 71
4.9 Qualitative comparison of CG-GNN, TG-GNN, and HACT-Net for 7-class classification 74
4.10 Feature attribution maps of HACT-Net on TG and CG for 7-class classification75
4.11 Qualitative performance of CG-GNN, TG-GNN, and HACT-Net on an image with tissue-tear and blur artifacts 75

5.1 A TRoI is transformed into a CG, and is processed by a CG-GNN to predict the cancer subtype. 78
5.2 Overview of CGExplainer. The original input CG is iteratively pruned until convergence of the optimization. 80
5.3 Qualitative comparison of original CG and CGExplainer generated explanations for 2-, 3-, and 5-class scenarios on a DCIS TRoI.. 83
5.4 Nuclei types annotation. Overlaid segmentation masks of nuclei from 5-class explanation in green. 84

6.1 Sample explanations produced by pixel- and entity-based explainability techniques for a ductal carcinoma *in situ* RoI.. 86
6.2 Overview of the quantitative framework for evaluating post-hoc graph explainability techniques 88
6.3 Detailed view of the quantitative assessment framework for graph explainers90
6.4 Qualitative results from different graph explainability techniques 95
6.5 Per-class histograms for different *concepts* across different graph explainers100
6.6 Visualizing the variation of pair-wise class separability score with respect to several nuclei importance thresholds 101
6.7 Qualitative results-I for the graph explainers on different breast cancer subtypes 102
6.8 Qualitative results-II for the graph explainers on different breast cancer subtypes 103

7.1 Overview of the co-representation learning, CoReL, for classification framework 110
7.2 Illustration of multiple similarities for a negative pair considering the spatial distribution of other samples in the embedding spaces 113
7.3 Illustrating the contribution of neighborhood awareness for sampling informative positive and negative pairs under neighborhood-aware multiple similarity sampling blanket 115
7.4 Classification and Davies-Bouldin scores on CRCHistoPhenoTypes and AMIDA13 datasets for different DML objectives and triplet sampling strategies 119
7.5 classification performances of CoReL with different DML objectives and triplet sampling strategies on six histopathology datasets 122

7.6 Qualitative assessment of the impact of training dataset size on CoReL classification performance . . . 128
7.7 Effect of loss trade-off parameter α on CoReL classification performance for different training dataset sizes . . . 131

8.1 Overview of the WholeSIGHT method for simultaneous WSI classification and segmentation . . . 143
8.2 Class distribution of the Karolinska, Radboud, and Sicap datasets. . . . 150
8.3 Confusion matrix of Gleason grade classification for the WholeSIGHT-DE method on the Karolinska, Radboud, and Sicap datasets . . . 156
8.4 Segmentation performances and uncertainty analysis of WholeSIGHT, WholeSIGHT-MCD, and WholeSIGHT-DE on Sicap dataset . . . 157
8.5 Uncertainty analysis in classification for WholeSIGHT, WholeSIGHT-MCD, and WholeSIGHT-DE models . . . 160
8.6 Reliability diagrams of WholeSIGHT, WholeSIGHT-MCD, and WholeSIGHT-DE tested on Karolinska and Radboud datasets for the primary and secondary Gleason classification heads . . . 161
8.7 Samples of ground truth and predicted segmentation maps using WholeSIGHT and WholeSIGHT(Multiplex, NC) on Sicap dataset . . . 162
8.8 t-SNE visualization of node-level feature representations from WholeSIGHT, and example patches corresponding to several regions on the two-dimensional t-SNE feature space for Sicap dataset . . . 163
8.9 Gleason grade, ISUP grade, primary-, and secondary-classification confusion matrices obtained for WholeSIGHT-DE on Karolinska, Radboud, and Sicap datasets . . . 164

9.1 Overview of HistoCartography modules and functionalities . . . 169
9.2 Qualitative explanations generated by the interpretability module on sample breast RoIs . . . 172
9.3 Implementation of Vahadane stain normalization and tissue mask detection with the *Preprocessing* functionalities in the HistoCartography API . . . 174
9.4 Implementation of cell-graph and tissue-graph generation using the graph builders in HistoCartography . . . 175
9.5 Implementation of the cell- and tissue- graph models by using the ML modules in HistoCartography . . . 175
9.6 Implementation of several post-hoc graph explainers in HistoCartography 176

List of Tables

3.1 Disease type confusion matrix for the best classifier trained with hand-crafted features.. 44
3.2 CNN architecture and network hyperparameters for disease type classification. 45
3.3 Machine learning and deep learning-based disease type classification results 45

4.1 Key statistics of the BRACS dataset 64
4.2 Ablation studies: Impact of node features in entity-graphs on breast cancer classification 67
4.3 Ablation: Impact of jumping knowledge. Mean and standard deviation of 7-class weighted F1-scores. Results expressed in %. 67
4.4 Ablation studies: Impact of GNN layers on breast cancer classification . . . 68
4.5 Quantitative results for fine-grained breast cancer TRoI classification 69
4.6 Quantitative results for breast cancer TRoI classification into different risk categories 71
4.7 Quantitative results for breast cancer TRoI classification by following binary diagnostic tree 72
4.8 Benchmarking of HACT-Net against pathologists for breast cancer subtyping on BRACS dataset 73
4.9 Concordance among three independent pathologists on BRACS test dataset 73
4.10 Classification results on BACH dataset. 74

5.1 Quantitative results for CG-GNN, CGExplainer compactness, CGExplainer and RGExplainer performances. ↑ and ↓ indicate higher and lower values are better, respectively. 82

6.1 Quantitative assessment of graph explainers: GnnExplainer, GraphGrad-CAM, GraphGrad-CAM++ and GraphLRP, using proposed *maximum*, *average*, and *correlated separability* metrics 96
6.2 Quantification of *concepts* for pair-wise and aggregated class separability in GnnExplainer 96
6.3 Statistics of BRACS dataset for quantifying graph explainers 98
6.4 Pathologically-understandable nuclear *concepts*, corresponding measurable *attributes*, and their expected behavior for three breast cancer subtypes . . . 99

7.1 Effect of embedding size on classification performance using various subsets of training data on CRCHistoPhenotypes and AMIDA13.. 121
7.2 Effect of base batch size on classification performance using various subsets of training data on CRCHistoPhenotypes and AMIDA13.. 121

7.3 Results on CRCHistoPhenotypes dataset for CoReL with several DML strategies 123
7.4 Results on AMIDA13 dataset for CoReL with several DML strategies 123
7.5 Results on ICPR12 dataset for CoReL with several DML strategies 124
7.6 Results on CoNSeP dataset for CoReL with several DML strategies 124
7.7 Results on Kather Multiclass datasets for CoReL with several DML strategies 126
7.8 Training time analysis of CoReL for different training dataset sizes on CRCHistoPhenoTypes and AMIDA13 130
7.9 Ablation studies: Impact of triplet sampling strategies on classification for CRCHistoPhenotypes and AMIDA13. 133
7.10 Ablation studies: Impact of DML losses on classification for CRCHistoPhenotypes and AMIDA13 133
7.11 CRCHistoPhenotypes dataset: Mean and standard deviation of classification for various learning strategies 133
7.12 AMIDA13 dataset: Mean and standard deviation of classification for various learning strategies 134
7.13 ICPR12 dataset: Mean and standard deviation of classification for various learning strategies 134
7.14 Kather Multiclass Internal datasets: Mean and standard deviation of classification for various learning strategies 134
7.15 Kather Multiclass External datasets: Mean and standard deviation of classification for various learning strategies 134
7.16 Impact of α on CoReL classification for different training dataset sizes, for CRCHistoPhenotypes and CoNSeP 135
7.17 Impact of α on CoReL classification for different training dataset sizes, for AMIDA13 and ICPR12 135
7.18 Impact of α on CoReL classification for different training dataset sizes, for Kather datasets 135

8.1 Classification and segmentation results of WholeSIGHT on Sicap dataset . 154
8.2 Classification and segmentation results of WholeSIGHT on Radboud dataset 154
8.3 Classification and segmentation results of WholeSIGHT on Karolinska dataset 155
8.4 Classification and segmentation results of WholeSIGHT on Radboud, Karolinska, and Sicap for models trained using both Radboud and Karolinska datasets 155

9.1 Overview of HistoCartography functionalities, with the i/o, CPU and GPU compatibility, and availability in extant libraries for individual module. . . 168
9.2 Run time analysis of HistoCartography core functionalities 171
9.3 Benchmarking HistoCartography for classification and segmentation tasks 172

1

Introduction

Histology is a branch of biology concerned with the composition and structure of tissues in relation to the specialized tissue functionalities. The fundamental aim of histology is to determine how tissues are organized at all structural levels, from cells and inter-cellular substances to organs. Histopathology is the study of tissues affected by diseases, such as cancer. A disease affects the tissue in a distinctive way depending on the type of the tissue as well as the progressive extent of the disease. Histopathology deals with diagnosing such diseased tissues to identify the disease, determine their severity, and thereby select pertinent treatment procedures. The following sections present an overview of a standard diagnostic workflow, and its transformation from conventional non-digital examination to modern computer-aided digital evaluation.

1.1 Histopathology diagnostic workflow

1.1.1 Conventional histopathological diagnosis

A tissue specimen (or biopsy) acquired from a patient undergoes a series of tissue preparations, i. e., *fixation*, *embedding*, *sectioning*, and *staining*, before being examined by a pathologist, as illustrated in Figure 1.1. During *fixation*, the specimen is placed into a small container of fixative to preserve the cells and the tissue from decaying. It maintains the cellular structure as close to the native state. Afterwards, the tissue is placed into a *cassette*, a unique numbered bar coded plastic box, and loaded into a tissue processor. The tissue is dehydrated with alcohol to effectively replace the water in the cells. The alcohol is then removed and infiltrated with paraffin wax to enable tissue preservation. In the *embedding* step, the cassettes are removed from the processor, and a histotechnologist orients them into a rectangular mold, filled with more wax, to create a tissue block. The block is laid on a solid tray of ice for even cooling and hardening the wax. Subsequently, the hardened block is *sectioned* by a microtome into thin slices and laid on a glass slide. A glass slide with a tissue slice is *stained* uniformly with a designated dye to highlight particular morphologies of interest in the tissue. Following a final quality check, the stained slides are assigned to a specialized pathologist for examination, who analyzes them under a microscope for tissue characterization. The observations are collected in a succinct report upon assessment. For a patient suspected of or diagnosed with a disease, advanced stainings are performed to highlight specific tissue characteristics for ascertaining the diagnosis and treatment selection.

1.1.2 Digital pathology and computer-aided diagnosis

With the advancements in slide-scanning technologies, glass slides with tissue specimens are digitized into WSIs with impressive resolution. These advancements have led to digital pathology (DP), which has not only replaced the microscopes by computers for tissue examination, but also has profoundly transformed the daily practice of pathologists. The digitalization facilitates the convenience of dealing with images instead of glass slides, and enables the process of interpretation, management, and automatic analysis using computational approaches. Telepathology is enabled by sharing the images with distant locations in real-time, which bridges the physical distance among hospitals, pathologists, and patients. The images can be stored and accessed from a central cloud-based repository, which promotes remote diagnosis, teleconsultation, workload efficiency, collaborations, central clinical review, and virtual education [Wilbur et al., 2009; Hamilton et al., 2012; Sagun et al., 2018; Nauhria et al., 2019; Pantanowitz et al., 2018; Hanna et al., 2019]. However, a digital diagnostic workflow requires pathologists to manually evaluate the digitized slides, analogous to a non-digital diagnostic workflow. The manual examination poses several challenges: (1) a cumbersome and time-consuming process, which demands the pathologists to analyze large volumes of information on slides, and (2) it is prone to high inter- and intra-observer variability, thereby reduced reproducibility of assessments. These challenges are further gaining prominence with the yearly increasing incidence rate of diagnostic cases [Sung et al., 2021].

Recently, computer-aided diagnosis (CAD) systems are developed to address the aforementioned challenges. A CAD system, empowered by DP, aspires to automate and assist pathologists in tissue examination with high accuracy, throughput, and reproducibility. Figure 1.1 illustrates DP and CAD workflows along with the fundamental tissue preparation and digitization steps. The basis of CAD is computational pathology (CP), which leverages innovative research in artificial intelligence (AI), especially ML and deep learning (DL), to address various histopathology tasks, such as:

- Detection, segmentation, and quantification of tissue constituents, e. g., (1) detection and segmentation of nuclei for analyzing nuclear morphology; (2) segmentation and analysis of glandular structures as the key criterion for cancer grading; (3) segmentation of various tissue types, such as epithelium, stroma, adipose tissue; and (4) segmentation of tumor regions in the tissue.
- Classification of histopathology images and its constituents, e. g., (1) nuclei subtyping; and (2) tumor proliferation and aggressiveness assessment, i. e., staging and grading, respectively, in different tissue types, e. g., breast, prostate, and colon.
- Disease prognosis, e. g., prediction of cancer susceptibility, recurrence, and survival.

1.2 Motivation

DL methods automatically identify sub-visual patterns in histopathology images and discover informative features which encode different explanatory factors to address a histopathology task. Recently, these methods have achieved performance comparable to expert pathologists for a number of tasks with high throughput and improved reproducibility [Businesswire, 2021; Bulten et al., 2021]. Despite the success of the DL methods,

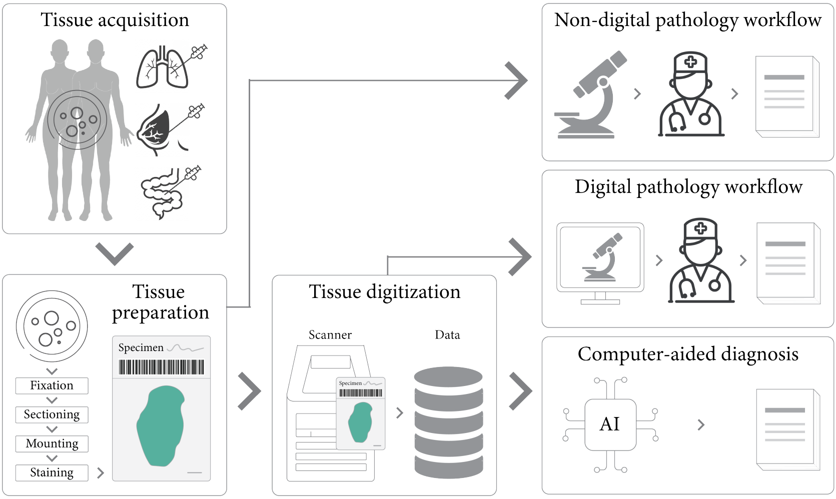


Figure 1.1: Overview of histopathological examination in a clinical setting. The figure illustrates the operational differences among the conventional non-digital pathology, the digital pathology, and the computer-aided diagnostic workflows. Tissue acquisition and preparation are the fundamentals of the workflow. Subsequent tissue digitization is the basis to enable digital pathology and computer-aided assessment.

they still need to address several key challenges for being adopted in routine pathological workflows [Tizhoosh et al., 2018], including:

- *Lack of labeled data*: DL methods typically require large sets of high-quality data that must ideally be “labeled”. This necessitates pathologists to manually generate large labeled datasets, which is tedious, time-consuming, expensive, and incurs observer-variability.
- *Large image sizes*: The advanced scanning technologies have enabled to yield gigapixel sized histopathology images. End-to-end processing of these images by a DL method is challenging due to high computational requirements. As a naive workaround, the images can be downsampled before processing. However, this would limit access to diagnostically relevant low-level information, thereby hampering the task performance.
- *Lack of transparency and interpretability of DL decisions*: A trained DL model acts as a “black box” [Castelvecchi, 2016], which impedes the fundamental requirement for a dependable diagnosis, i. e., transparency and interpretability of the decision-making procedure. Although several researchers have investigated creative ways to explain the model decisions, at present there is no established way to comprehensively explain a specific model decision for a given histopathology scan.

Another set of limitations of DL methods in CP concern with the architectures of the DL models. Most DL models in CP so far are built on CNNs [Deng et al., 2020] due to their large-scale success in computer vision [Li et al., 2021]. These models have achieved

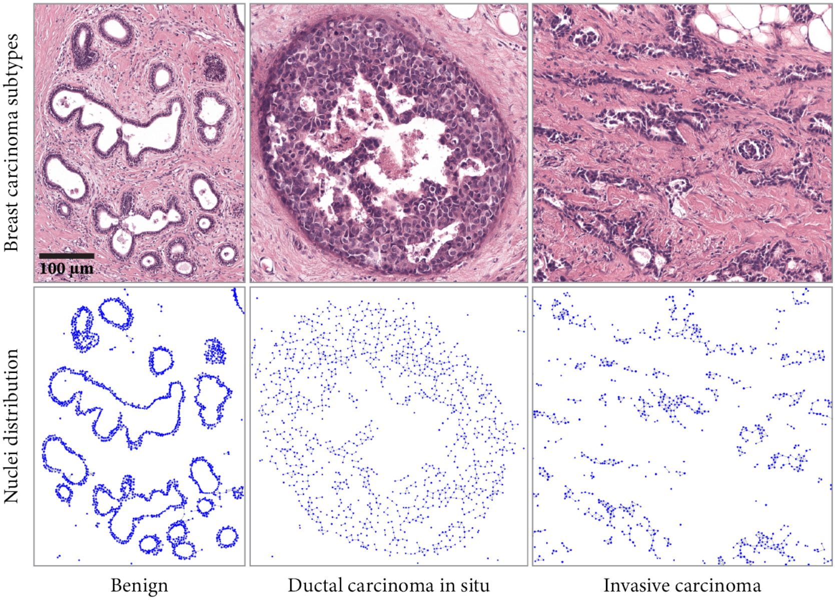


Figure 1.2: H&E stained breast tumor regions from different carcinoma subtypes (top row). Respective cell-graphs illustrating the distribution of epithelial nuclei (thresholded k-nearest neighbor topology ($k = 5$)) (bottom row).

remarkable advances in addressing various histopathology tasks [Srinidhi et al., 2021]. However, they pose several limitations, such as:

- *Detachment from histological entities*: A tissue specimen constitutes of several *histological entities*, e. g., nuclei, cells, tissue types, and glands. The phenotype and topological distribution of these entities characterize the tissue functionality. For instance, Figure 1.2 illustrates three Hematoxylin & Eosin (H&E)-stained breast carcinoma subtypes and their respective nuclei distributions. It can be observed that the benign tissue is composed of glandular organization of epithelial nuclei, whereas the invasive carcinoma includes a fragmented distribution. Thus, to aptly comprehend a tissue composition, computational techniques are imperative to operate on tissue representations which adequately encode the characteristics of the entities. However, CNNs construe a tissue as a set of pixels, which disregards the notion of entities and their organization [Hägele et al., 2020].

  Further, a pathologist examines a tissue specimen in terms of the type and organization of the constituting histological entities. For instance, pathologists focus on necrotic areas built up inside tumors to estimate the aggressiveness of ductal carcinoma in situ in breast tissue [Salvatorelli et al., 2020]; and they analyze lymphocyte infiltration into tumor epithelium for disease prognosis [Idos et al., 2020], as shown in Figure 1.3. Such prior knowledge, included as inductive bias into tissue modeling, can improve the diagnostic performance of DL methods. However, CNNs are

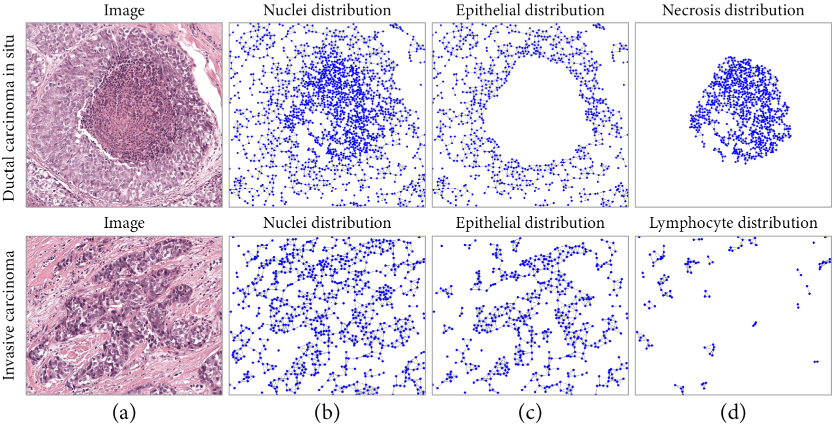


Figure 1.3: (a) H&E stained breast tumor regions from ductal carcinoma in situ and invasive carcinoma, (b) cell-graphs of nuclei distributions, (c) cell-graphs of epithelial nuclei distributions, (d) cell-graph of necrosis (top row) and lymphocytes (bottom row) distribution. Cell-graphs are constructed with k-nearest neighbor topology ($k = 5$).

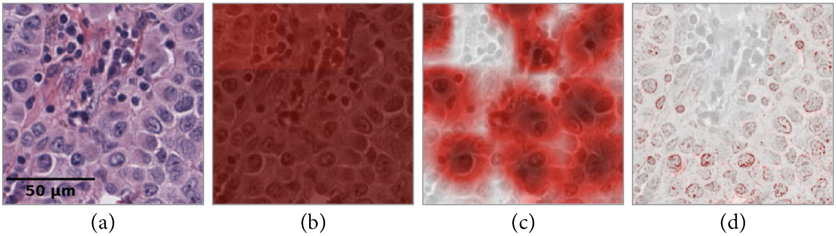


Figure 1.4: Explanation heatmaps for class cancer produced by post-hoc feature attribution techniques applied to a CNN classifier. Input H&E image (a) is split into nine patches and heatmaps are produced by using (b) probability map, (c) Grad-CAM, and (d) Layer-wise Relevance Propagation.

limited to utilize such knowledge due to their inflexibility to selectively analyze entities and their distribution. Moreover, a disregard to histological entities in pixel-based analysis curtails the interpretability and explainability of CNNs. To discern the focus of a CNN during the diagnosis, various post-hoc feature-attribution techniques are employed by [Hägele et al., 2020], as illustrated in Figure 1.4. The presented heatmaps highlighting a subset of pixels are non-localized, i. e., they do not specify the set of relevant entities. Further, they impart distinct emphasis on pixels that belong to an individual entity. Therefore, these heatmaps are not thoroughly explainable to pathologists. Hence, DL methods are desired to operate on comprehensible entity-guided tissue representations, which the pathologists can directly relate to and reason with. This way, prior pathological knowledge can be seamlessly incorporated into the analysis for an enhanced model performance and interpretation.

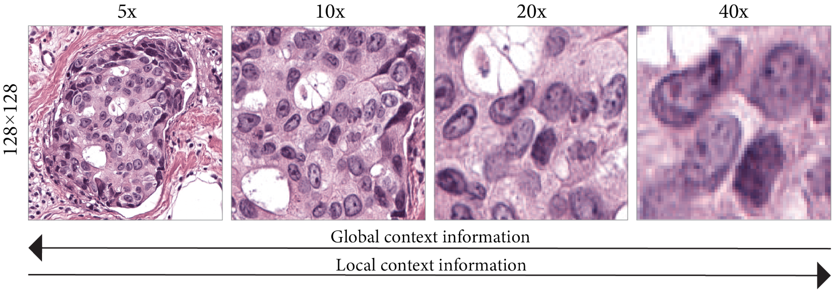


Figure 1.5: Field-of-view of fix-sized sized concentric patches extracted from a tissue region at different resolutions. The arrows at the bottom depict the variation in embedded global- and local-context information in the patches across different resolutions.

- *Trade-off between capturing local- and global-context*: A CNN typically processes a histopathology image in a patch-wise manner due to its large size. The model splits the image into a set of fixed-size patches at a particular resolution, extracts patch-wise representations, and aggregates patch-level features to perform image-level tasks. However, a patch-wise processing incurs the trade-off between capturing adequate local- and global-context information, as illustrated in Figure 1.5. Addressing this challenge is crucial for diagnostic performance [Bejnordi et al., 2017a; Sirinukunwattana et al., 2018]. To highlight, operating at a lower resolution captures global tissue microenvironment but hinders the resolvability of cells and cellular properties, whereas, operating at a higher resolution captures local cellular characteristics but constrains the access to global tissue microenvironment. An effort to simultaneously capturing both information encounters computational bottlenecks. The methods by [Bejnordi et al., 2017a; Sirinukunwattana et al., 2018; Pinckaers et al., 2020] address this trade-off by increasing the visual context per patch, but still are restricted to incorporate information from arbitrary distanced tissue regions. Though, the method by [Tellez et al., 2021] captures both information, it is confined to work with only rectangular input images. Since tissues in the images can be of arbitrary shapes and sizes, working with rectangular inputs obligates to process both informative tissue and uninformative non-tissue regions. Therefore, DL methods are desired to efficiently capture and analyze both local- and global-context simultaneously.
- *Scalability to large histopathology images*: Most of the natural images processed by CNNs contain millions of pixels, whereas a WSIs contains billions of pixels. Such large size requires to develop specific DL methods that can operate on an entire WSI. To this end, Multiple Instance Learning (MIL) [Campanella et al., 2019] and compression-based methods [Tellez et al., 2021; Shaban et al., 2020] are proposed. MIL methods treat a WSI as a bag of patches, and aims to characterize the bag. This ignores the spatial relationship among the patches, and incurs the aforementioned context trade-off. In contrary, compression-based methods demand high computational resources as they tend to process both informative and uninformative tissue regions. Therefore, DL methods are desired to directly operate on histopathology images of arbitrary shape and size with high efficiency.

To address the aforementioned limitations of CNN-based methods, histological entity-based methods are proposed by [Demir et al., 2004]. An entity-based method transforms a histopathology image into an entity-graph, where the nodes and edges denote histological entities and entity-to-entity interactions, respectively. Such representations explicitly encode meaningful histological entities, thus are closer to pathological comprehension. An entity-graph simultaneously captures both local- and global-context information in the graph nodes and topologies, respectively. These representations are flexible enough to describe tissues of arbitrary shapes, and can accommodate large number of nodes and edges to encode arbitrary sizes. Notably, the analysis of these graphs is more computationally efficient than the pixel-based counterparts, as a graph data structure is memory efficient than images. Subsequent to the entity-graph representations, traditional ML [Demir et al., 2004; Sharma et al., 2015] and modern DL [Zhou et al., 2019a; Wang et al., 2019a; Adnan et al., 2020] methods are employed to process the graph-structured data for various histopathology tasks. Considering these advantages, DL methods on entity-graphs are gaining popularity in CP [Ahmedt-Aristizabal et al., 2021]. However, these methods have only been applied to image-level tasks, such as tumor classification, and survival analysis. Further, the designed entity-graphs are limited to only nuclei, as the *units of explanation*, and simple graph topologies, such as k-Nearest Neighbors (k-NN). These representations are inadequate to comprehensively capture a tissue composition.

1.3 Thesis Goals

Though DL methods have achieved remarkable success in CP, they incur several challenges and limitations, as described above. This thesis aims to tackle the above-mentioned issues, in order to develop enhanced methods on several histopathology tasks. To this end, this thesis advocates for an entity-based analytical workflow, as illustrated in Figure 1.6, instead of the conventional patch-based analytical workflows. The specific objectives of the thesis can be summarized as:

- To develop alternate representations of histopathology images that can comprehensively encode tissue compositions embedded in the images. These representations are desired to be interpretable and customizable to seamlessly accommodate task-relevant prior pathological knowledge;
- To develop suitable DL methods that can operate on the above alternate tissue representations and learn appropriate mappings between the tissue compositions and any targeted task;
- To develop DL-based methodologies that can directly operate on histopathology images of arbitrary shapes and sizes;
- To develop DL-based methodologies that can simultaneously take advantage of multi-scale information embedded in histopathology images, and produce contextualized tissue representations.
- To develop DL-based methodologies that can learn from weak- or limited-annotations, thereby alleviating the annotation burden;
- To interpret the DL methods in CP, and express their decisions in pathologist-understandable terminologies.

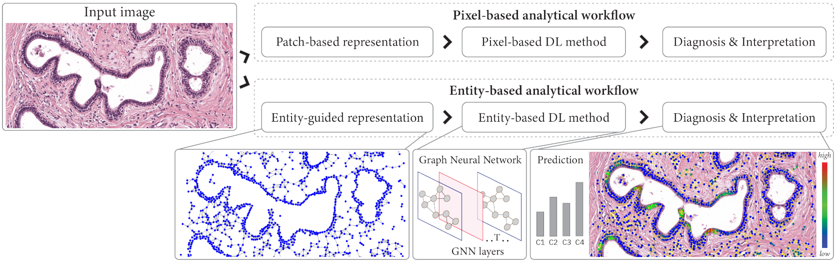

Input image
Pixel-based analytical workflow
Patch-based representation
Pixel-based DL method
Diagnosis & Interpretation
Entity-based analytical workflow
Entity-guided representation
Entity-based DL method
Diagnosis & Interpretation
Graph Neural Network
..T..
GNN layers
Prediction
C1 C2 C3 C4
high
low

Figure 1.6: Illustration of pixel- and entity-based analytical workflows in computational pathology.

1.4 Thesis Outline and Contributions

In the pursuit of the aforementioned goals, the structure of this thesis is outlined below.

Chapter 2 presents a literature review of extant DL methods in CP that concern with learning from limited labeled data, processing of large size images, and interpreting the DL models for various histopathology tasks. The chapter highlights the limitations of these methods to emphasize on the thesis contributions. Further, this chapter presents some essential technical preliminaries into graph representation learning, Graph Neural Networks (GNNs), and post-hoc graph interpretability techniques, to help put in context the proposed methods in this thesis.

Chapter 3 describes the motivation behind encoding a histopathology image in terms of its constituting histological entities. Specifically, diagnostically relevant cell membranes in HER2-stained immunohistochemisty images are delineated, encoded into feature representations, and processed by ML methods to define disease-specific staining quality metrics. Subsequently, a sensitivity analysis of the staining quality variations over the process parametric space is performed to determine disease-specific optimal staining protocols. The findings are substantiated by validating against clinical staining protocols. This work has been published in a peer-reviewed journal as a joint first author [Arar†[1], **Pati**† et al., 2019]. In this work, my contributions were the end-to-end development of the methodology, designing and conducting the experiments, and collaborating with biologists for data curation and result validation.

Considering the significance of building on histological entities, **Chapter 4** proposes a hierarchical entity-graph, a hierarchical cell-to-tissue (HACT) graph, to comprehensively characterize tissue composition. A HACT graph encodes the phenotype, and inter- and intra-entity interactions among a multiset of entities. Afterwards, a novel hierarchical GNN, HACT-Net, hierarchically learns on the HACT graphs for tumor subtyping. A large cohort of H&E stained breast tumor regions is curated to benchmark the proposed method against state-of-the-art CP methods and expert pathologists. Further, detailed ablation studies and qualitative interpretations are conducted to demonstrate the advantages of hierarchical representation and learning. This work has been published in a peer-reviewed journal as a joint first author [**Pati**†, Jaume† et al., 2021]. In this work, my contributions include the idea conceptualization, development of image-to-graph

---

[1] † denotes equal contribution in a shared first authorship.

transformation modules, implementation of uni-level GNNs, software development for efficient data curation, experiments, and validation with pathologists.

Following the classification superiority of entity-guided GNNs, **Chapter 5 & 6** explore the aspects of interpreting and explaining the model decisions. **Chapter 5** proposes a perturbation-based post-hoc interpretability technique (or explainer), CGExplainer, to generate an explanation for a model decision. The explanation denotes the identification of a subset of entities in an input entity-graph which drive the model prediction. The method is applied to understand cell-graph based GNNs for breast tumor classification. Quantitative evaluations demonstrate the sparsity of task-relevant information in the cell-graphs. Qualitative evaluations indicate agreement between the generated explanations and pathological reasonings. This work has been published in a peer-reviewed workshop paper as a joint first author [Jaume†, **Pati**† et al., 2020]. **Chapter 6** explores further into other graph explainers to interpret entity-guided GNNs. Qualitative evaluations illustrate that the differences in underlying mechanisms of the explainers render distinct explanations for an input entity-graph. Further, there is no definite measure to identify the best explanation or the explainer except qualitative analysis by a pathologist. Therefore, this chapter proposes a set of novel quantitative metrics to characterize graph explainers. The metrics are based on the statistics of class separability in terms of pathologically measurable concepts. These metrics, validated by expert pathologists, express the model predictions in pathologically understandable terminologies. This work has been published in a peer-reviewed conference as [Jaume†, **Pati**† et al., 2021]. In these works, my contributions include the conceptualization and development of the quantitative metrics, implementation of a number of pixel- and graph-explainers, design of the experiments, analyses and validations by the pathologists.

Chapter 7 & 8 propose methods to address the challenge of learning from limited labeled data in CP. Specifically, **Chapter 7** proposes CoReL, a classification framework that simultaneously captures class-label information and spatial distribution information of the data points in the embedding space for improving model performance. To exploit the spatial information, DML is leveraged with a novel context-aware pair mining strategy and a novel soft-multi-pair objective. The framework achieves state-of-the-art performance on five benchmark datasets across three histopathology tasks. This work has been published in a peer-reviewed journal as [**Pati** et al., 2021].

Chapter 8 develops a novel entity-graph based weakly-supervised semantic segmentation method, WholeSIGHT, to segment WSIs from image-level labels. WholeSIGHT aggregates the capabilities of GNNs, graph explainers, and DML, to simultaneously achieve state-of-the-art prostate WSI classification performance and weakly-supervised Gleason patterns segmentation performance. The results also demonstrate the scalability of entity-graph based DL methods to large images. Further, a Bayesian extension of WholeSIGHT is proposed to for improved generalization to out-of-domain datasets. The generalizability is quantified in terms of performance assessment, uncertainty analyses, and estimation of model calibration. A preliminary version of this work has been presented in a peer-reviewed conference paper as [Anklin†, **Pati**†, Jaume† et al., 2021]. The extended version is to be submitted in a peer-reviewed journal as [**Pati**†, Jaume† et al., 2021]. In this work, I am responsible for the conceptualization, design and implementation of the methodology, and experimentation and analyses.

In view of the advances in entity-graph based representations and graph-based DL methods in CP, **Chapter 9** introduces HISTOCARTOGRAPHY, a standardized python-based open source library. The library aims to alleviate researchers' effort in developing boilerplate code to perform graph analytics in CP by including adequate preprocessing, machine learning, and explainability tools. The chapter benchmarks the library in terms of computational time and performance on multiple datasets across various histopathology tasks and image scales to illustrate its applicability for building CP workflows. This work has been presented in a peer-reviewed workshop as [Jaume†, **Pati**† et al., 2021]. Both the authors have contributed equally towards the design and development of the library and the modules.

This thesis is concluded in **Chapter 10** with a brief summary of the proposed methods and contributions, their limitations, and a discussion on potential future research directions.

2

Background

Computer-aided diagnosis (CAD) in histopathology aspires to assist pathologists with interpreting histologic findings of interest in a tissue specimen. The core of an effective CAD system in histopathology is computational pathology (CP), which typically uses AI, especially machine learning (ML) and deep learning (DL), to analyze the histopathology images. This chapter reviews DL-based CP solutions for various histopathology tasks, while emphasizing on the researches pertaining to learning from limited labeled data, handling of large size histopathology images, and the interpretability of DL methods. In this course, we also describe the pixel- and the entity-paradigms in histopathology, and review the DL methods in the entity-paradigm. Additionally, we present the technical preliminaries of DL methods operating on graph-structured data, which are relevant for understanding the DL of entity-guided representations in this thesis.

2.1 A Review of Deep Learning in Computational Pathology

Recent advancements in DL have significantly contributed to CP in terms of automatically discovering patterns and inter-pattern relationships from complex histopathology images. The DL methods have the ability to handle gigantic quantity of histopathology data created throughout the patient-care lifecycle, and encode different explanatory factors of variation behind the data to improve the diagnosis, prediction, and disease prognostication. The methods operate in a data-driven manner, and include complex and autonomous techniques for rendering a better mapping between the underlying data distribution and a targeted histopathology task. The quality of the mapping relies heavily on the level of human supervision in the ingested data. The DL methods using labeled data, i. e., paired input data and desired output labels acquired via human supervision, produce superior mapping, thus superior task-specific performance, compared to the methods based on unlabeled data, i. e., standalone input data. Although, there exist several learning schemes to handle specific data ingestion scenarios, the DL methods in CP can be majorly categorized into supervised learning and unsupervised learning. This section first presents several DL methods in CP associated with the aforementioned learning schemes. Subsequently, this section describes specific methods that are developed to address the challenges and limitations of applying DL in CP, as detailed in Section 1.2.

2.1.1 Deep learning methods based on training supervision

Supervised learning: Supervised learning methods are trained using labeled data to map the input-output pairs. The type of ingested supervision depends on the histopathology task. For example, image-level labels are required for image classification, whereas pixel-level labels are required for segmenting objects in the image. Further, point or bounding box labels are required for object detection, but instance-level pixel labels are required for segmenting object instances. Typically, the task-specific annotations are acquired by the consensus of multiple pathologists to alleviate observational bias in the labeling process. Notably, the type of the annotation determines the design of the DL methods. For instance, the design of nuclei detection methods differ based on point-wise or bounding box annotations.

Numerous supervised DL methods are proposed to address several histopathological tasks across different tissue types, e. g., breast, prostate, colon, lung, and stomach, and stainings, e. g., H&E, immunohistochemistry, and immunofluorescence, by using different annotation types. A few examples are showcased as follows, nuclei detection [Sirinukunwattana et al., 2016; Romo-Bucheli et al., 2016; Sornapudi et al., 2018; Höfener et al., 2018; Xing et al., 2019; Yang et al., 2020], nuclei segmentation [Kumar et al., 2017; Naylor et al., 2019; Zhou et al., 2019b; Graham et al., 2019a; Verma et al., 2021], nuclei classification [Sirinukunwattana et al., 2016; Li et al., 2018d; Zhou et al., 2018; Graham et al., 2019a; Verma et al., 2021], mitosis detection [Roux et al., 2013; Veta et al., 2015; Roux, 2014; Veta et al., 2019], semantic segmentation of cellular objects [Ciresan et al., 2012; Ronneberger et al., 2015; Song et al., 2015; Zhang et al., 2017], cell detection [Xie et al., 2015; Kashif et al., 2016; Wang et al., 2016; Xie et al., 2018], tissue segmentation [Xu et al., 2016a], tissue classification [Kather et al., 2019; Xu et al., 2019b], gland segmentation [Li et al., 2016; Xu et al., 2017a; Kainz et al., 2017; Graham et al., 2019b; Binder et al., 2019], tumor segmentation [Bejnordi et al., 2017b; Cruz-Roa et al., 2017; Sharma et al., 2017b; Qaiser et al., 2019a; Wei et al., 2019; Tokunaga et al., 2019], tumor classification [Hou et al., 2016; Couture et al., 2018; Shaban et al., 2019; Nagpal et al., 2019; Bulten et al., 2020a; Rathore et al., 2020], and survival prediction [Zhu et al., 2017; Bychkov et al., 2018; Mobadersany et al., 2018; Courtiol et al., 2019; Kather et al., 2019].

To promote supervised DL methods in CP, a number of labeled datasets are released. Some of the notable datasets for different histopathology tasks are, mitosis detection [Roux et al., 2013; Veta et al., 2015; Roux, 2014; Veta et al., 2019], gland segmentation [Sirinukunwattana et al., 2017; Graham et al., 2019b], nuclei detection [Sirinukunwattana et al., 2016], nuclei segmentation [Graham et al., 2019a; Kumar et al., 2020; Verma et al., 2021], nuclei classification [Graham et al., 2019a; Verma et al., 2021], metastasis detection [Bejnordi et al., 2017b; Bandi et al., 2018], tissue classification [Kather et al., 2019], cancer subtyping [Li et al., 2018e; Aresta et al., 2019], HER2 scoring in breast cancer [Vandenberghe et al., 2017; Qaiser et al., 2018], and Gleason scoring and ISUP grading in prostate cancer [Arvaniti et al., 2018; Silva-Rodrìguez et al., 2020]. These well-defined public datasets have drawn significant interest of the computer vision community for designing CP solutions.

Despite the success of supervised DL methods in CP, the main challenge lies in the acquisition of task-specific high-quality large labeled datasets that capture real-world data variations. To address this challenge, various methods are proposed to learn from limited labeled data, as presented in Section 2.1.2.

Unsupervised learning: Unsupervised learning methods are trained using unlabeled data and aspire to identify intrinsic data representations that can disentangle relationships among the data points. Consequently, these methods aim to group (or cluster) the data points into separate categories by deciphering the underlying data distribution. These methods are desirable as they can be interpreted in terms of their quality of understanding about the underlying data distribution. However, without any labels, the learning task is ambiguous as it can possibly map the input data into infinitely many subsets of representations. Therefore, most unsupervised approaches aim to construct data representations under certain constraints, such that the potential representation subsets are limited and achieve a desired grouping. Some popular categories of unsupervised DL approaches and their learning constraints are presented as follows:

- *Auto-encoder* learns compressed representation by using an encoder and a decoder sub-model. It is constrained to reconstruct the ingested input, where the encoder compresses the input into a latent representation and the decoder reconstructs the input from the representation. Advances in modeling the stochasticity [Kingma et al., 2014] and more robust feature disentanglement [Higgins et al., 2016; Chen et al., 2018] have made autoencoders more attractive.

- *Generative adversarial network* [Goodfellow et al., 2014] uses two neural networks, a generator and a discriminator, to optimize an explicit min-max objective such that the instances generated by the generator are indistinguishable compared to real instances by the discriminator. The features from the discriminator are typically considered as the data representation [Mao et al., 2019]. However, a number of advanced generative adversarial networks are dedicated to produce other types of informative representations [Larsen et al., 2016; Donahue et al., 2017].

- *Self-supervised learning* designs unsupervised auxiliary supervision tasks to exploit some intrinsic information available in the data. Existing methods vary in the design of the auxiliary tasks, such as spatial context [Doersch et al., 2015; Noroozi et al., 2016], spatio-temporal continuity [Wang et al., 2015; Wang et al., 2017], colour patterns [Zhang et al., 2016a; Larsson et al., 2016], image inpainting [Pathak et al., 2016], and pair similarity in embedding space [Chen et al., 2020b; Caron et al., 2020]. Recently, a few histopathology-specific auxiliary tasks are also proposed to produce better domain-specific data representations [Koohbanani et al., 2021].

- *Deep clustering* methods [Xie et al., 2016; Jiang et al., 2016] combine feature extraction, dimensionality reduction, and clustering into an end-to-end model, allowing deep neural networks to learn data representations that adapt to various clustering criteria.

Several unsupervised DL methods based on aforementioned approaches are employed to tackle various histopathology tasks, such as, nuclei detection [Xu et al., 2016b] nuclei segmentation [Yao et al., 2021b; Liu et al., 2020], tissue classification [Ciga et al., 2020; Koohbanani et al., 2021], tissue segmentation [Ciga et al., 2020; Mahapatra et al., 2021a], tumor classification [Dercksen et al., 2019; Muhammad et al., 2019; Li et al., 2019b; Ciga et al., 2020; Koohbanani et al., 2021; Mahapatra et al., 2021c; Wang et al., 2021], tumor segmentation [Gadermayr et al., 2018; Roy et al., 2021; Mahapatra et al., 2021b], survival prediction [Zhu et al., 2017; Yamamoto et al., 2019; Abbet et al., 2020], histology image registration [Awan et al., 2018; Hecht et al., 2020], and anomaly detection [Schlegl et al., 2017; Pocevičiūtė et al., 2021]. Though unsupervised methods are desired for unbiased representation learning, they are limited to model the distribution of complex and diverse histopathology images in the absence of labeled data. However, the recent researches in unsupervised DL approaches, especially self-supervised approaches, hold promising potential for advancing unsupervised methods in CP [Ciga et al., 2020].

2.1.2 Learning from limited labeled data

DL methods are achieving new state-of-the-art performances on various histopathology tasks. These gains are achieved at the cost of acquiring large annotated high-quality training data. However, acquiring such annotations are often infeasible due to tedious, time-consuming, and expensive labeling procedure, and the unavailability of expert pathologists. Ideally, we expect DL methods to address any histopathology task in an unsupervised manner, but unsupervised DL methods at the moment are unable to deliver the desired performance for clinical use. Hence, learning from limited labeled data is gaining popularity to balance the trade-off between generating labeled data and learning performance. To this end, several research directions are being explored, such as,

- *Semi-supervised learning:* These methods learn from a few labeled and a large number of unlabeled data points. To maximally utilize the labeled and unlabeled data, semi-supervised learning methods make certain assumptions about the underlying structure of the data, such as smoothness of the data point neighborhood and clustering of similar data points in the embedding space. Another popular approach is to incrementally train a DL method by predicting labels for the confident data points in the unlabeled set, and retraining the DL method with the augmented labeled dataset. These methods are applied to various histopathology tasks, such as histopathology image classification [Foucart et al., 2019; Marini et al., 2021; Su et al., 2021; Peikari et al., 2015], and segmentation [Li et al., 2018b; Yu et al., 2021].
- *Active learning:* These methods aims to learn an algorithm which can automatically select the informative unlabeled data points for the training of the method. Over multiple training iterations, these methods select informative data points, from a large unlabeled dataset, to be annotated, get the data points labeled by human experts, and include the labeled data into the training procedure. Labeling of only a subset of the dataset, identified by the learning method, reduces the overall expense of building an effective model. Primarily, active learning methods differ in terms of their querying strategies to select the relevant data points to be labeled. A few applications of their applications in CP are, tumor segmentation [Folmsbee et al., 2021], nuclei segmentation [Wen et al., 2018], nuclei detection and classification [Carse et al., 2019], analysis of nuclei pleomorphism [Cosatto et al., 2008], and gland segmentation [Yang et al., 2017].
- *Weakly-supervised learning:* These methods exploit coarse-grained labels, e. g., image-level labels, to automatically infer fine-grained, e. g., pixel/patch-level, information. These methods are well suited for histopathology applications where the coarse-grained labels are readily available from clinical reports, but acquiring fine-grained labels are tedious, expensive, or infeasible. Among several weakly-supervised approaches, MIL is a very popular method for analyzing giga-pixel sized WSIs to perform region-level tasks [Jia et al., 2017; Liang et al., 2018] and whole-slide-level tasks [Ilse et al., 2018; Campanella et al., 2019; Wang et al., 2019d; Wang et al., 2019e; Lu et al., 2021; Yao et al., 2020]. In MIL, a whole-slide is decomposed into a bag of small high-resolution patches (or instances), and bag-level labels are used to predict both bag-level and instance-level labels. Other types of weakly-supervised methods

alleviate the need for expensive labeling by proposing various loss functions [Li et al., 2019a; Silva-Rodrìguez et al., 2021], feature encoding strategies [Akbar et al., 2018; Tellez et al., 2021; Shaban et al., 2020], loss balancing mechanisms [Bokhorst et al., 2018], to use feature attribution techniques [Chan et al., 2019], and to derive pseudo labels from weak annotations [Qu et al., 2019] for training supervised learning methods.

- *Augmentation and data generation:* Augmenting the available labeled data or synthesizing labels for unlabeled data aim to increase the labeled dataset for training effective DL methods. Data augmentation techniques, such as morphological transformation, e. g., rotation, translation, and scaling, and stain color augmentation, i. e., perturbing images in respective staining color space [Liu et al., 2017b; Tellez et al., 2019], are applied to the available labeled images and annotations to increase the size of the labeled dataset. However, these techniques are limited by their inability to address the inherent problem of dealing with a small training labeled set, which does not comprehensively represent the underlying data distribution. Therefore, generative methods are proposed to generate synthetic labeled data which can augment the size and the diversity of the training dataset. To this end, Generative Adversarial Networks [Goodfellow et al., 2014] are used to synthesize realistic ground-truth label maps from unlabeled histopathology images. These methods treat the label synthesis as an image-to-image translation task. Applications of such methods have addressed nuclei segmentation [Mahmood et al., 2019; Hou et al., 2019] and nuclei detection [Bug et al., 2019; Hou et al., 2019]. A summary of recent state-of-the-art developments and potential future applications of image synthesis in histopathology is detailed in [Tschuchnig et al., 2020].
- *Transfer learning:* These methods have become the de-facto methods for applying DL under limited labeled data scenarios. Transfer learning aims to extract knowledge from a well-defined source domain and apply it to a target domain to expedite and improve the learning. The knowledge is encoded in the form of weights of the DL method. Most of these methods in CP leverage the knowledge extracted from natural images, and are applied for cell detection [Valkonen et al., 2020], genomic prediction [Coudray et al., 2018], tumor classification [Tabibu et al., 2019; Ström et al., 2020], survival prediction [Tabibu et al., 2019], etc. Recently, relevant knowledge is being extracted from other well-defined histopathology tasks [Mormont et al., 2020; Khan et al., 2019], or by pretraining the DL method directly on the target data via unsupervised learning strategies, such as self-supervision, contrastive learning, and metric learning. However, there exist several challenges in applying transfer learning to a target domain. First, the target domain images need to be transformed as per the input requirements of the pretrained models on the source domain. Such transformation may cause significant information loss, thus affecting the task performance. Second, the principles of transfer learning concerning, (1) the domain gap between target and source domain, and (2) the amount of labeled target domain data, need to be carefully followed for a meaningful transfer of knowledge. Third, the transferred knowledge may significantly improve the learning speed of the method on the target domain but may not necessarily conclude a better performance. Evaluations of the effect of transfer conclude that transfer offers

little benefit to performance, which may be due to the over-parameterization of pretrained models [Raghu et al., 2019].

The aforementioned methods pose individual operating requirements based on the amount of available labeled and unlabeled data, potential human participation, type of label, and diversity of the labeled data. Consequently, each method pose individual advantages and disadvantages for certain operating environment. In case of an operating environment suiting the requirements of multiple methods, the combinations of methods can be developed to complement each other for maximally utilizing the available data [Wang et al., 2019a; Chen et al., 2020a; Lai et al., 2021; Otálora et al., 2021].

2.1.3 Processing large histopathology images

The advancements in scanning technologies have enabled scanning glass slides with tissue specimens at high resolution to produce high-quality WSIs. The WSIs are typically of the order of 100,000×100,000 pixels and contain more than 1 million descriptive objects. In a clinical setting, multiple tissue specimens are collected from a patient resulting several WSIs per patient. Therefore, the WSIs datasets are considered as large-scale image analysis applications. The inherent large-scale property of such datasets pose several challenges for applying DL methods on them. DL methods need to be time, memory, and computation efficient, while extracting as much information as possible from the WSIs. Predominantly, two types of DL methods described below are applied in this case: MIL and compression-based learning.

The MIL technique decomposes a WSI into a bag of patches, where the bag holds the image-level label. The patches are encoded by a CNN to produce patch-level features, which are aggregated to produce WSI-level features for further associating with the WSI-level label via a DL model. The CNN backbone and the ultimate DL model can be trained end-to-end to learn meaningful image representations. Some applicatiions of MIL in CP are, tumor segmentation [Jia et al., 2017; Liang et al., 2018], whole-slide tumor classification [Ilse et al., 2018; Campanella et al., 2019; Wang et al., 2019d; Lu et al., 2021], and survival prediction [Yao et al., 2020]. The advantages of MIL methods are, they can scale to arbitrary image sizes and perform end-to-end learning with low memory requirements. However, the performance of MIL is impacted by the trade-off between resolution of patch extraction and the patch-level captured context. These methods also disregard the spatial distribution of the tissue in the process of creating a bag representation of a WSI. Further, MIL assumes that a single patch correlating with the bag label can predict the image label, which makes MIL prone to produce false positive predictions because of a single adversarial patch.

A compression-based method operates on a WSI in three steps. First, it extracts consecutive patches from a WSI and encodes them into low-dimensional features via a CNN. Second, it stitches the patch-wise features to form a WSI-level feature-cube representation, which is a compressed version of the WSI due to the low-dimensional encoding of the patches. Finally, a supervised learning method is applied on the feature-cube to map to the slide-level label. The CNN backbone and the supervised learning method can be trained in an end-to-end manner. Some examples of compression-based methods are, WSI tumor classification [Tellez et al., 2021; Shaban et al., 2020], and WSI tissue segmentation [Tellez et al., 2021]. Compared to MIL, the compression-based methods encode dependencies among patches. However, as these methods operate on all the patches constituting a WSI, they require more computations and memory during processing. Further, these methods are obligated to process patched from uninformative background regions, which constitute a large percentage of areas in WSIs.

A few other techniques are being researched to handle large histopathology images, such as, memory-efficient methods that enable CNNs to be trained with large image-tiles [Kong et al., 2007; Pinckaers et al., 2021], and reinforcement learning methods to increase the evaluated field-of-view in a WSI [Dong et al., 2018; BenTaieb et al., 2018; Qaiser et al., 2019b]. However, all these methods inherit the limitations of CNNs operating on histopathology images, such as disregard to histological entities, and inability to efficiently process arbitrary image shapes, as described in Section 1.2.

2.1.4 Interpretability of deep learning methods

DL methods have successfully addressed several histopathology tasks with high predictive performance and throughput while ensuring objectivity and reproducibility of the assessment. These benefits come at the cost of reduced transparency in decision-making process [Holzinger et al., 2017; Tizhoosh et al., 2018; Hägele et al., 2020] due to the inherent "black-box" nature of the DL methods. Since, transparency, interpretability, and explainability are fundamental requirements of any clinical decision, it is imperative to enable the interpretability and explainability of DL decisions to pathologists for the adoption of DL methods in a clinical setting. These requirements for a DL method are less important when carrying out low-level tasks, e. g., nuclei detection, classification, and segmentation, which can be readily verified by pathologists. However, these requirements are indispensable for a DL method that performs high-level tasks, e. g., grading, staging, survival prediction, and treatment selection.

Inspired by the interpretability techniques for DL methods on natural images [Simonyan et al., 2013; Zeiler et al., 2014; Yosinski et al., 2015; Bach et al., 2015; Montavon et al., 2015; Selvaraju et al., 2017; Kindermans et al., 2015; Zintgraf et al., 2017; Chattopadhay et al., 2018; Kim et al., 2018], several interpretability techniques have been developed to acquire insights about DL methods in CP. Primarily, these techniques in CP can be categorized into, feature attribution-based methods [Korbar et al., 2017; Binder et al., 2018; Hägele et al., 2020], concept attribution-based methods [Graziani et al., 2020], attention-based learning [Lu et al., 2021], and image captioning [Zhang et al., 2019]. The feature attribution techniques are post-hoc in nature, i. e., they operate on trained DL models, and produce visual explanations in terms of heatmaps for an input image. A feature attribution technique, applied to a DL method operating on an image, aims to identify a subset of pixels in the image that positively influence the prediction of the DL model. However, the pixel-level explanations produced by these techniques pose several notable issues, including: (1) a pixel-wise analysis disregards the notion of histological tissue entities, their topological distribution, and inter-entity relationships, thus any explanations generated are detached from pathological comprehension; and (2) the generated explanations are often blurry. Further, the feature attribution techniques applied to a typical DL method, which processes a large histopathology image in a patch-wise manner, fail to accommodate complete tissue microenvironment information; and produce unintuitive patchy heatmaps. Differently, a concept-attribution technique, applied to a DL method operating on an image, evaluates the sensitivity of the DL model's prediction with respect to quantifiable pathological *concepts*, extracted from the image, to highlight the relevance of the *concepts*. A typical DL method processes a large histopathology image, e. g., a WSI, in a patch-wise manner, thus to apply the concept-attribution techniques, the *concepts* need to be extracted at patch-level. However, patch-level concepts are neither fit nor meaningful to interpret the WSI, which contain various localized *concepts*. Furthermore, attention-based learning and image captioning techniques, i. e., a multimodal mapping between an image and corresponding diagnostic report, are devised to localize the focus of a DL model during inference on parts of an input image. However, the pixel-wise and patch-based processing in these techniques incur the same issues as the feature attribution-based methods.

Notably, all the aforementioned interpretability techniques, when applied to a DL model for an input histopathology image, produce unique qualitative explanations. To select the most suitable explanation, all the generated explanations are required to be assessed by an expert pathologist. However, the qualitative assessment involves several challenges, they are, (1) the explanations being detached from pathological comprehension are difficult to be evaluated by a pathologist; (2) to identify the best interpretability technique, the generated explanations by all the techniques need to be benchmarked on a dataset of histopathology images, which is tedious, time-consuming, expensive, and prone to observer variability.

2.1.5 Deep learning of entity-guided representations

The DL-based histopathology applications described in Sections 2.1.1, 2.1.2, 2.1.3, and 2.1.4 analyze histopathology images in pixel-paradigm. These DL methods consider an image as a composition of multiset of pixels, and operate at pixel-level to map the tissue structure to a histopathology task. However, a histopathology image constitutes of several histological entities, which are organized in a specific order to characterize the tissue. Operating in pixel-paradigm entirely disregards the notion of these entities, which consequently hampers the interpretability, scalability, and ability to incorporate prior knowledge and adequate context information into the DL-based analysis, as detailed in Section 1.2. In view of these limitations with pixel-based processing, DL of entity-guided representations are gaining popularity in CP [Ahmedt-Aristizabal et al., 2021]. An entity-guided tissue representation explicitly encodes the histological entities and their interactions in form of nodes and edges in an entity-graph. An entity-graph is motivated by pathological diagnostic procedure, where pathologists analyze the phenotype and topological distribution of various histological entities, such as nuclei, cells, tissue regions, and glands, to characterize a tissue. Therefore, an entity-graph based processing provides an interpretable input space to the pathologists and enables them to recommend task-specific prior knowledge for encoding inductive bias in tissue modeling and computation. The entity-graph based processing further allows to scale to histopathology images of arbitrary shape and size while incorporating both local- and global-context information.

In view of these advantages, several entity-graph based methods have been proposed across a variety of histopathology tasks. The proposed entity-graph structures primarily differ in terms of the encoded histological entities, and their embedded phenotypes and topologies. Additionally, the methods have leveraged classical machine learning and DL techniques to process the entity-graphs. Interestingly, a number of graph interpretability techniques are proposed recently to interpret and visualize an entity-guided DL model.

In this section, we present a brief overview of the entity-graph based methods in CP, and their interpretability for various histopathology tasks.

Entity-graphs in computational pathology: Entity-graphs are first introduced by [Demir et al., 2004] to characterize brain cancer tissues, where an entity-graph encodes cells as the entities and a Waxman model-based cell-to-cell interactions. Afterwards, classical machine learning methods using the statistics from cell distribution are used to characterize the cells and the tissues. Since then a variety of entity-graphs have been proposed by using, different types of entities, such as nuclei and image patches, different entity-graph topologies, such as k-nearest neighbor, region-adjacency, radial distance, probabilistic models. Further, different types of features, such as hand-crafted features and DL features, are included to characterize the nodes of the entity-graphs. After constructing meaningful entity-graphs from tissues, classical machine learning and deep graph learning methods are employed to address different histopathology tasks. Recent advancements in entity-graph based DL methods have been applied across different tissue types, such as breast [Anand et al., 2019; Ye et al., 2019; Aygüneş et al., 2020; Lu et al., 2020; Ozen et al., 2021], prostate [Wang et al., 2019a; Chen et al., 2020a], colon [Zhou et al., 2019a; Javed et al., 2020; Raju et al., 2020; Zhao et al., 2020a; Studer et al., 2021], lung [Li et al., 2018c; Zheng et al., 2019; Adnan et al., 2020], to address various histopathology tasks, such as disease classification, image retrieval, cellular community detection, and survival prediction. Further, the entity-graph based methods have been applied to extract and combine multiple rich visual representations of the same input data (unimodal fusion) [Shi et al., 2021], or integrate information from various input modalities (multimodal fusion) [Chen et al., 2020a] to enable more accurate and robust decisions.

Interpretability of entity-graphs: Inspired by the interpretability techniques for DL model decisions on graph-structured data [Baldassarre et al., 2019; Pope et al., 2019; Ying et al., 2019], a few developments have been made for histopathology applications. [Wu et al., 2019] proposed a GCN propagated supervisory information over patches to learn patch-aware interpretability in the form of a probability score. [Zhou et al., 2019a] analyzes cluster assignment of nodes in a cell-graph to group them according to their appearance and tissue types. [Sureka et al., 2020] proposed a robust spatial filtering with an attention-based GNN architecture and node occlusion to highlight the cell contributions. [Levy et al., 2020] introduced Graph Mapper, a topological data analysis tool, to compress histological information to its essential structures and capture meaningful histology regions. Notably, majority of the aforementioned approaches have been limited to interpreting only cell-graphs. These works however lack the definition of objectives to validate the quality, utility, and effectiveness of the generated explanations by the interpretability techniques. Further, these explanations are not expressed in pathologically comprehensive terminologies, which is crucial for bridging trust between pathologists and computational approaches.

2.2 Technical Preliminaries on Graph Representation Learning

Graphs are a ubiquitous data structure and a universal language for describing complex systems. In the most general view, a graph is a collection of objects, i. e., nodes, and a set of inter-object interactions, i. e., edges. For instance, a molecule can be represented as a graph, where nodes and edges denote atoms and chemical bonds, respectively. A social network can be represented as a graph, where nodes and edges denote users and user-to-user interactions. A physical system can be designed as a graph, where objects, i. e., nodes, interact with each other through physical forces, i. e., edges. Considering the ever-increasing scale and complexity of graph-structured data in recent times, the analysis of graphs in terms of modeling, comprehension, and prediction is of high significance. For this purpose, the advancements in ML and DL have been leveraged to produce GNNs.

In this section, we begin with formally introducing graphs and the associated relevant properties. We overview various graph related tasks that can be accomplished by neural networks. Afterwards, we introduce GNNs and Message Passing Neural Networks (MPNNs), a type of GNN model. Finally, we discuss several post-hoc techniques to interpret trained GNNs. To highlight, in this section, we only describe a subset of research topics from graph representation learning, that are relevant to understand the contributions of this thesis. The reader can refer to [Wu et al., 2020; Zhou et al., 2020; Hamilton, 2020] for a detailed review on graph representation learning.

2.2.1 Graphs: definitions and notations

A graph $G := (V_G, E_G)$ constitutes of a set of nodes V_G and a set of edges E_G. A directed edge $e_{vu} \in E_G$ for $v, u \in V_G$ is an edge starting from v and ending in u. When there is no ambiguity, for simplicity, the node and edge sets are denoted as V and E, respectively. For each node $v \in V$, we define its neighborhood as $\mathcal{N}(v) := \{u \in V | e_{vu} \in E \vee e_{uv} \in E\}$. While dealing with directed graphs, we distinguish between incoming neighbors $\mathcal{N}^I(v) := \{u \in V | e_{uv} \in E\}$ and outgoing neighbors $\mathcal{N}^O(v) := \{u \in V | e_{vu} \in E\}$. Naturally, $\mathcal{N}(v) = \mathcal{N}^I(v) \cup \mathcal{N}^O(v)$. The cardinalities $d_v = |\mathcal{N}(v)|$, $d_v^I = |\mathcal{N}^I(v)|$, and $d_v^O = |\mathcal{N}^O(v)|$ refer to *degree*, *in-degree*, and *out-degree* of node v, respectively. In this thesis, we are concerned with undirected graphs, i. e., $e_{vu} = e_{uv}, \forall e \in E$.

Further, we are concerned with *attributed* graphs $G := (V, E, H)$, where each node is associated with *attributes*. The node attributes are denoted as, $H \in \mathbb{R}^{|V| \times d}$, and defines at node-level as $H_{v,:} := h(v) \in \mathbb{R}^d$. In this thesis, we do not incorporate edge attributes for our graph data. We refer to *discrete* node attributes as *labels*, i. e., atoms in a molecule, and user names in a social graph. Multi-dimensional *continuous* node attributes are referred to as *features*. Note that, the graph signal processing community usually use the term *signal* to refer to attributes, and the term *embeddings* to refer to processed labels and features. In the literature, the terms graph and network, and the terms nodes and vertices, are often used interchangeably. To avoid confusion with neural networks, and in agreement with the graph community, we only use the terms graphs and nodes.

2.2.2 Machine learning on graphs

In this section, we present an overview of common graph learning tasks (see Figure 2.1).

- **Graph classification task:** It aims to identify the label associated with a graph, analogous to an image classification task. In Figure 2.1, we show a supervised graph classification task that predicts the molecular property of chemical compounds. In CP, graph classification can be used to predict the stage or grade of histopathology images encoded as graphs.

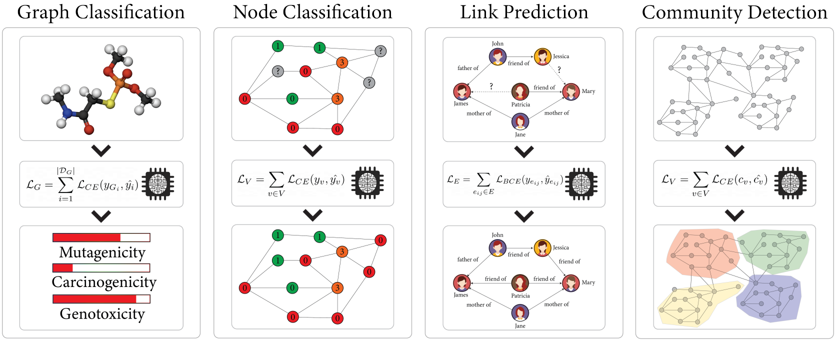

Graph Classification
Node Classification
Link Prediction
Community Detection
$\mathcal{L}_G = \sum_{i=1}^{|\mathcal{D}_G|} \mathcal{L}_{CE}(y_{G_i}, \hat{y}_i)$
$\mathcal{L}_V = \sum_{v \in V} \mathcal{L}_{CE}(y_v, \hat{y}_v)$
$\mathcal{L}_E = \sum_{e_{ij} \in E} \mathcal{L}_{BCE}(y_{e_{ij}}, \hat{y}_{e_{ij}})$
$\mathcal{L}_V = \sum_{v \in V} \mathcal{L}_{CE}(c_v, \hat{c}_v)$
Mutagenicity
Carcinogenicity
Genotoxicity
John
Jessica
James
Patricia
Mary
Jane
father of
friend of
mother of
?

Figure 2.1: An overview of major deep graph learning tasks, i. e., Graph classification: learning graph-level representations to predict graph-level properties; Node classification: operating in a semi-supervised manner to predict unknown node labels by learning from known node labels; Link prediction: predicting missing connections in an incomplete graph; and Community detection: identifying clusters of similar nodes according to the graph topology, and node- and edge-attributes.

- **Node classification task:** It aims to identify the labels of constituting nodes in a graph. Typically, a node classification task operates under a semi-supervised setting, where a model is trained using the known node labels in a graph, and is used to predict the unknown node labels in the graph. This setting breaks the i.i.d. assumption of DL as the nodes to be classified in a graph are dependent on other nodes via edges that are used for model training. While theoretically limiting, such models can still be trained on large graphs without any issue, when the receptive field of the network is smaller than the graph diameter. This is a reasonable assumption in knowledge graphs, social networks, etc. Node classification can be applied to citation network labeling, predicting user's preference on social networks, and recommender systems on retail websites. In CP, this task can be applied to classify nuclei in a cell-graph representation for a H&E stained tissue region.
- **Link prediction task:** It aims to infer missing edges in a large and incomplete graph. It is also referred to as graph completion or relational inference. The setting is similar to a node classification task with a difference that the model is trained to predict the presence or absence of edges between pairs of nodes. Link prediction can be applied in social networks to recommend appropriate new connections, pages, contents, etc., to a user.
- **Community detection task:** It aims to identify clusters of nodes in a graph that belong to the same community or category. This task can be trained in a supervised setting with ground truth node-level labels, similar to node classification, or in an unsupervised manner via graph partitioning. In CP, community detection can be applied to identifying cellular communities for tissue phenotyping.

All these tasks require to build graph- and node-level features by encoding both node attributes and the topological distribution of the nodes. The prime objective of applying ML and DL methods on graph dataset is to understand the composition of the nodes and their design properties.

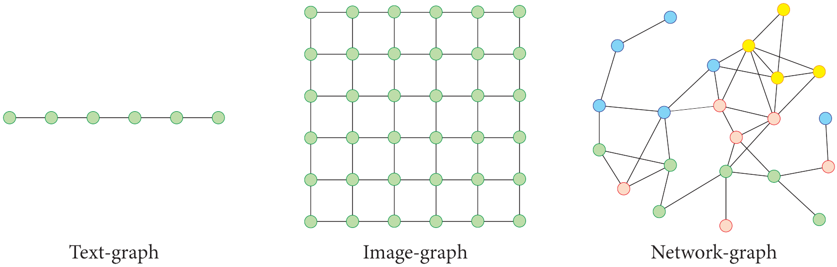


Figure 2.2: Topological distribution of units of explanation, i. e., words, pixels, and objects, in text-, image-, and network-graphs, respectively.

2.2.3 Graph Neural Networks

To analyze graph-structured data using ML or DL methods, a natural question arises to understand whether existing methods designed for handing different data types, such as sequence and images, can work on graphs, and to realize the limitations of these methods.

A first class of neural networks are Recurrent Neural Networks (RNNs) that are designed to operate on sequences. A sequence can be considered as a directed path graph where the nodes and the directed edges lie on a single straight line, as shown in Figure 2.2. Such type of graphs implicitly assume a pre-defined ordering of nodes, which do not hold for generic graphs, where the nodes are neither numbered nor ordered. Therefore, RNNs can be used to model certain types of graphs, i. e., directed path graphs, or when the graphs can be approximated by directed path graphs, e. g., in chemistry, molecular graphs can be transformed into sequences using SMILE representations [Weininger, 1988], and further processed by RNNs [Schwaller et al., 2018]. Thus, RNNs enforce a sequential inductive bias in the network. Another class of neural network are CNNs that are designed to operate on images, which is a regular grid of pixels. Grids can be considered as graphs with a fixed node neighborhood, where each node is connected to its eight nearest neighbors, as shown in Figure 2.2. A fixed node neighborhood allows to apply a fixed-size convolutional kernel to the entire grid, thus inducing a local inductive bias in the network. However, such property does not hold in generic graphs. Feed-forward neural networks can also be seen as operating on a graph-structured data. These networks operate on vectorized inputs by building all-to-all connections among the input features, which is analogous to processing a fully-connected graph.

These considerations highlight that existing neural networks are insufficient to operate on arbitrary sized and complex graph structures that do not have any fixed node ordering or reference point. This motivates the development of a novel class of neural networks, i. e., Graph Neural Networks GNNs. GNNs are expected to generalize to some existing neural networks, as the input to the existing networks can be represented in form of graphs, as shown in Figure 2.2.

2.2.3.1 Desiderata for Graph Neural Networks

We provide a list of requisites that a GNN should meet to effectively learn on graph-structured data.

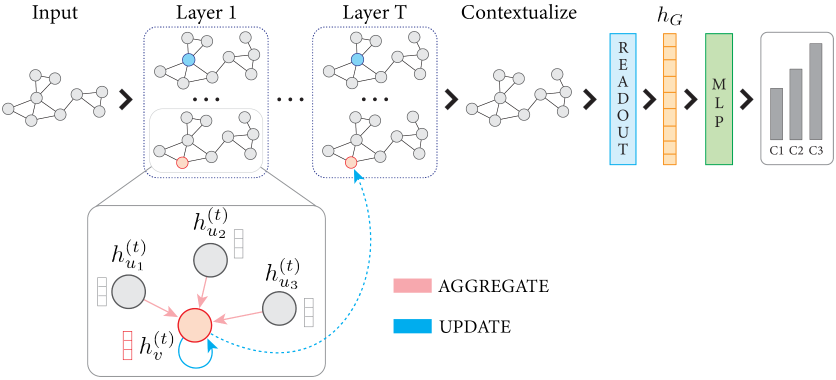


Figure 2.3: An overview of a Message Passing Neural Network. The AGGREGATE and UPDATE steps for node v are illustrated in the zoom.

- **Permutation invariance:** The node ordering of a graph is arbitrary. Re-ordering the nodes does not change the graph. Therefore, a GNN should be invariant to node permutations and produce the same embedding for all such permutations.
- **Scalable and adaptive:** A GNN should be scalable to an arbitrary large input graph, with arbitrary number of nodes, edges, node-, and edge-attributes. Moreover, all the graphs (as defined in Section 2.2.1), i. e., with and without directed edges, with and without node- and edge-attributes, should be able to be encoded by the same type of models, i. e., only minor architectural changes should be needed to adapt to different graph types. Also, no prior knowledge beyond the mathematical description of the graph should be required to train the GNN, i. e., the network should remain application-agnostic.
- **Locality principle:** A GNN should follow a locality principle which states that nearby neighboring nodes and edges share more information than distant neighbors. Intuitively, a GNN should aggregate information from local topological patterns, similar to the concept of convolution in image representation learning. To build arbitrary deep networks, a GNN should be composed of layers that can be stacked, thus increasing the receptive field of the model.
- **Encode graph properties:** A GNN should leverage all the information encoded in a graph, i. e., the graph adjacency, that encodes the graph topology, and the node- and edge-attributes. All information should be jointly encoded by a single GNN.

2.2.3.2 Message Passing Neural Networks

A MPNN [Gilmer et al., 2017] is a type of GNN that follows the message passing paradigm. It is designed to operate on attributed graphs and contextualize the node- and edge-attributes. In relevance to this thesis, we only describe message-passing GNNs among other types of GNNs. Further, we only describe the operations on node-attributed graphs. As we demonstrate MPNN, we put it in relation with the aforementioned list of desiderata. An overview of MPNN framework is illustrated in Figure 2.3.

The node attributes $h(v)$, $v \in V$ are iteratively updated in two phases, i. e., AGGREGATE and UPDATE steps. In the AGGREGATE step for node v, the attributes of neighboring nodes $\mathcal{N}(v)$ are aggregated into a single feature representation $a(v)$. In order to be invariant to node permutation (*Desideratum 1*), the AGGREGATE step is chosen to be a permutation invariant function, e. g., sum, and mean. In the UPDATE step, the attributes of node v are updated by using the current node attributes of v, i. e., $h(v)$, and the aggregated feature representation $a(v)$. Typically, the UPDATE step is a trainable feed-forward neural network. This step is building local (*Desideratum 3*) representations by jointly encoding the graph topology and attributes (*Desideratum 4*). A series of T such iterations, denoted as T GNN layers, are stacked to obtain updated node attributes $\forall v \in V$, incorporating information up to T-hops from each node. Therefore, increasing the number of layers increases the receptive field of the network (*Desideratum 3*), which is analogous to CNNs. Finally, we build a fix-sized graph-level embedding, denoted as h_G, by pooling the node attributes $h^{(T)}(v)$ in a READOUT step. The READOUT step is only employed for graph classification tasks, where a graph embedding is required. Similar to the AGGREGATE step, the READOUT needs to be permutation invariant. This ensures the algorithm to provide graph embeddings of the same dimension, irrespective of the graph size (*Desideratum 2*). To allow back-propagation and GNN training, the AGGREGATE, UPDATE, and READOUT operations must be differentiable. Formally, the three steps are presented as

$$\begin{aligned} a^{(t+1)}(v) &= \textsc{Aggregate}(\{h^{(t)}(u) : u \in \mathcal{N}(v)\}) \\ h^{(t+1)}(v) &= \textsc{Update}(h^{(t)}(v), a^{(t+1)}(v)) \\ h_G &= \textsc{Readout}(\{h^{(T)}(v) : v \in V\}) \end{aligned} \tag{2.1}$$

where, $t = 0, \ldots, T$ denotes the stacked GNN layers.

Following the MPNN framework, different GNN architectures are designed by varying the AGGREGATE, UPDATE, and READOUT operations. A simple message-passing GNN can be designed by using a sum operator for AGGREGATE and READOUT, and a shallow MLP for UPDATE. It can be expressed as

$$h^{(t)}(v) = \sigma\Big(h^{(t-1)}(v) + \sum_{u \in \mathcal{N}(v)} h^{(t-1)}(v)\Big)\, W^{(t)} \tag{2.2}$$

where σ is the ReLU activation function, $W^{(t)} \in \mathbb{R}^{d^{(t)} \times d^{(t+1)}}$ are trainable weights, and $d^{(t)}$ and $d^{(t+1)}$ are the node embedding dimensions at layers t and $t+1$, respectively.

2.2.3.3 Expressivity of Message Passing Neural Networks

An important aspect of designing a GNN is the characterization of its *expressivity*. The expressive power is measured by the GNN's ability to map non-isomorphic graphs to unique graph embeddings, which denotes an injective mapping between the graph- and the embedding-space. Powerful GNNs are expected to be expressive as they can encode non-isomorphic input graphs to distinct locations in the embedding space. A line-of-research exploring the expressive power of GNNs ([Morris et al., 2018; Xu et al., 2019b]) have independently proven the connection between iterative message passing steps of a MPNN and the popular Weisfeiler-Lehman (WL) ([Weisfeiler et al., 1968]) test for graph isomorphism. These results are similar, in spirit, to the Universal Approximation Theorem for neural networks [Cybenko, 1989]. In practice, it means that there exist MPNNs which can learn unique representations for (almost) all undirected node-attributed graphs. Further, the design choices of AGGREGATE, UPDATE, and READOUT operations in a MPNN determine its *expressivity*.

It is established that MPNN architectures such as GIN [Xu et al., 2019b] can perform as well as the 1-dimensional WL test for *discrete* node attribute spaces. However, for graphs with *continuous* node attributes, e. g., CNN-based attributes, the use of multiple permutation-invariant aggregators, e. g., sum, max and mean, can build more expressive GNNs ([Dehmamy et al., 2019; Corso et al., 2020]). To this end, [Corso et al., 2020] proposed PNA by using a combination of aggregators with *degree-scalers*. The series of aggregators replace the sum operation in GIN, and the degree-scalers scale neighboring aggregated-messages according to the node degree.

Specifically, the GIN node update function for node $v \in V$ is defined as

$$h^{(t+1)}(v) = \mathrm{MLP}\Big((1+\epsilon^{(t)})h^{(t)}(v) + \sum_{u \in \mathcal{N}(v)} h^{(t)}(u)\Big) \tag{2.3}$$

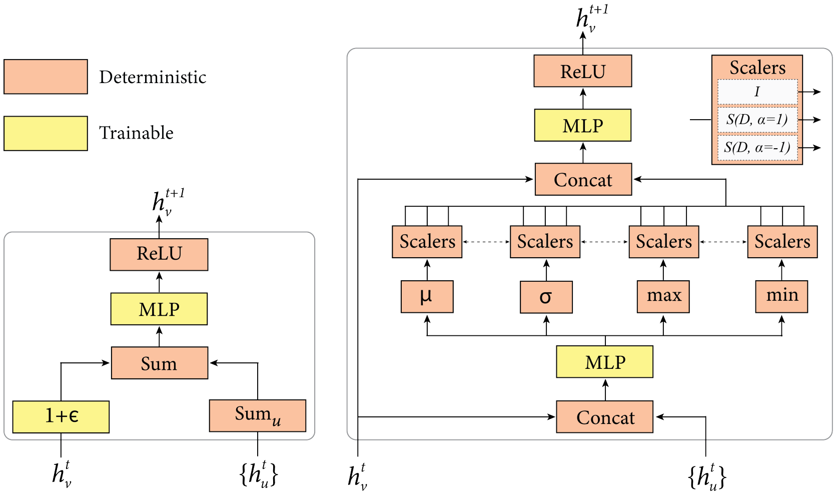
where $\epsilon^{(t)}$ is an optional trainable parameter. The GIN architecture is illustrated in Figure 2.4(a). Similarly, the PNA node update function for node $v \in V$ is defined as

$$\begin{aligned} a^{(t+1)}(v) &= \bigoplus_{u \in \mathcal{N}(v)} M^{(t)}\Big(h^{(t)}(v), h^{(t)}(u)\Big) \\ h^{(t+1)}(v) &= U^{(t)}\left(h^{(t)}(v), a^{(t+1)}(v)\right) \end{aligned} \tag{2.4}$$

As shown in Figure 2.4(b), for a node v, first, the set of neighboring node features $\{h^{(t)}(u)\}, \forall u \in \mathcal{N}(v)$ is concatenated with $h^{(t)}(v)$, and processed by M^t, a MLP, to produce a set of neighborhood-aware features. Then, multiple aggregators with degree-scalers denoted by $\bigoplus$ operate on the set of MLP features to extract a set of multivariate information, which expresses the neighborhood distribution of node v. Finally, the set of information is concatenated to produce the aggregated message $a^{(t+1)}(v)$ for node v. Afterwards, $a^{(t+1)}(v)$ and $h^{(t)}(v)$ are concatenated and processed by U^t, a MLP, to update the node embedding, i. e., $h^{(t+1)}(v)$. Details of $\bigoplus$ is presented as

$$\begin{aligned} \bigoplus &= \Big[I, \mathcal{S}(D, \alpha=1), \mathcal{S}(D, \alpha=-1)\Big] \bigotimes \Big[\mu, \sigma, \max, \min\Big] \\ \mathcal{S}(D,\, \alpha) &= \frac{\log\,(D+1)^{\alpha}}{\delta}, \quad \delta = \frac{1}{|V_{train}|} \sum_{i \in V_{train}} \log\,(d_i+1) \end{aligned} \tag{2.5}$$

where I is identity matrix, S is degree-scaler matrix, D is node degree matrix, δ is normalization constant, α is scaling variable, and V_{train} is nodes in the training dataset. $[I, \mathcal{S}(D, \alpha=1), \mathcal{S}(D, \alpha=-1)]$ and $\Big[\mu, \sigma, \max, \min\Big]$ denote the list of scalers and the list of aggregators, respectively. The aggregators compute statistics on neighboring multiset of nodes, and the injective scalers discriminate between the multisets of various



(a) Graph Isomorphism Network (GIN) (b) Principal Neigborhood Aggregation (PNA)

Figure 2.4: An overview of GIN and PNA layers. h_v^t and $\{h_u^t\}$ denote the representations of node v and its neighbors at layer t, respectively. h_v^{t+1} denotes the representation of node v at layer $t+1$.

sizes. $\alpha = \{-1, 0, 1\}$ for each of these three settings respectively controls the reduction, no scaling, or amplification of the scaling. $\bigotimes$ denotes tensor product between scalers and aggregators, and produces twelve operations that extract the set of multivariate information.

2.2.4 Interpretability of Graph Neural Networks

Though the development of GNNs has primarily focused on improving task performance on graph-structured data, interpretability of GNNs still remains an open research question. Analogous to other classes of neural networks, e. g., CNN, and RNN, GNNs are "black-box" networks, where the process leading to a prediction is too complex for being understood by humans. This lack of transparency can hinder the adoption of GNNs in real-life applications, especially for applications that demand explainable and reliable predictions. In this section, we first present a set of desiderata for deep graph explanations, and then provide an overview of existing post-hoc graph interpretability techniques (or explainers) that produce the graph explanations.

2.2.4.1 Desiderata for graph explanations

The goal of an explainer is to identify a subset of nodes, edges, and node attributes that are important towards a GNN's prediction for a certain task. The identified subset is denoted as the *explanation* for the prediction. An explanation is considered to be "good", if it matches the task-specific prior knowledge. Specifically, there are four requisites for building a graph explainer:

- **Fidelity:** The prediction by the GNN for using the explanation and the original graph should be consistent.
- **Sparsity:** The explanation should be as small as possible, i. e., the explainer should identify the most relevant subset of graph components with high fidelity.
- **Stability:** The explainer should produce similar explanations for similar input graphs, i. e., a small perturbation to an input graph should marginally affect the output explanation.
- **Accuracy:** The explanation needs to align with corresponding ground truth explanation. However, for majority of real-world tasks, ground truth explanations are neither accessible nor uniquely defined, i. e., multiple convincing explanations can exist for an observation. Therefore, this requirement can be relaxed by stating that an explanation is required to align with the experts' understanding of the task.

2.2.4.2 Taxonomy of deep graph explainers

Graphs can be interpreted at instance- and model-level. For *instance-level* interpretability, a graph explainer identifies important input objects for a query graph. Differently, *model-level* interpretability aspires to extract representative graph patterns that drive certain behaviors. In this work, we focus on instance-level methods, which can be further categorized into the following four groups:

- **Gradient-based methods** define node importance by measuring the gradient of an output class, e. g., the predicted class, with respect to input graph components. A positive or high gradient for a component denotes a positive relevance, whereas a negative or low gradient for a component denotes a negative or less relevance of the component on the prediction [Baldassarre et al., 2019; Pope et al., 2019]. Example methods are GraphGrad-CAM and GraphGrad-CAM++.
- **Perturbation-based methods** study the influence of small input perturbations on the output. Intuitively, the removal of discriminative graph components should change the model prediction, whereas the removal of uninformative graph components should not impact the prediction. By characterizing these changes, instance-level explanations are proposed by [Ying et al., 2019; Luo et al., 2020; Yuan et al., 2020b; De Cao et al., 2020].
- **Decomposition-based methods** decompose the original model predictions from the predicted logits, and backpropagates the logits to the input features to understand the relationship between the input-space and the logit-space [Baldassarre et al., 2019; Pope et al., 2019; Schwarzenberg et al., 2019]. An example method is GraphLRP that backpropagates the output logits in a layerwise manner by following certain propagation rules.
- **Surrogate methods** aim to explain a complex model prediction via a simple and interpretable surrogate model, e. g., a linear model [Huang et al., 2020] or a probabilistic graphical model [Vu et al., 2020]. Specifically, these methods first generate perturbed graphs around a query graph, and then approximate the original model's predictions on the perturbed graphs using a surrogate model.

The reader can refer to [Yuan et al., 2020a] for a thorough and detailed review on deep graph interpretability.

2.2.4.3 Deep graph explainers

In this section, we formally present four post-hoc graph explainers, namely, GraphLRP, GraphGrad-CAM, GraphGrad-CAM++, and GnnExplainer, studied in this thesis. These methods form the theoretical foundations of Chapter 6, where (1) we demonstrate the potential of deep graph explainers for explaining predictions from histopathology images, and (2) we emphasize the importance of developing quantitative metrics for measuring the quality of graph explanations in the absence of ground truth explanations.

☐ **Notations**

We define an attributed graph $G := (V, E, H)$ as a set of nodes V, edges E, and node attributes $H \in \mathbb{R}^{|V| \times d}$. d denotes the number of attributes per node, and $|.|$ denotes set cardinality. The graph topology is defined by an adjacency matrix, $A \in \mathbb{R}^{|V| \times |V|}$, where $A_{uv=1}$ if $(u, v) \in E$. $H_{n,k}$ expresses the k-th attribute of the n-th node. The forward prediction of a graph G is denoted as, $y = \mathcal{M}(G)$, where $\mathcal{M}$ is a GNN operating on the graph, and $y \in \mathbb{R}^{|\mathcal{T}|}$ are output logits. Notation $y(t), t \in \mathcal{T}$ denotes the output logit of the t-th class. We refer to the logit of the predicted class as $y_{\max} = \max_{t \in \mathcal{T}} y(t)$, and the predicted class as $t_{\max} = \text{argmax}_{t \in \mathcal{T}} y(t)$.

☐ **Graph explainer setting**

All the graph explainers operate in a similar setting, described as follows:

- Input is an attributed graph G.
- GNN model $\mathcal{M}$ is trained a priori and can be used for inference. Note that different graph learning models could also be combined with the presented graph explainers, but this is beyond the scope of this work.
- Explanations are always generated by explaining one output logit, e. g., predicted class $y_{\max}$. But it can also be generated by explaining any output query logit.
- Each explainer returns normalized node-level importance scores that characterizes the relevance of each node for predicting a certain class, e. g., for classifying $t_{\max}$.
- Node importance scores can be thresholded to retain the most relevant subset, defined as the explanation $G_s = (V_s, E_s, H_s) \subset G$. The explanation graph topology is derived by keeping all the edges connected to the subset of identified nodes, i. e., $E_s = (u, v) | u, v \in V_s, e_{uv} \in E$.

☐ **Graph Layerwise Relevance Propagation** (GraphLRP)

Layerwise Relevance Propagation (LRP) [Bach et al., 2015] is a decomposition-based method. LRP explains an output logit by decomposing the individual contributions of each input element, i. e., each node, to the query logit value. An output logit, defined as the output relevance for a given class, is layerwise back-propagated until the input to compute the positive or negative impact of the input elements on the output logit. LRP was initially formulated for operating on fully connected layers (LRP-FC), and works as follows. Given a pre-trained fully connected layer $W \in \mathbb{R}^{z_1 \times z_2}$ between layer 1 and layer 2, where z_1 and z_2 are the number of neurons in layer 1 and layer 2, respectively, we compute the contributions of a neuron i , $i \in \{1, ..., z1\}$ using propagation rules introduced in [Montavon et al., 2015]. In this work, we are interested in identifying input elements positively contributing to the prediction. To this end, we use the z^+ propagation rule that back-propagates the positive neuron contribution from layer 2 to layer 1 as:

$$R_i = \sum_j^{z_2} \frac{f_i |w_{ij}|}{\sum_k^{z_1} f_k |w_{kj}|} R_j \tag{2.6}$$

where $|w_{ij}|$ is the absolute value of the weight between i-th and j-th neuron in layer 1 and 2, respectively. f_i denotes the activation of the i-th neuron in layer l.

The extension from LRP-FC to LRP for GIN layers (GraphLRP) is achieved by following the observations in [Schwarzenberg et al., 2019]. First, the *aggregate step* in GNN corresponds to projecting the graph's adjacency matrix on the node attribute space. For simplicity, assuming a 1-layer Multi-layer Perceptron (MLP) as an update function, the GIN layer with *mean* aggregator can be re-written in its global form as:

$$H^{(l+1)} = \sigma\Big(W^{(l)}(I + \tilde{A})H^{(l)}\Big) \tag{2.7}$$

where $\tilde{A}$ is the degree-normalized graph adjacency matrix, i. e., $\tilde{A}_{ij} = \frac{1}{|\mathcal{N}(i)|} A_{ij}$. σ is the ReLU activation. Second, this representation allows to treat the term $(I + \tilde{A})$ as a regular, fully connected layer. We can apply the z^+ propagation rule with weights w_{ij} defined as:

$$w_{ij} = 1 \quad \text{if } i = j \tag{2.8}$$

$$w_{ij} = \frac{1}{|\mathcal{N}(i)|} \quad \text{if } e_{ij} \in E \tag{2.9}$$

$$w_{ij} = 0 \quad \text{otherwise} \tag{2.10}$$

GraphLRP outputs an importance score for each node i in the input graph.

☐ **Graph Gradient-weighted Class Activation Mapping** (GraphGrad-CAM)

Grad-CAM [Selvaraju et al., 2017] is a feature attribution post-hoc explainer that identifies salient regions of the input that drives a neural network prediction. It assigns importance to each element of the input to produce a Class Activation Map [Zhou et al., 2016]. While originally developed for explaining CNNs operating on images, Grad-CAM can be extended as GraphGrad-CAM to GNNs operating on graphs [Pope et al., 2019].

GraphGrad-CAM processes in two steps. First, it assigns an importance score to each channel of a graph convolutional layer. The importance of channel k in layer t is computed by looking at the gradient of the predicted output logit $y_{\max}$ with respect to the node attributes at layer t of the GNN. Formally it is expressed as:

$$w_k^{(t)} = \frac{1}{|V|} \sum_{n=1}^{|V|} \frac{\partial y_{\max}}{\partial H_{n,k}^{(t)}} \tag{2.11}$$

Intuitively, large positive gradients are evidences of the presence of the class under consideration, while small gradients do not confirm the presence. In the second step, a node-wise importance score is computed using forward node feature activations $H^{(t)}$ as:

$$L(t, v) = \mathrm{ReLU}\Big(\sum_{k}^{d^{(t)}} w_k^{(t)} H_{n,k}^{(t)}\Big) \tag{2.12}$$

where $L(t, v)$ denotes the importance of node $v \in V$ in layer t , and $d(t)$ denotes the number of node attributes in layer t. As we are interested in the positive node contributions, i. e., nodes that positively influence the class prediction, we apply a ReLU activation to the node importance scores. Following the prior work by [Pope et al., 2019], we take the average scores obtained over all the GNN layers to obtain smooth representations, i. e.,

$$L(v) = \frac{1}{T} \sum_{t \in \{1,\ldots,T\}} L(t, v), \forall v \in V \tag{2.13}$$

Node-level scores are thresholded to identify the most important subset of nodes.

☐ **Graph Gradient-weighted Class Activation Mapping++** (GraphGrad-CAM++)

GraphGrad-CAM++ extends GRAD-CAM++ [Chattopadhay et al., 2018] to graph-structured data. It improves the node importance localization of GRAD-CAM by introducing node-wise contributions to the channel importance score computation. It builds on the work by Zhou et al. (2016), that empirically proved to have localization properties. Specifically, Equation 2.11 is modified as:

$$w_k^{(t)} = \frac{1}{|V|} \sum_{n=1}^{|V|} \alpha_{n,k}^{(t)} \frac{\partial y_{max}}{\partial H_{n,k}^{(t)}} \tag{2.14}$$

where $\alpha_{n,k}^{(t)}$ are node-wise weights expressed for each attribute k at layer t. The closed-form solution for $\alpha_{n,k}^{(t)}$ is analogous to the derivation in [Chattopadhay et al., 2018], where the graph size, i. e., number of nodes, replaces the spatial dimensions of a channel as:

$$\alpha_{n,k}^{(t)} = \frac{\frac{\partial^2 y_{max}}{(\partial H_{n,k}^{(t)})^2}}{2\frac{\partial^2 y_{max}}{(\partial H_{n,k}^{(t)})^2} + \sum_{n=1}^{|V|} H_{n,k}^{(t)} \big(\frac{\partial^3 y_{max}}{(\partial H_{n,k}^{(t)})^3}\big)} \tag{2.15}$$

The subsequent node importance computation in GraphGrad-CAM++ is similar to GraphGrad-CAM, i. e., use Equation 2.12 to derive $L(t, v)$, and Equation 2.13 to get $L(v)$.

☐ **Perturbation-based** GnnExplainer

GnnExplainer [Ying et al., 2019] is based on graph pruning. It is model-agnostic and can explain any flavor of GNN. Intuitively, GnnExplainer tries to find the minimum sub-graph $G_s \subset G$, i. e., the minimum set of nodes and edges, while retaining the model prediction. It enforces explanation sparsity while ensuring high explanation fidelity. The inferred sub-graph G_s is regarded as the explanation for the graph G.

Formally, the sub-graph $G_s = (V_s, E_s, H_s) \subset G$ is created such that the mutual information (MI) between the original prediction $y_{\max}$ and the sub-graph G_s is maximized, i. e.,

$$\max_{G_s} MI(\hat{y}, G_s) = \mathcal{H}(\hat{Y}) - \mathcal{H}(\hat{Y}|G = G_s) \tag{2.16}$$

which is equivalent to minimizing the conditional entropy,

$$\min_{G_s} \mathcal{H}(\hat{Y}|G = G_s) = -\mathbb{E}_{\hat{Y}|G_s}[\log(P_{\mathcal{M}})(\hat{Y}|G_s)] \tag{2.17}$$

Intuitively, the sub-graph G_s is extracted which maximizes the probability of $y_{\max}$. Exhaustively searching G_s in the space created by nodes V and edges E is infeasible due to the combinatorial nature of the task. Instead, GNNEXPLAINER formulates the task as an optimization problem that learns a mask to activate or deactivate parts of the graph. In this regard, this approach can be seen as a feature attribution method with binarized node and edge importance scores, i. e., a node $v \in V$, edge $e \in E$, has importance one if $v \in V_s$, $e \in E_s$, and zero otherwise.

The formulation by [Ying et al., 2019] is developed for explaining *node classifiers*, where the aim is to explain the classification prediction of a query node. Specifically, a mask $M_E \in \mathbb{R}^{|V| \times |V|}$ is learned over the edges, i. e., over the adjacency matrix A. Masking edges will cut connections between the query node and its neighbors. Formally, the mask is searched such that,

$$\min_{M_E} - \sum_{c=1}^{C} \mathbb{1}_{[y=c]} \log(P_{\mathcal{M}}(\hat{Y}|G = A \odot \sigma(M_E), H)) \tag{2.18}$$

where C denotes the number of classes, σ is the sigmoid activation, and $\odot$ denotes element wise multiplication. Heuristically, these constraints can be enforced by minimizing,

$$\mathcal{L} = \mathcal{L}_{\text{KD}}(y_{\max}, y^{(t)}) + \alpha_{M_E} \sum_{i}^{|E|} \sigma(M_{E_i}^{(t)}) + \alpha_{\mathcal{H}} \mathcal{H}^e(\sigma(M_E^{(t)})) \tag{2.19}$$

where, t denotes the optimization step. The first term is a knowledge-distillation loss L_{KD} between the new logits $y^{(t)}$ and the original prediction $y_{\max}$, to preserve explainer fidelity. The second term enforces explainer sparsity by minimizing the mask size M_E. The third term binarizes M_E by minimizing its element-wise entropy $\mathcal{H}^e$. Following [Hinton et al., 2015], L_{KD} is defined as a combination of distillation and cross-entropy loss,

$$\mathcal{L}_{\text{KD}} = \lambda \mathcal{L}_{\text{CE}} + (1 - \lambda) \mathcal{L}_{\text{DIST}} \text{ where } \lambda = \frac{\mathcal{H}^e(y^{(t)})}{\mathcal{H}^e(\hat{y})} \tag{2.20}$$

As the element-wise entropy $\mathcal{H}^e(y^{(t)})$ increases, L_{CE} gains importance and avoids a change in predicted label. M_E, produced by optimizing Equation 2.19, is learned with iterative gradient descent until convergence. Note that, the original formulation can be extended to prune features along the node dimension as well. As this extra step is not relevant for the proposed downstream tasks, we let the reader refer to Section 2.1 in [Ying et al., 2019] for an in-depth formulation.

3

High-Quality Immunohistochemical Stains through Computational Assay Parameter Optimization

Accurate profiling of tumors using immunohistochemistry (IHC) is essential in cancer diagnosis. The inferences drawn from IHC-stained images depend to a great extent on the quality of immunostaining, which is in turn affected strongly by assay parameters. To optimize assay parameters, the available tissue sample is often limited. Moreover, with current practices in pathology, exploring the entire assay parameter space is not feasible. Thus, the evaluation of IHC stained slides is conventionally a subjective task, in which diagnoses are commonly drawn on images that are suboptimal. In this work, we introduce a framework to analyze IHC staining quality and its sensitivity to process parameters. To that extent, first histopathological sections are segmented into diagnostically relevant and contextually immaterial signals histological entities. Then, machine learning techniques based on the histological entities are employed to extract disease-specific staining quality metrics (SQMs) targeting a quantitative assessment of staining quality. Lastly, an approach to efficiently analyze the parameter space is introduced to infer sensitivity to process parameters. We present results on microscale IHC tissue samples of five breast tumor classes, based on disease state and protein expression. A disease-type classification F1-score of 0.82 and a contrast-level classification F1-score of 0.95 were achieved. With the proposed SQMs an area under the curve of 0.85 was achieved on average over different disease types. Our methodology provides a promising step in automatically evaluating and quantifying staining quality of IHC stained tissue sections, and it can potentially standardize immunostaining across diagnostic laboratories.

3.1 Introduction

Malignancies are often studied and detected by acquiring a protein expression profile on a tissue section. Such a protein expression map on a tissue is obtained by immunohistochemical (IHC) staining thereby generating a visual signal while retaining the tissue structure of tissues (histology). IHC has been an invaluable tool in the field of both cancer

---

This chapter has been published as: Nuri Murat Arar[†], Pushpak Pati[†], Aditya Kashyap, Anna Fomitcheva Khartchenko, Orcun Goksel, Govind V. Kaigala, and Maria Gabrani, "High-Quality Immunohistochemical Stains through Computational Assay Parameter Optimization", In: IEEE Transactions on Biomedical Engineering, vol. 66, no. 10, pp. 2952-2963, 2019, doi: 10.1109/TBME.2019.2899156

diagnostics and research, owing to a rapidly obtainable snapshot of status of cells within tissue samples. In this paper, we focus on a new methodology for realizing high-quality immunostaining both at the micrometer-length scale and for conventional whole-tissue staining for tumor stratification.

IHC is implemented by exposing a tissue to antibodies which bind to a specific protein, thus identifying prognostic and treatment-related biomarkers. The commonly used IHC protocol is a multi-step and multi-parametric process [Taylor, 2000] and involves binding of a primary antibody specific to a protein of interest on the tissue, followed by a secondary antibody that binds to the primary. The colored signal on the tissue is obtained using a chromogenic moiety coupled with the secondary antibody, where the chromatic signal strength is a function of the density of the proteins of interest in the tissue, their accessibility, and the concentration of the antibody that is exposed to the antigen, among several other parameters. The IHC signal can provide vital information in a diagnosis workflow. However, when the process parameters are not optimal, this may lead to difficult-to-interpret images and potential misdiagnosis, e.g., false positive and false negative staining as demonstrated in Figure 3.1.

Although IHC has been used now for decades, standardization and reproducibility remain two major concerns. Pathology laboratories manually determine the parameters leading to a good staining quality. Such a manual process comprising trial-and-error is cumbersome and tissue exhaustive. Besides, it is characterized by high inter- and intra-laboratory variability, leading to poor reproducibility. Nordic Immunohistochemical Quality Control (NordiQC), an international external quality-assurance organization, found that about 20% of the staining results in a breast-cancer IHC cross-lab examination were insufficient for diagnostic use [Vyberg et al., 2016]. Inaccurate and/or equivocal results are mostly obtained because of inappropriate parameters used in the staining process (protocol), less specific antibodies, insufficiently calibrated antibody dilutions, variable fixation processes and erroneous epitope retrieval methods. To improve the standardization in immunostaining, efforts have been made by ad-hoc committees on pathology [Goldstein et al., 2007; Yaziji et al., 2008; Wolff et al., 2007; Wolff et al., 2014; Torlakovic et al., 2014; Torlakovic et al., 2015], by external quality-assurance schemes [Von Wasielewski et al., 2008; Copete et al., 2011; Howat et al., 2014; O'Hurley et al., 2014], and by field researchers [Pinard et al., 2012; Grunkin et al., 2019] through addressing one or more of the factors affecting the staining results. The effect of specific process parameters on the quality is hard to deconvolve owing to limited tools that allow for the scanning of a range of process parameters on the same tissue. Thus, strategies that perform automated analysis of process parameter sensitivity and contextual quantitative analysis are crucial in improving the IHC standardization, and thus reproducibility.

More recently, the advent of digital pathology has prioritized the extraction of quantitative information from scanned histopathological sections to aid pathologists in the diagnostic process, while attempting to reduce or eliminate observer biases [Masmoudi et al., 2009; Rizzardi et al., 2012]. Furthermore, computational pathology aims at automating the analysis of stained sections, as manually analyzing numerous biopsy slides can be tedious and labor intensive. Recent advances enabled the automated recognition of pathological patterns, which has the potential to provide valuable assistance to a pathologist. There exist several studies which demonstrate the agreement between digital image analysis-based methods and pathologists'visual examination. For instance, Dobson et al. [Dobson

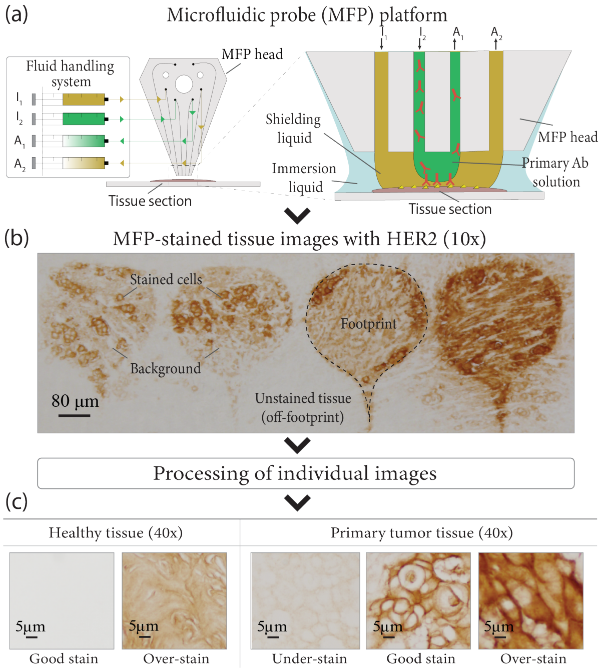


Figure 3.1: Immunostaining process and variability in staining quality due to process parameter variations in IHC. **(a)** illustration of microfluidic probe platform for microscale IHC, **(b)** sample HER2-stained tissue images using an MFP, **(c)** staining quality variability on healthy and primary tumor tissues. For HER2 non-expressing healthy tissue, low and high HER2 expression indicates high-quality staining (true negative) and over-staining (false positive), respectively. For HER2 overexpressing primary tumor tissue, low HER2 expression indicates under-staining (false negative). HER2 expression solely on membranes implies high-quality staining (true positive), whereas expression in cytoplasm and stroma indicates over-staining (false positive).

et al., 2009] and Brugmann et al. [Brügmann et al., 2012] demonstrated that HER2 antibody protein expression can be classified with a high accuracy by analyzing the staining intensity and membrane connectivity on IHC images with optimal staining quality. Differently from the previous work, this work deals with IHC-stained tissue images with both optimal and sub-optimal staining quality. The combination of such quantification-aided diagnosis with quantified grading has the potential to improve diagnostic accuracy.

Limited prior work exists on the quantitative analysis of the immunostaining quality. Pinard et al. [Pinard et al., 2012] proposed a system that extracts quantitative quality indicators and compares them with the respective user-defined minimum acceptable quality thresholds. Failure of one or more of the indicators to meet its respective threshold suggests that the sample is unsuitable for a subsequent automated pathological evaluation. Similarly, Grunkin and Hansen [Grunkin et al., 2019] described a method for assessing the staining quality of specimens in a working laboratory. Their system compares the quality parameters, e.g., staining intensity, connectivity, number of cells, Allred-score, Nottingham index, obtained from a reference specimen prepared at a standardized laboratory according to a predetermined staining protocol with the quality parameters obtained from a specimen prepared at the working laboratory. The relative quality measure is computed using a distance metric between the quality parameters of the test and reference specimen. Both studies output relative quality estimates with respect to either a user-defined threshold or a reference specimen, which limits the standardization of the process and thus the reproducibility. Instead, we propose to use reference standards for quality labeling during the training phase and use the trained quality metrics during the testing phase, thereby removing the need of posterior standards. The proposed methodology does not completely remove the need of an external standard but reduces the dependency on it on a daily practice. In addition, neither of the aforementioned studies takes into account the diagnostical relevance of the signals on the stained images, which can potentially hamper the computed quality indicators. For an alternate perspective, our automated methodology first segments the diagnostically relevant and the contextually immaterial signals in an IHC-stained image, followed by machine learning models for estimating the quality indicators.

Addressing IHC assay limitations requires technologies that enable precise control of the various steps of the assay, including the ability to create multiple assay conditions on the same tissue section. Here we use a microfluidic probe (MFP) [Kaigala et al., 2011], a scanning microfluidic device that localizes nanoliter volumes of antibodies on micrometer scale areas of tissue sections. By leveraging the ability of the MFP to perform multiple microscale IHC tests on the same tissue section [Lovchik et al., 2012; Taylor et al., 2016], we not only can perform experimental parameter optimization of IHC by exposing adjacent areas on a sample to different experimental conditions (antibody concentration, incubation time), but can also be conservative of the tissue sample.

In this work, we introduce a complete methodology to quantify and analyze the staining quality and its sensitivity to IHC process parameters using well-established image processing and machine learning techniques. The proposed methodology first extracts quantitative information from scanned histopathological sections using an automated diagnostically relevant signal segmentation algorithm. It then learns multiple metrics for the quantitative assessment of the staining quality. Lastly, it performs an analysis of the sensitivity of staining quality to process parameters for the optimal parameter-space determination. Preliminary results of this work were presented in [Arar et al., 2017]. These have been extended herein with improvements on the methodology and validation. First, we refined our framework in order to account for different disease types. To achieve this, we conducted a comprehensive analysis of the impact of various image representation and classification-related parameters of the framework. We additionally explored alternative feature extraction and classification techniques, including deep learning strategies. We provide herein a comprehensive validation on a cohort of annotated breast cancer tissues from five different disease types. Moreover, we compared the proposed approach against the current clinical staining approach and demonstrate the superiority of the proposed staining approach.

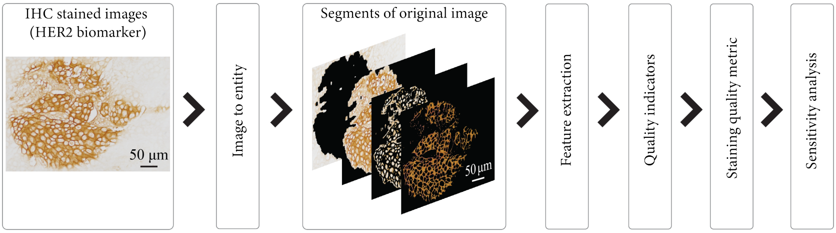


Figure 3.2: Overview of immuno-staining quality assessment methodology. Images are segmented to extract different levels of information, which is used for generating various staining quality indicators. These are fed into a machine learning algorithm to learn multiple staining quality metrics. Lastly, an analysis of the quality sensitivity to the staining process parameters is performed for the identification of process parameter space resulting in optimal staining quality.

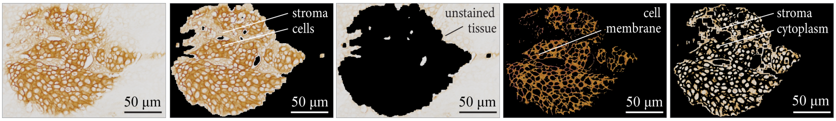
3.2 Methodology

The proposed methodology for staining quality and sensitivity assessment has 4 main components: a) separation of diagnostically relevant and contextually immaterial signals, b) staining quality metric learning, c) image and quality representation, and d) sensitivity analysis to staining parameters. An overview of the methodology is shown in Figure 3.2.

3.2.1 Diagnostically relevant signal segmentation

Staining quality is directly proportional to the diagnostically relevant signal, i. e., the staining on interesting cell structures or regions (true positive staining), and is inversely proportional to the contextually immaterial signal, i. e., the staining on the remaining areas (false positive staining). An optimal staining quality is achieved when the ratio of the relevant signal to the immaterial signal is the highest. Therefore, our methodology essentially focuses on a good delineation of the two signals in IHC-stained tissue images prior to further analysis. Note that the definition of the two signals may vary depending on the choice of the biomarker in the staining process, as each biomarker binds to a specific antigen present in stipulated cell structures. For instance, HER2 biomarker binds to the HER2 antigen in the cell membrane; developing diagnostically relevant signal on the cell membrane and arising immaterial signal on the remaining cell structures, i. e., cytoplasm and stroma. Whereas, p53 biomarker produces the relevant signal on the nuclei of tumor cells and develops the immaterial signal on cell membrane and stroma.

Our methodology begins with an automatic segmentation algorithm based on a combination of well-known image processing techniques for separating aforementioned two signals in the images of μIHC-stained breast tissue. The algorithm firstly segments an image into two regions as *off-footprint* and *footprint*. The latter is further partitioned into two: *foreground* and *background*, as shown in Figure 3.3 for a HER2-stained tissue. The segmentation process begins with finding and delineating the localized *footprint*, the tissue area where the MFP head is applied. To that end, we first estimate the *footprint* by Otsu binarization, followed by a morphological opening to generate highly confident masks for both the *footprint* and *off-footprint*; with the remaining regions considered as uncertain. Second, the obtained masks are fed into the Watershed algorithm to assign the



(a) (b) (c) (d) (e)

Figure 3.3: Sample outputs of the segmentation algorithm: **(a)** input IHC-stained tissue, **(b)** stained region (*footprint*), **(c)** unstained region (*off-footprint*), **(d)** diagnostically relevant signal, e. g., staining of cell membrane (*foreground*), **(e)** contextually immaterial signal (*background*).

uncertain areas into either *footprint* or *off-footprint* (Figure 3.3(b,c)). Next, the *footprint* is subdivided into the relevant (*foreground*) and immaterial (*background*) regions as shown in Figure 3.3(d,e). Considering the intensity distribution difference between two regions, global thresholding with a robust threshold value is sufficient to extract the *foreground*. Here, we set the threshold value as the mean of the most frequent and maximum intensity values within the *footprint* region. To determine the threshold value more robustly, particularly in the presence of experimental or imaging artifacts (often resulting in a significantly high intensity), we calculate a 16-bin intensity histogram of the inverted gray-scale *footprint*, and extract the corresponding values from the histogram bins. Subsequently, we extract the *background*, i.e., false positive stain. Assuming that false positive staining highly occurs around the *foreground*, we subtract the binary *foreground* mask from the dilated *foreground* mask to extract the *background* within a close proximity of the *foreground*. We then ensure the connectivity of the *background* through a morphological closing operation and remove any remnants of the *foreground* pixel. Lastly, we derive the statistics on the amount of true positive and false positive staining within the segmented regions as part of the quality features and for an early assessment of the tissue sufficiency.

3.2.2 Staining quality metric learning

Optimal staining of cell structures reveal the disease type of an IHC-stained tissue, therefore, we consider that the definition of staining quality varies across disease types for a particular tissue type. Figure 3.4 presents high-quality HER2-staining of 5 disease types. HER2 is a transmembrane receptor, thus the quantification of its overexpression can be modelled as detecting 'peaks' (cell membranes) versus 'valleys' (cell cytoplasm and stroma), as also depicted in Figure 3.4. The 'peaks', and 'valleys' model represents the intensity profiles along the cross-sectional view of a cell for each disease type. The model indicates that a HER2+ tumorous tissue exhibits a high contrast, whereas a HER2- tumorous tissue or a healthy tissue exhibits low or no contrast between the 'peaks' and 'valleys'. Considering the variability in staining quality expectations, a unique SQM per disease type must be developed. Note that, previous works on staining quality assessment employed a reference-based staining quality estimation, e.g., [Pinard et al., 2012], [Grunkin et al., 2019]. In contrast, herein we propose a machine learning-based *no-reference* SQM learning method, which enables to assess the staining quality of a tissue without the need of any reference specimen or user-defined quality threshold.

As per our experimental observations across various disease types, we hypothesize that an immunostaining can be of high-quality, a) if it contains sufficient information (signal) to reflect its disease type, and b) if the contrast level between diagnostically relevant and contextually immaterial signals aligns with the expected contrast level for

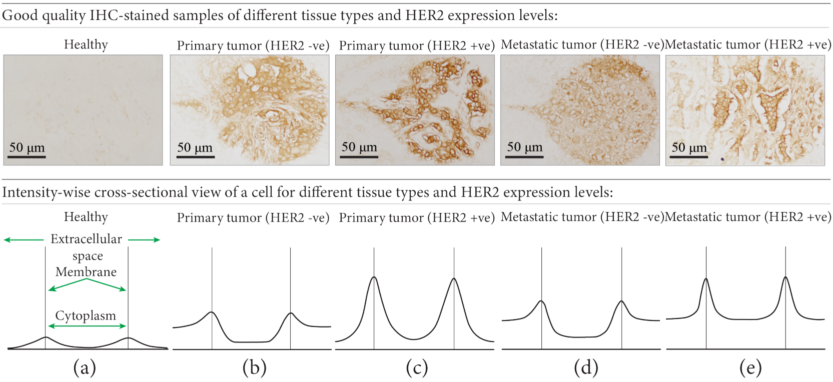

Good quality IHC-stained samples of different tissue types and HER2 expression levels:
Healthy
Primary tumor (HER2 -ve)
Primary tumor (HER2 +ve)
Metastatic tumor (HER2 -ve)
Metastatic tumor (HER2 +ve)
50 µm
Intensity-wise cross-sectional view of a cell for different tissue types and HER2 expression levels:
Extracellular space
Membrane
Cytoplasm
(a)
(b)
(c)
(d)
(e)

Figure 3.4: Variability in immunostaining quality expectations with HER2 antibody for different disease types: (top row) Samples of high-quality immunostaining, and (bottom row) cross-sectional intensity profiles across a cell for each category. Note that for transmembrane HER2 antibody, high expression is expected only on the cell membrane for tumorous tissues.

the corresponding disease type. We develop our quality assessment metrics based on these two quality indicators (Figure 3.5). For an IHC-stained tissue, we first capture the disease type information via a probability map indicating its likelihood of being a certain disease type. Secondly, we acquire the contrast information via another probability map indicating the relevant-to-immaterial signal contrast level irrespective of its disease type. We then learn disease type-specific SQMs based on these two pillars in our proposed staining quality assessment framework. Through further analysis, additional quality indicators may be included to improve the framework.

Disease type quality indicator: Breast tissues can be categorized into 3 types, namely, healthy tissue adjacent to the tumor (HT), primary tumor tissue (PT), and lymph-node metastasis tissue (MT). On staining the tissues with HER2 biomarker, the latter two can present either an overexpression (HER2+) or a weak overexpression (HER2-) based on the aggressiveness of the cancer. Thereby, HER2-stained breast tissues can be categorized into 5 disease types, namely, HT, HER2+ PT (PT+), HER2- PT (PT-), HER2+ MT (MT+) and HER2- MT (MT-). We propose to train a 5-class supervised probabilistic classifier to identify the disease type of an IHC-stained tissue, and capture the first quality indicator.

Contrast level quality indicator: HER2-stained breast tissues exhibits a certain degree of contrast between the cell membranes and, the cytoplasm and extra-cellular space depending on the aggressiveness of cancer, as presented by the ‘peaks’, and ‘valleys’ model in Figure 3.4. To obtain the second quality indicator, we propose to train a binary-class supervised probabilistic classifier to identify the contrast level between the diagnostically relevant membrane and contextually immaterial background.

SQM learning & quality assessment: We propose to learn disease-type specific SQMs considering the unique expectation of staining quality per disease type. The quality indicators acquired for the samples of a particular disease type are used to train an individual SQM. An SQM is learned in a supervised manner using the quality labels for the respective samples obtained from a group of experts. In general, the experts evaluated each sample with various metrics, namely tissue type, antibody expression status, tissue

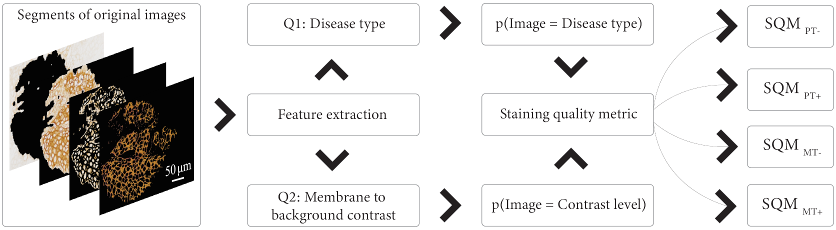


Figure 3.5: Overview of the disease-specific staining quality metric (SQM) learning. It involves learning various quality indicators, which currently rely on the output probability maps of two classifiers such as a disease type classifier and a membrane-to-background contrast classifier.

sufficiency, membrane-to-background contrast etc., to assign a quality label $\in$ {Acceptable, NotAcceptable}. An SQM is a probabilistic classifier that learns to predict the quality labels of images using respective two quality indicators. The $p(image = \text{Acceptable})$ is termed as the quality value (QV) of the image that indicates the acceptability of the image. For instance, a PT+ sample with low membrane-to-background contrast depicts a low-quality staining. Thus, a QV obtained from SQM_{PT+} will have a low value for that sample indicating its low acceptability. In summary, the SQM learning can be defined as:

$$\begin{aligned} QI_1 &= p(image = \text{DiseaseType}_i),\ \text{DiseaseType}_i \in \{\text{HT, PT-, PT+, MT-, MT+}\}, \\ QI_2 &= p(image = \text{ContrastLevel}_j),\ \text{ContrastLevel}_j \in \{\text{High, Low}\}, \\ SQM_i &= p(image = \text{QualityLabel}_k \mid QI_1, QI_2, image \in \text{DiseaseType}_i), \\ &\quad \text{QualityLabel}_k \in \{\text{Acceptable, NotAcceptable}\}, \\ QV_i &= p(image = \text{Acceptable} \mid QI_1, QI_2, SQM_i) \end{aligned} \tag{3.1}$$

3.2.3 Image and quality representation

The supervised probabilistic classifiers for identifying disease type and contrast level in an immunostained tissue are trained using a set of quality relevant features extracted from HER2-stained training images. We propose a comprehensive feature extraction followed by feature selection to obtain a more efficient representation for individual classification task. We experiment with traditional machine learning and deep learning approaches for training the classifiers. The machine learning-based system relies on hand-crafted features, which are shown to be successful in the prior work, whereas the deep architecture is trained with features extracted from a pre-trained network. The individual feature sets are discussed in details in following sections.

Hand-crafted features: Hand-crafted features are extracted both holistically, features from individual segmented regions to capture information about relevant and immaterial signals in the whole image, and locally, patch-wise features from relevant regions to capture local structural and morphological information. Local features are extracted from patches containing a sufficient amount of *foreground*, as segmenting each cell for the analysis is not feasible.

Holistic features: Intensity-based features directly relate to the amounts of relevant and immaterial signal in an image. We extract mean *foreground* intensity (relevant signal strength), mean *footprint* intensity (immaterial signal strength), mean *footprint* intensity from segmented regions, and relative intensity of relevant to immaterial signal. Note that relative intensity feature is used a major quality indicator in [Pinard et al., 2012] and [Grunkin et al., 2019].

Percentile features, namely % of *foreground* in the image, % of *foreground* within the *footprint* and % of *footprint* within the image are included to encode the amount of relevant area in the whole image.

Difference of Gaussians is used to detect keypoints on the *foreground* of an image, and SIFT features [Lowe, 2004] are extracted around the keypoints. We combine the SIFT features using K-means clustering, K decided by Bayesian information criterion, with bag-of-words to define a fixed-dimensional feature representation for the image.

Local features: We extract texture, spatial and frequency domain features in a patch-wise manner. Texture features in terms of contrast and entropy statistics are obtained from Gray-level co-occurrence matrices with two distance values and four orientations. Additional image-gradient-based sharpness features namely, mean gradient magnitude, mean and standard deviation of blur difference, sharpness and Tenengrad response are extracted as suggested in [Lopez et al., 2013].

We acquire morphological and topological clues in the neighborhood of cells using spatial and frequency domain wavelet features at multi-scale resolution. Gabor wavelet based rotation- and scale-invariant features are extracted using complete and non-orthogonal basis set of Gabor filters with eight rotations and five scaling factors, as in [Han et al., 2007]. Discrete Haar wavelet transformation at 3 levels is performed per patch to extract mean, variance, rotation-invariant energy and anisotropy of energy features along horizontal, vertical and diagonal sub-bands, as in [Livens et al., 1996] and [Hu et al., 2014]. Visual perceptual directionality, contrast and coarseness features are extracted using [Jian et al., 2009]. Shift invariant Haralick features are extracted from individual sub-images obtained via Dual tree complex wavelet transform of patches [Yang et al., 2016].

The patch-wise features are extracted from patches with sufficient amount of *foreground*, thereby, making them homogeneous in nature over an image. Hence firstly, we exclude the outlier patches for an image, based on the distance between the per patch feature representation and the mean feature representation, computed using all patches from the image. Secondly, we compute the mean feature representation across all the remaining patches to define the final feature representation for the complete image.

Feature selection: Classification-task specific feature selection is performed on the extracted set of features to remove irrelevant and redundant attributes. We use *recursive feature elimination* with Random forest feature importance to select the optimal set of features.

Deep learning-based features: Popular pre-trained networks on ImageNet dataset, such as, VGG19 [Simonyan et al., 2014] and ResNet50 [He et al., 2016], are used to extract feature representations for the images. Considering the difference between HER2-stained image dataset and ImageNet, we extract more generalizable lower level features using the pre-trained networks. Subsequently, the extracted features are used to train supervised convolutional neural networks (CNNs) to generate the desired staining quality indicators.

3.2.4 Sensitivity to process parameters

Sensitivity analysis of the staining quality to the staining process parameters benefits in obtaining optimal range for the process parameters for high-quality immunistaining. As the staining expectations differ across disease types, the optimal parameter space depends on the disease type. Therefore, we utilize the quality values (QV) of samples per disease type to perform the sensitivity analysis to parameters, namely antibody concentration/dilution (C) and residence time (RT).

First, QVs are interpolated for all C and RT configurations over the entire parametric space to have a dense and smooth distribution of QV. We triangulate the input data (C, RT, QV), available at specific configurations, with Quickhull algorithm [Barber et al., 1996] and construct a piece-wise cubic interpolating Bezier polynomial on each triangle [Farin, 1986] for interpolating at desired C and RT configurations using a Clough–Tocher scheme [Alfeld, 1984]. Second, a smooth 3D manifold is fitted to the QVs on a 3D coordinate system with C, RT and QV as the major axes. The 3D manifold enables a better visualization and more comprehensive statistical analysis of the sensitivity of QV with respect to the staining process parameters.

We perform sensitivity analysis at every configuration using variation quantification, similar to [Seguin et al., 2014]. At a point $p_i{=}(C_i, RT_i, QV_i)$ on the surface, we calculate the difference vector, v_i, between p_i and its 8-connected neighbors. Then, covariance matrix is computed for v_i, as $\mathcal{C}_i = v_i v_i^H$, where v_i and $\mathcal{C}_i$ signify the degree of change in QV in the neighboring configurations. Eigenvalue decomposition of $\mathcal{C}_i$ quantifies the degree of variation at p_i in different directions. The maximum eigenvalue indicates the degree of maximum variation at p_i and the corresponding eigenvector indicates the direction of maximum deviation. The higher the eigenvalue at a point, the higher the degree of variation, implying a higher sensitivity of staining quality to slight variations in corresponding process parameters at the point. Subsequent to obtaining the disease-type specific sensitivity information at all parametric configurations, we can select the operational parameter bounds that produce a high staining quality with a low sensitivity of the quality to variations in the process parameters.

3.3 Materials

Tissue microarrays (TMAs) (Novusbio, USA) of HT, PT, and MT from different patients were obtained to perform HER2-staining. HER2 is a clinically relevant protein, as it is related to an aggressive tumor progression and is the target of immunoherapeutic agent trastuzumab. TMA cores were graded as HER2+ or HER2- by the vendor according to their protein expression levels. TMAs were dried at 60°C for 15 min, dewaxed, rehydrated, and processed with heat induced antigen retrieval. Peroxidase and protein blocks were applied to the TMA prior to staining as per vendor's recommendation. Monoclonal HER2 antibodies (Thermo Fisher Scientific, USA) with concentrations of 6.25, 12.5 or 25 μg/mL were exposed on to the core using an MFP head, that stained the tissue section in a diameter of 300 μm. Each TMA core was patterned with 8 footprints of increasing incubation time between 12 and 289 seconds to generate a gradient. Images of each stained regions were acquired at 40x magnification using a bright field microscope. Exposure time was set to 24 ms with a lamp voltage of 6V, field stop is set to 30.5 mm, and aperture stop to 30.5 mm. The neutral density filter was adjusted for 5.8% transmittance.

White balance was automatically adjusted with a region clear of cells on the tissue as a reference prior to imaging. Several tissue specimens were collected per TMA core and each specimen was stained for only a particular antibody concentration and residence time configuration.

3.4 Results

The image dataset used for empirical evaluation of the proposed staining quality assessment and sensitivity analysis to process parameters methodology consisted of 488 IHC-stained images from 61 TMA cores across five disease types, namely, HT, PT-, PT+, MT-, MT+. Each image is annotated as Acceptable and NotAcceptable regarding the quality of immunostaining. Our methodology starts with a segmentation of each image into *footprint*, *off-footprint*, *foreground* and *background* regions. Subsequently, hand-crafted and deep learning-based features were extracted to train disease type and membrane-to-background contrast level identifying supervised probabilistic classifiers that returns the two quality indicators. The conducted experiments and the impact of experimental hyperparameters on the extraction of individual quality indicators are explained in detail in the following subsections.

3.4.1 Extraction of first quality indicator

The first quality indicator conveys the disease type information for an image, which is obtained via a 5-class supervised classifier. A balanced subset of 267 images across all disease types was selected that contained sufficient cell materials and represented the respective disease types for both poor (over- and insufficient staining) and high-quality staining. Both statistical machine learning-based and deep learning-based disease type classifiers were trained to maximize the 10-fold cross-validation F1 score. Details on the training and tuning of individual classifiers are presented below.

Traditional machine learning-based classifier: We extracted 584 hand-crafted features for each image, namely intensity (5), segmentation statistics (3), SIFT (128), texture (22), Gabor (26), discrete wavelet transform (100) and dual-tree complex wavelet transform (300) based features. The acquired features and disease type labels, obtained from the vendor, were used to train a Support Vector Machine (SVM) classifier. To obtain the optimal classifier, several hyperparameters, such as patch size for extracting local features, feature categories, and feature combinations, kernel types and hyperparameters, were fine-tuned as described below in order.

Most of the features are extracted from local patches, hence the choice of patch size has a significant impact on the overall classification performance. We examined with different patch sizes, i.e., 48×48, 64×64, 96×96, 128×128 and 160×160 pixels. The best F1 score was achieved for the classifier trained with a patch size of 64×64 pixels.

We evaluated the impact of individual feature categories on the disease type classification by training separate classifiers for each feature group. Rotation- and scale-invariant Gabor wavelet features performed the best as individual feature types, followed by texture and dual-tree complex wavelet features. Combination of different feature groups significantly improved the classification performance. The combination of intensity, texture, Gabor wavelet and dual-tree complex wavelet feature groups (353 features in total) achieved the best F1 score. A further improvement was attained by ranking and selecting the top

Table 3.1: Disease type confusion matrix for the best classifier trained with hand-crafted features.

	HT	PT-	PT+	MT-	MT+
HT	52	2	1	0	3
PT–	0	41	2	6	1
PT+	0	2	51	1	6
MT–	1	5	1	51	1
MT+	1	2	9	3	25

features. After feature selection, we obtained the best accuracy by including 70 features, which reduced the total number of features by 5-folds and provided an increment of 5% in overall accuracy, with the optimal set of hyperparameters.

Different types of kernels, namely linear, polynomial with degrees of 3, 5, 7 and 10, sigmoid, radial-basis function (RBF), Hellinger, Jensen-Shanon, with appropriate fine-tuning of hyperparameters were examined with SVM classifier. The best F1 score of **0.823** was achieved for an SVM classifier trained with RBF kernel and optimal feature set. Table 3.1 presents the confusion matrix obtained for 5-class disease type classification for the best-trained classifier. The confusion matrix indicates the efficacy of the trained classifier in identifying the tissue-types and HER2-expression status. As expected, most of the confusion occurs between PT- and MT- and between PT+ and MT+, which corresponds to a high similarity in the staining behaviors of the HER2- and HER2+ disease types.

Deep Learning-Based Classifier. We augmented the dataset for training a deep network by extracting 50 random patches per image, which were of size 224×224 pixels and hold more than 70% overlapping with the *foreground*. The patches extracted from an image were annotated with the disease type label of the complete image. We employed a ConvNet, pre-trained on ImageNet, to process the patches. Considering the dissimilarity between the HER2-stained IHC patches and ImageNet, we extracted more generalizable features from a lower layer of the network. Subsequently, another shallow CNN was trained using the per-patch extracted representations and disease type labels.

We experimented with two pre-trained ConvNets, VGG19 and ResNet50, in Keras. The features were extracted after the third block in both the architectures that resulted in outputs of size 28×28×256 and 28×28×512 for VGG19 and ResNet50 respectively. The subsequent CNN architecture and network training parameters are presented in Table 3.2. In the testing phase, the trained network predicted disease types for 50 extracted patches from a test image, and majority voting was performed to assign the final disease type to the image. The trained networks with features from VGG19 and ResNet50 pre-trained models achieved **0.758** and **0.834** F1 scores respectively.

The results in Table 3.3 indicate that the hand-crafted feature-based method performs similarly to CNN in the disease type, HER2 expression and tissue type identification tasks. The F1 score and Cohen’s kappa coefficient indicates very good agreement between the original labels and the predicted labels. The kappa coefficient conveys that the trained classifiers perform well in spite of class-imbalance in the training dataset. SVM is faster and easier to train, easier to tune hyperparameters and possess easier explainability compared to a deep network. As the results from the SVM and the (ResNet50 + CNN)

Table 3.2: CNN architecture and network hyperparameters for disease type classification.

Type	Output Size	Block	Strides
convolution	28x28x16	[1x1, 16]	1
max pool	14x14x16	-	2
convolution	14x14x8	[3x3, 8]	1
max pool	7x7x8	-	2
convolution	7x7x4	[3x3, 4]	1
flatten	196	-	-
dropout (50%)	196	-	-
linear	196	-	-
softmax	5	-	-

convolution layer = (convolution + ReLU + batch normalization),
batch size = 128, He uniform initialization, Adam optimizer,
cross-entropy loss, learning rate=0.01

Table 3.3: Machine learning and deep learning-based disease type classification results.

Approach	Task	F1	Kappa
Hand-crafted + SVM	5-class disease type	0.823	0.779
	3-class HER2 expression	0.921	0.878
	3-class tissue type	0.854	0.773
VGG19 + CNN	5-class disease type	0.761	0.699
	3-class HER2 expression	0.884	0.820
	3-class tissue type	0.802	0.692
ResNet50 + CNN	5-class disease type	0.835	0.793
	3-class HER2 expression	0.925	0.884
	3-class tissue type	0.865	0.791

5-class disease type = (HT, PT-, PT+, MT-, MT+)
3-class HER2 expression = (HT, HER2+, HER2-)
3-class tissue type = (HT, PT, MT)

were comparable, we proceeded with SVM-based development for subsequent tasks. With the inclusion of more images in the dataset, the deep networks may become the method of choice. Figure 3.6 presents the misclassified images by the SVM. Overstained PT- and MT- samples (a, b respectively) displayed comparable membrane staining to HER2+, thus being misclassified as PT+ (false positive). Understained and overstained PT+ (c, d respectively) lack sufficient staining contrast between the foreground and the background, similar to HER2- images, thus being misclassified as PT- and MT- (false negative). The bottom row presents the ambiguous PT+ (e), MT+ (f) and PT- (g), MT- (h) samples. Due

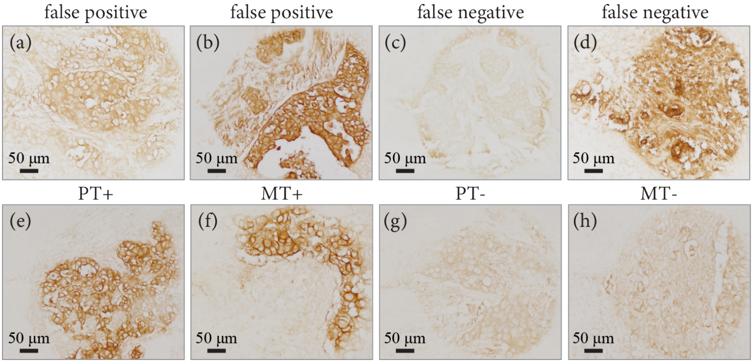


Figure 3.6: Misclassified samples by the Hand-crafted + SVM classifier. (top row) False positives, where PT- and MT- samples were interpreted as PT+ (a, b) and false negatives, where PT+ samples were interpreted as PT- and MT- respectively (c, d). (bottom row) Ambiguous PT+ (e), MT+ (f) and PT- (g), MT- (h) samples.

to the similarity in staining between HER2 overexpressing tissues (PT+ and MT+) and HER2 no over-expressing tissues (PT- and MT-), these images were misclassified.

3.4.2 Extraction of second quality indicator

The second quality indicator conveys the membrane-to-background contrast level information of a stain, which is obtained via a binary-class supervised probabilistic classifier. A balanced subset of 77 images were selected that clearly represented high and low contrast levels irrespective of the disease type. We used the 584 features extracted per image to train the contrast level classifier and tuned the same hyperparameters similar to the first quality indicator. The best trained classifier achieved a 5-fold cross validated F1 score of **0.947**. The best classifier was an SVM trained with RBF kernel and with 63 features, which predominantly included features from intensity, Gabor wavelet and dual-tree complex wavelet transform categories.

3.4.3 Staining quality assessment

The disease-type specific SQMs were trained with the two quality indicators extracted from the two supervised probabilistic classifiers. We selected sets of 91, 106, 96 and 60 images from PT-, PT+, MT- and MT+ disease categories respectively with balanced sets of images from both Acceptable and NotAcceptable staining qualities. Individual SQMs were trained in supervised manner using quality labels acquired from the consensus of three experts. Individual SQMs were analyzed using the area-under-the-curve (AUC) measure of the respective receiver operating characteristic (ROC) curves. The optimal SQMs were obtained for SVM with RBF kernel, which achieved AUC scores of **0.84**, **0.83**, **0.82** and **0.90** for the PT-, PT+, MT- and MT+ disease types respectively. An average AUC score of **0.85** is achieved for the proposed methodology with individual SQMs. For comparison, we trained an SQM that learns the staining quality labels for samples directly using their respective feature representations. After tuning all the hyperparameters of the

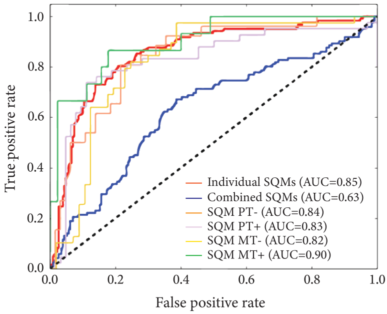


Figure 3.7: ROC curves and AUC scores for individual SQMs, aggregated result of individual SQMs and direct SQM, trained directly with sample features and sample quality labels.

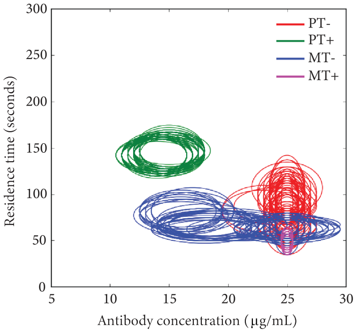


Figure 3.8: Disease-type specific 95% confidence ellipses fitted to quality values over the process parameter space of antibody concentration and residence time.

direct SQM, we achieved an overall AUC score of **0.63**. The ROC curves for individual SQMs, the aggregated SQM and the direct SQM are presented in Figure 3.7.

The QVs for the samples were acquired from the individual SQMs and the QVs were interpolated over the entire parameter space of C and RT. Subsequently, the disease-type specific 3D manifolds were fitted to the QVs, C and RT configurations, as shown in Figures 3.9(a, d). Figure 3.9(e) displays the 2D contour plot of fitted 3D manifold for SQM$_{PT+}$, where the yellow region corresponds to the parameter space with $\geq$95% staining quality. Fitting an ellipse to this region corresponds to the 95% confidence interval of QV. We evaluated the robustness of our SQM algorithm by generating confidence ellipses for individual SQMs using 200 over-sampled bootstrap datasets for different disease type categories. Figure 3.8 displays the confidence ellipses for all the trained disease-type specific SQMs. For PT+ and MT+, the confidence ellipses are concentrated in a specific parameter region, whereas for PT- and MT- they are more dispersed. The consistency of the confidence ellipses for PT+ and MT+ indicate that the staining process parameters can be confidently confined to a specific parameters space to achieve high-quality staining, whereas this sort of confinement is not possible for PT- and MT- disease types. Usually in HER2+ tissue sections, the overexpression of the HER2 protein is consistent, resulting in consistent stained-expressions. For HER2- tissue samples, the HER2 protein has weak HER2-overexpression, which can have a high variability in staining expressions. The degree of variability in HER2 overexpression can explain the resulting behavior of the confidence ellipses.

3.4.4 Sensitivity to process parameters

The disease-type specific SQM manifolds were used to evaluate the variability of the staining quality scores with respect to variations in the process parameters. Using the eigenvalue based variational quantification approach, we inspected the sensitivity of the staining quality for all possible process parameter configurations. For instance, Figure 3.9(e) and Figure 3.9(f) present the 2D contour plot of the staining quality and the

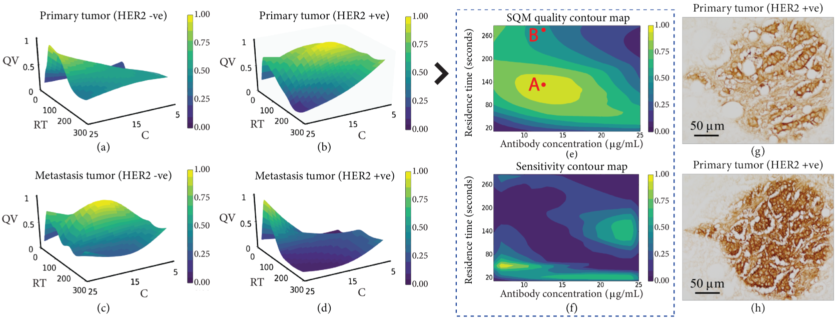


Figure 3.9: Manifold fitting to quality values (QVs) acquired using disease-type specific SQMs, **(a)** PT-, **(b)** PT+, **(c)** MT-, and **(d)** MT+. The manifolds are fitted using QVs of samples across all configurations of antibody concentrations (C) and residence times (RT). Also shown is, the sensitivity analysis for PT+: **(e)** 2D contour map of QVs and **(f)** 2D contour map of the staining sensitivity to process parameters (C and RT).

2D contour plot of the sensitivity analysis for SQM_{PT+}. Figure 3.9(e) shows that high-quality staining can be obtained when operating in the range of $9 < C < 17\,\mu g/mL$ and $90 < RT < 160\,s$ and that the best quality is obtained for $C = 14\,\mu g/mL$ and $RT = 120\,s$. It also illustrates that the staining quality is low for low-end and high-end C and RT values, which is consistent with the concepts of under-staining and over-staining, respectively. These observations can help reduce false negative and false positive staining, but this map does not convey the stability in operating with these parameter settings. Figure 3.9(f) presents the sensitivity information by plotting the degree of variation in the staining quality for each C and RT configuration. It shows that the staining quality is slightly sensitive towards the lower-end and the upper-end of the aforementioned range of C values. Combining the knowledge from both the maps, an operational range of $11 < C < 15\,\mu g/mL$ and $90 < RT < 130\,s$ can be selected for generating stable and high-quality stains for PT+. Figure 3.10 presents stained images of a PT+ TMA core for the entire parametric configurations of antibody concentration and residence time showing the staining and quality value variability within a core. The PT+ specific quality values per image are indicated at the top left corner of each image. The best QVs, corresponding to the optimal staining region, are highlighted in red. Similarly, optimal parameter configurations and best practices for other tissue categories and biomarkers can be inferred from their sensitivity analyses.

3.4.5 Comparison with a clinical staining protocol

To gauge the value of the proposed method, we aimed to understand whether the methodology can be transferred to a clinical setting. Thus, we performed staining M1, using a protocol used currently in a hospital, and staining M2, the proposed optimized parametrization. These evaluations were performed using in vitro diagnostics antibodied (Herceptest, Dako) with an on-bench approach, i.e. without the use of an MFP for primary antibody deposition. Since the antibody for these tests differs from the antibody used for developing the proposed methodology (anti-Her2 antibody, ThermoFisher), the trained

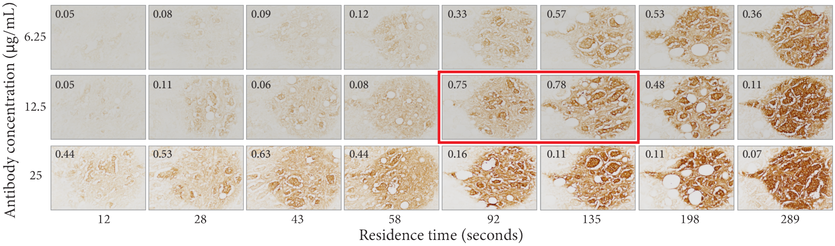


Figure 3.10: IHC stained images of a sample PT+ TMA core for the entire parametric configurations of antibody concentration and residence time. The PT+ specific quality values are noted at the top left corner of each image. The best QVs, corresponding to the optimal staining region, are highlighted in red.

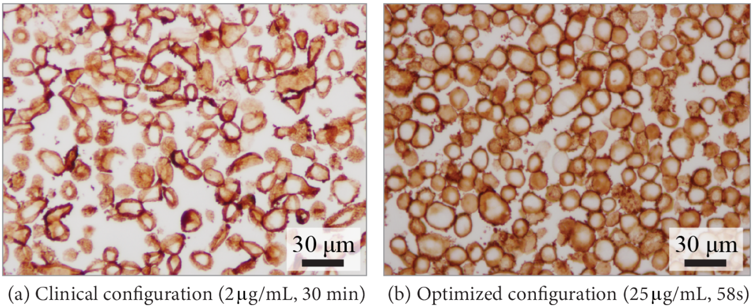
classifiers, SQMs and derived optimum staining configurations could potentially present variations, due to differences in antibody kinetic parameters. However, we expected that the results can be transferable between different antibodies, showing the robustness of the method. Therefore, we analyzed M1 and M2 using the parametric space from our development dataset. To remove bias to antibody selection, we extracted feature representations for the stained cell blocks and scaled the features to the same range as of the features used during the training phase.

For M1, SKBR3 cell blocks were stained using an antibody concentration of 2 μg/mL for 30 minutes, as recommended by the provider (Figure 3.11(a)). The proposed disease type classifier categorized SKBR3 to be MT+, as expected since SKBR3 cells were acquired from a metastatic site. Then for M2, we applied the MT+ specific optimal staining condition, 25 μg/mL for 58 s, as derived by the proposed method (Figure 3.11(b)). The QVs were computed using SQM$_{MT+}$, and resulted to QVs of 0.12 and 0.88 for M1 and M2 respectively. Both visual inspection and qualitative assessment demonstrate that M2 produced clearer diagnostically relevant information, namely sharper stained membrane signal compared to M1.

Despite the change in antibody, the optimized staining approach delivered a better QV. Hence, we proceeded to stain PT+ tissues with the PT+ specific optimized configuration (antibody concentration of 12.5 μg/mL for 135 s) using clinically validated antibodies on-bench. Figures 3.12(a-c) depicts three stained PT+ images. Their QVs were computed to be 0.83, 0.51 and 0.87 respectively using SQM$_{PT+}$. Despite the much shorter residence time than recommended by the provider (135 s vs 30 minutes), these samples present appropriate staining on the cell membranes and possess good relative intensity between the membrane and the cytoplasm, as confirmed by our experts. The estimated QVs classified the images to be in the good staining range.

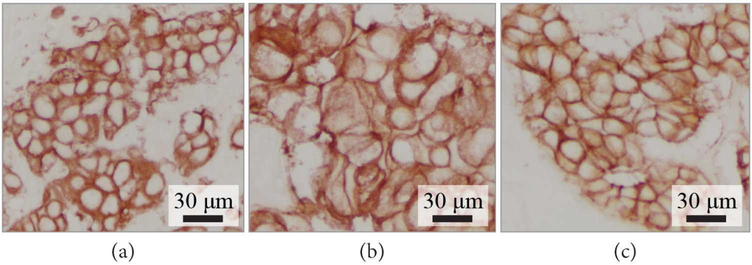
3.5 Conclusion

In this paper, we introduce a methodology to analyze the immunostaining quality sensitivity with respect to the staining process parameters, i.e., antibody dilution and residence time. The proposed methodology initially delineates the diagnostically relevant and contextually immaterial signals in a given immunostained tissue. It then learns machine



(a) Clinical configuration (2 µg/mL, 30 min) (b) Optimized configuration (25 µg/mL, 58s)

Figure 3.11: SKBR3 cell blocks stained with clinical protocol and optimized staining protocol.



(a) (b) (c)

Figure 3.12: IHC images of PT+ tissues stained with the proposed optimized staining protocol.

learning based disease-type specific staining quality metrics using extracted comprehensive features relevant to staining quality. Subsequently, it performs statistical sensitivity analysis of the staining quality with respect to the process parameters. The proposed quantitative quality metric and the sensitivity analysis contribute to the process parameter optimization to achieve high-quality staining for various disease types. As a model system, the proposed methodology is validated on a cohort of HER2-stained breast cancer tissues from five different disease types, stained using μIHC under various parameter configurations of the MFP. Utilization of the MFP allowed to stain a small fraction of a tissue section for the analysis and extrapolation of suitable process parameters.

We believe that the entire methodology can be extended for other types of staining, disease types, and tumor types as it does not involve any prior assumptions about the staining method. It can easily be applied to the conventional whole-slide-staining and staining with other biomarkers. This allows the comparison of different antibodies leading to the choice of the best and finding the corresponding optimal staining protocols.

With the proposed method, the number of false positives and false negatives produced by incorrect parametrization can be reduced substantially. For instance, when the disease state is unknown, the optimal configuration of PT+ could be applied as a first approximate set of parameters. In case of a low-quality result for PT+, one of the parameters can be kept fixed, while the other is modified to the closest optimal value from MT+. Performing this sequentially with the aim to maximize the quality metric, a set of optimal parameters can easily be scanned on a tissue. Comparing the information known about the disease type and the one obtained with the algorithm can provide potential information on the developmental status of the tumor.

Digital and computational pathology have received increasing attention from the medical community as they can aid the accuracy of decisions made by pathologists, while also reducing workloads and removing subjective artifacts. However, the underlying aspects of staining quality still remain only partially solved. The suite of methods outlined in this paper can assist pathologists and improve the reproducibility since it establishes an objective metric and reduces the human factor.

4

Hierarchical Graph Representations in Digital Pathology

Cancer diagnosis, prognosis, and therapy response predictions from tissue specimens highly depend on the phenotype and topological distribution of constituting histological entities. Thus, adequate tissue representations for encoding histological entities are imperative for computer aided cancer patient care. To this end, several approaches have leveraged cell-graphs, capturing cell-microenvironment, to depict the tissue. These allow for utilizing graph theory and machine learning to map the tissue representation to tissue functionality, and quantify their relationship. Though cellular information is crucial, it is incomplete alone to comprehensively characterize complex tissue structure. We herein treat the tissue as a hierarchical composition of multiple types of histological entities from fine to coarse level, capturing multivariate tissue information at multiple levels. We propose a novel multi-level hierarchical entity-graph representation of tissue specimens to model the hierarchical compositions that encode histological entities as well as their intra- and inter-entity level interactions. Subsequently, a hierarchical graph neural network is proposed to operate on the hierarchical entity-graph and map the tissue structure to tissue functionality. Specifically, for input histology images, we utilize well-defined cells and tissue regions to build HierArchical Cell-to-Tissue (HACT) graph representations, and devise HACT-Net, a message passing graph neural network, to classify the HACT representations. As part of this work, we introduce the BReAst Carcinoma Subtyping (BRACS) dataset, a large cohort of H&E stained breast tumor regions-of-interest, to evaluate and benchmark our proposed methodology against pathologists and state-of-the-art computer-aided diagnostic approaches. Through comparative assessment and ablation studies, our proposed method is demonstrated to yield superior classification results compared to alternative methods as well as individual pathologists. The code, data, and models can be accessed at https://github.com/histocartography/hact-net

4.1 Introduction

Breast cancer is the most commonly diagnosed cancer and registers the highest number of deaths for women with cancer ([Sung et al., 2021]). A study by [Allemani et al., 2015]

This chapter has been published as: Pushpak Pati†, Guillaume Jaume†, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel and Riccio, Maurizio Di Bonito, Giuseppe De Pietro, Gerardo Botti, Jean-Philippe Thiran, Maria Frucci, Orcun Goksel, Maria Gabrani, “Hierarchical Graph Representations in Digital Pathology”, In: Medical Image Analysis, vol. 75, pp. 102264, 2022, doi: 10.1016/j.media.2021.102264, [arXiv]

exhibits that intensive early diagnostic activities have improved 5-year survival to 85% during 2005–09 for breast cancer patients. Early diagnosis of cancer, primarily through manual inspection of histology slides, enables the acute assessment of risk and facilitates an optimal treatment plan. Though the diagnostic criteria for breast cancer are established, the continuum of histologic features phenotyped across the diagnostic spectrum prevents distinct demarcation. Thus, manual inspection is tedious and time-consuming with significant intra- and inter-observer variability ([Gomes et al., 2014; Elmore et al., 2015]). The increasing incidence rate of breast cancer cases per year ([Siegel et al., 2020]) and the challenges in manual diagnosis demand for automated computed-aided diagnostis.

Whole-slide scanning systems empowered rapid digitization of pathology slides into high-resolution whole-slide images (WSIs) and profoundly transformed pathologists' daily practice ([Mukhopadhyay et al., 2017]). Further, they enabled computer aided diagnostics to leverage artificial intelligence ([Litjens et al., 2017; Deng et al., 2020]), especially deep learning (DL), to address various pathology tasks, such as nuclei segmentation ([Kumar et al., 2017; Graham et al., 2019a]), nuclei classification ([Pati et al., 2021; Verma et al., 2021]), gland segmentation ([Graham et al., 2019b; Binder et al., 2019]), tissue segmentation ([Mehta et al., 2018; Mercan et al., 2019b]), tumor detection ([Aresta et al., 2019; Bejnordi et al., 2017b; Pati et al., 2018]), tumor staging ([Aresta et al., 2019; Mercan et al., 2019a]), and survival analysis ([Zhu et al., 2017; Yao et al., 2021a]). DL techniques primarily use Convolutional Neural Networks (CNNs) ([Madabhushi et al., 2016; Parwani, 2019]) to process histology images in a patch-wise manner. CNNs extract representative patterns from patches and aggregate them to perform image-level tasks. However, patch-wise processing suffers from the trade-off between the resolution of operation and the utilization of adequate context ([Bejnordi et al., 2017a; Sirinukunwattana et al., 2018]). Operating at a higher resolution captures local cellular information but limits the field-of-view due to computational burden and limits the access to global tissue microenvironment. In contrast, operating at a lower resolution hinders resolvability of cells and access to cellular properties. [Bejnordi et al., 2017a; Sirinukunwattana et al., 2018; Tellez et al., 2021] have proposed CNN methods to address such trade-off by leveraging visual context, however, CNNs, which operate on fix-sized input patches, are confined to a fixed field-of-view and are restricted to incorporate information from varying spatial distances. Further, pixel-based processing in CNNs disregards the notion of histologically meaningful entities ([Hägele et al., 2020]), such as cells, glands, and tissue types. The inattention to histological entities severely limits the interpretability of CNNs, and any utilization of established entity-level prior pathological knowledge in the CNN-based frameworks. Additionally, CNNs disregard the structural composition of tissue, where fine entities hierarchically constitute to form coarse entities, such as, epithelial cells organize to form epithelium, which further constitutes to form glands. Such hierarchical structure is relevant both for diagnostics and interpretation.

In this paper, we address the aforementioned limitations by shifting the analytical paradigm from pixel to entity-based processing. In an entity paradigm, a histology image is described as an entity-graph, where nodes and edges of a graph denote biological entities and inter-entity interactions, respectively. An entity-graph can be customized in various aspects, e. g., in terms of the type of entity set, entity attributes, and graph topology, by incorporating any task-specific prior pathological knowledge. Thus, the graph representation enables pathology-specific interpretability and human-machine co-learning. In addition, the graph representation is memory efficient compared to pixelated images and can seamlessly describe a large tissue region. [Demir et al., 2004] first introduced cell-graphs using cells as the entity type. Though a cell-graph efficiently encodes the cell microenvironment, it cannot extensively capture the tissue microenvironment, i. e., the distribution of tissue regions such as necrosis, stroma, and epithelium. Similarly, a tissue-graph comprising of the set of tissue regions cannot depict the cell microenvironment. Therefore, an entity-graph representation using a single type of entity set is insufficient to comprehensively describe the tissue composition. To address this, we propose a multi-level entity-graph representation, i. e., HierArchical Cell-to-Tissue (HACT), consisting of multiple types of entity sets, i. e., cells and tissue regions, to encode both cell and tissue microenvironment. The multiset of entities is inherently coupled depicting tissue composition at multiple scales. The HACT graph encodes individual entity attributes and intra- and inter-entity relationships to hierarchically describe a histology image. Upon the graph construction, a graph neural network (GNN), a DL technique operating on graph-structured data, processes the entity-graph to perform image analysis. Specifically, we introduce a hierarchical GNN, HierArchical Cell-to-Tissue Network (HACT-Net), to sequentially operate on HACT graph, from fine-level to coarse-level, to provide a fixed dimensional embedding for the image. The embedding encodes morphological and topological distribution of the multiset of entities in the tissue. Interestingly, our proposed methodology resembles the tissue diagnostic procedure in clinical practice, where a pathologist hierarchically analyzes a tissue.

We propose a methodology that consists of HACT graph construction and HACT-Net based histology image analysis of breast tumor regions-of-interest (TRoIs). A preliminary version of this work was presented as [Pati et al., 2020]. Our substantial extensions herein include, (1) an improved HACT representation and HACT-Net architecture, (2) a larger evaluation dataset (twice the earlier size), (3) detailed ablation studies and evaluation on public data, and (4) a benchmark comparison against independent pathologists. Specifically, the major contributions of this paper are:

- A novel hierarchical entity-graph representation (HACT) and hierarchical learning (HACT-Net) methodology for analyzing histology images;
- Introducing BReAst Carcinoma Subtyping (BRACS[1]), a large cohort of breast TRoIs annotated into seven breast cancer subtypes. BRACS includes challenging atypical cases and a variety of TRoIs to represent a realistic breast diagnosis;
- Benchmarking of HACT-Net against three independent pathologists on BRACS. An extensive assessment to demonstrate the superiority of HACT-Net over recent CNN and GNN approaches for cancer subtyping, while being comparable to pathologists's classification performance.

4.2 Related work

Tumor subtyping in digital pathology: Several DL algorithms have been proposed to categorize histopathology images into cancer subtypes ([Komura et al., 2018; Srinidhi et al., 2021; Deng et al., 2020; Spanhol et al., 2016a; Araùjo et al., 2005; Aresta et al., 2019]). For this task, most algorithms employ CNNs in a patch-wise manner. In [Araùjo et al., 2005; Bardou et al., 2018; Roy et al., 2019; Mercan et al., 2019a], CNNs are used to classify breast

---

[1] BRACS dataset for breast cancer subtyping: https://www.bracs.icar.cnr.it

histology images. These methods use single stream patch-wise approaches to capture local patch-level context, aggregate the patch-level information, and classify the image using aggregated information. However, single-stream approaches do not capture adequate context from the tissue microenvironment to aptly encode a patch. [Sirinukunwattana et al., 2018] address this issue by including multi-scale information from concentric patches across different magnifications. [Tellez et al., 2021] propose neural image compression, where WSIs are compressed using a neural network trained in an unsupervised fashion, followed by a CNN trained on the compressed representations to classify the images. [Shaban et al., 2020] include an attention module with an auxiliary task to improve neural image compression for histology image classification. [Bejnordi et al., 2017a] propose a stacked CNN architecture to capture large contexts and perform end-to-end processing of large histology images. [Pinckaers et al., 2020] propose a streaming CNN to accommodate multi-megapixel images. [Campanella et al., 2019] utilize a multiple-instance learning approach to process whole-slide images in an end-to-end manner, which is extended by [Lu et al., 2021] to automatically identify sub-regions of high diagnostic value via an attention mechanism. Though the aforementioned methods use different strategies to encode a tissue, they all operate on a square and fix-sized patches. However, actual TRoIs can be of diverse dimensions depending on the cancer subtype and the site of tissue extraction. Our proposed entity-graph methodology can acquire both local and global context from arbitrary-sized TRoIs to address the aforementioned limitations.

Graphs in digital pathology: Entity-graph-based tissue representations can effectively describe the tissue composition by incorporating morphology, topology, and interactions among biologically comprehensible entities, unlike CNNs. Using cells as entities, [Demir et al., 2004] introduced a cell-graph (CG) representation, where cell morphology can be embedded in the nodes via hand-crafted ([Demir et al., 2004; Zhou et al., 2019a; Pati et al., 2020]) or deep-learning based features ([Chen et al., 2020a]). The graph topology is often heuristically defined, e. g., using k-Nearest Neighbors, probabilistic modeling, or a Waxman model ([Sharma et al., 2015]). Then, a CG is processed by classical machine learning techniques ([Sharma et al., 2016; Sharma et al., 2017a]) or GNNs ([Zhou et al., 2019a; Pati et al., 2020; Chen et al., 2020a; Anand et al., 2019]) for mapping to tissue function. Recently, patches ([Aygüneş et al., 2020]) and tissue regions ([Pati et al., 2020; Anklin et al., 2021]) as entities have been used for better tissue representation. Other graph-based applications in computational pathology include cellular community detection ([Javed et al., 2020]), WSI classification ([Zhao et al., 2020b; Adnan et al., 2020]), WSI segmentation ([Anklin et al., 2021]). Notably, entity-graphs consist of biological entities to which the pathologists can readily relate. So, the entity-graph paradigm enables to incorporate pathologically-defined, task-specific entity-level prior knowledge in constructing "meaningful" tissue representations. This implicitly enables *interpretability* and *explainability* of graph-based networks for pathologists. To this end, [Zhou et al., 2019a] analyzes the clustering of nodes in a CG to group cells according to their appearance and tissue types. [Jaume et al., 2020] introduces a post-hoc graph-pruning explainer to identify decisive cells and interactions. [Sureka et al., 2020] employs robust spatial filtering that utilizes an attention-based GNN and node occlusion to highlight cell contributions. [Jaume et al., 2021b] propose quantitative metrics leveraging pathologically relevant cellular properties to characterize graph explainability for CG analysis.

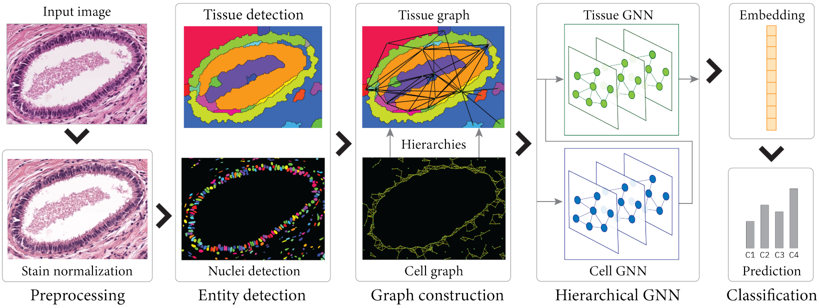


Figure 4.1: Overview of the proposed hierarchical entity-graph based tissue analysis methodology. Following some pre-processing, a hierarchical entity-graph representation of a tissue is constructed, and it is processed via a hierarchical graph neural network to learn the mapping from tissue compositions to respective tissue categories.

4.3 Methodology

In this section, we detail our proposed methodology for hierarchical tissue analysis, as illustrated in Figure 4.1. For an input (H&E) stained histology TRoI image, first, we apply pre-processing to standardize the input. Then, we identify pathologically relevant entities and construct a HACT graph representation of the TRoI by incorporating the morphological and topological distribution of the entities. Finally, HACT-Net, a hierarchical GNN, is devised to map the HACT graph to a corresponding category, e. g., cancer subtype.

4.3.1 Notations

We define an attributed, undirected entity-graph $G := (V, E, H)$ as a set of nodes V, edges E, and node features H. Each node $v \in V$ is represented by a feature vector $h(v) \in \mathbb{R}^d$, thus, $H \in \mathbb{R}^{|V| \times d}$. d denotes the number of features per node, and $|.|$ denotes set cardinality. An edge between two nodes $u, v \in V$ is denoted as e_{uv}. The graph topology is described by a symmetric adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$, where $A_{u,v} = 1$ if $e_{uv} \in E$. The neighborhood of a node $v \in V$ is denoted as $\mathcal{N}(v) := \{u \in V | v \in V, e_{uv} \in E\}$.

4.3.2 Pre-processing

H&E stained images exhibit appearance variability due to different specimen preparation techniques, staining protocols, fixation characteristics, imaging device characteristics etc. Such variability adversely impacts computational methods for downstream diagnosis [Veta et al., 2014; Tellez et al., 2019]. To alleviate the variability, we use the unsupervised, reference-free stain normalization algorithm proposed by [Macenko et al., 2009]. The algorithm is based on the principle that RGB color of each pixel is a linear combination of two unknown stain vectors, Hematoxylin and Eosin, that need to be estimated. First, the algorithm estimates the stain vectors of a TRoI by using a Singular Value Decomposition of the non-background pixels. Second, the algorithm applies a correction to account for the intensity variations due to noise. The algorithm requiring no model training is

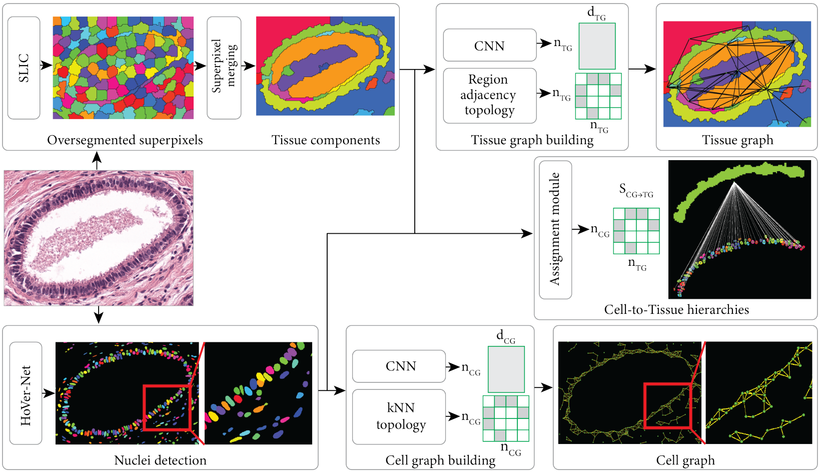


Figure 4.2: Overview of hierarchical cell-to-tissue (HACT) graph construction for a TRoI. Our HACT graph representation consists of a cell-graph, a tissue-graph, and cell-to-tissue hierarchies, while encoding the phenotypical and topological distributions of tissue entities to describe the cell and tissue microenvironments.

computationally inexpensive. Specifically, for stain normalization we employ the scalable and fast pipeline proposed by [Stanisavljevic et al., 2018].

4.3.3 Graph representation

A stain normalized TRoI is processed to identify relevant entities and construct a hierarchical entity-graph representation. In this work, we consider nuclei and tissue regions as the entities. Therefore, the HACT graph consist of three components: 1) a low-level *cell-graph*, capturing cell morphology and interactions, 2) a high-level *tissue-graph*, capturing morphology and spatial distribution of tissue regions, and 3) cells-to-tissue hierarchies, encoding the relative spatial distribution of cells with respect to the tissue distribution. The details of the components are presented in the following subsections.

4.3.3.1 Cell-graph representation

A cell-graph (CG) characterizes cell microenvironment, where nodes denote cells and encode cell morphology, and edges denote cellular interactions and encode cell topology. It is constructed in three steps, i) nuclei detection, ii) nuclei feature extraction, and iii) topology configuration, as shown in Figure 4.2.

Precise nuclei detection enables reliable CG representation. To this end, we use HoVer-Net, a nuclei segmentation network proposed by [Graham et al., 2019a], pre-trained on MoNuSeg dataset by [Kumar et al., 2017]. HoVer-Net leverages the instance-rich information encoded in the vertical and horizontal distances of nuclear pixels to their centers of mass. These distances are used to accurately segment clustered nuclei, particularly in areas with overlapping nuclei. The centroids of the segmented nuclei form the spatial coordinates of nodes in CG.

Following nuclei detection, morphological features are extracted by processing patches of size $h \times w$ centered around nuclei centroids via ResNet [He et al., 2016] pre-trained on ImageNet dataset [Deng et al., 2009]. Spatial features of the nuclei are extracted as the spatial coordinates of the nuclei, normalized by the TRoI dimensions. Morphological and spatial features together constitute the nuclei features, which are collocated for all nodes as the node-feature matrix $H_{\text{CG}} \in \mathbb{R}^{|V_{\text{CG}}| \times d_{\text{CG}}}$.

For the CG topology E_{CG}, we utilize the fact that spatially close cells have stronger interactions ([Francis et al., 1997]) with distant cells having weaker cellular interactions. Accordingly, we connect nearby cells with edges to model their interactions. To this end, we use the k-Nearest Neighbors (kNN) algorithm to build an initial topology, that we subsequently prune by removing edges longer than a threshold distance d_{min}. We use Euclidean distances between nuclei centroids in the image space to quantify cellular distances. Formally, for each node v, an edge e_{vu} is built if

$$\begin{aligned} u \in \{w \mid \text{DIST}(v,w) \leq d_k \wedge \text{DIST}(v,w) < d_{\text{min}},\ \forall w, v \in V_{\text{CG}}, \\ , d_k = k^{\text{th}} \text{ smallest distance in } \text{DIST}(v,w)\} \end{aligned} \tag{4.1}$$

CG topology is presented by a binary adjacency matrix $E_{\text{CG}} \in \mathbb{R}^{|V_{\text{CG}}| \times |V_{\text{CG}}|}$. Figure 4.2 illustrates the CG representation for a sample TRoI. Formally, a CG representation is formulated as $G_{\text{CG}} := \{V_{\text{CG}}, E_{\text{CG}}, H_{\text{CG}}\}$.

4.3.3.2 Tissue-graph representation

A tissue graph (TG) depicts a high-level tissue microenvironment, where the nodes and edges denote tissue regions and their interactions, respectively. A TG is constructed by first identifying tissue regions (e. g., epithelium, stroma, lumen,and necrosis), followed by encoding the tissue regions, and finally the topology building. The steps are illustrated in Figure 4.2. A parallel approach involving superpixel detection and neighborhood information aggregation is adopted by [Mercan et al., 2018] to semantically segment tissue regions in histology images.

Tissue regions are identified in two-steps. First, we oversegment the tissue to detect non-overlapping homogeneous superpixels. We operate at a low magnification to avoid noisy pixels and compute efficiently. Specifically, we use the Simple Linear Iterative Clustering (SLIC) algorithm ([Achanta et al., 2012]). SLIC follows an unsupervised approach by associating each pixel with a feature vector and merging the pixels using a localized version of k-means clustering. Next, we iteratively merge neighboring superpixels that have similar color attributes, i. e., channel-wise mean, to create superpixels that capture meaningful tissue information. A sample tissue-region instance-map is shown in Figure 4.2.

To extract feature representations of tissue regions, we follow a two-step procedure: first, we extract CNN-based features for oversegmented superpixels, i. e., patches of size $h \times w$ centered around the superpixel centroids are processed by ResNet. Second, morphological features of a tissue region are obtained by averaging the deep features of its constituting superpixels. Similar to CG, we include spatial features as the normalized

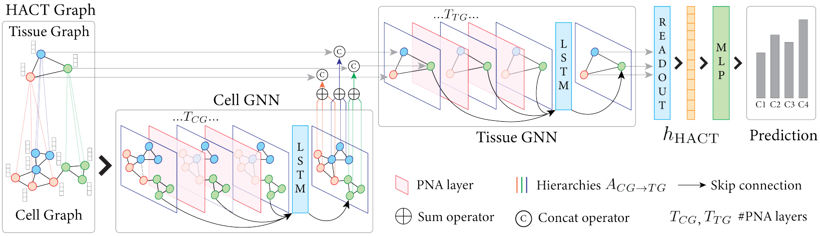


Figure 4.3: Overview of the proposed HACT-Net architecture. The network processes an input HACT graph representation in a hierarchical manner, from fine cell-level to coarse tissue-region level, to obtain a contextualized graph embedding, and consequently classify the input graph.

centroids of the tissue region. For a TRoI with a set of V_{TG} tissue regions, we denote the TG node-feature matrix as $H_{\text{TG}} \in \mathbb{R}^{|V_{\text{TG}}| \times d_{\text{TG}}}$.

We assume adjacent tissue regions to biologically interact the most, and thus connect in the TG topology. To this end, we construct a Region Adjacency Graph ([Potjer, 1996]) where an edge is built between adjacent tissue region. The topology is presented by a binary adjacency matrix $E_{\text{TG}} \in \mathbb{R}^{|V_{\text{TG}}| \times |V_{\text{TG}}|}$. Formally, a TG representation is formulated as $G_{\text{TG}} := \{V_{\text{TG}}, E_{\text{TG}}, H_{\text{TG}}\}$.

4.3.3.3 Hierarchical Cell-to-Tissue graph representation

A histopathology tissue can be considered as a hierarchical organization of biological entities ranging from fine-level, i. e., cells, to coarse-level, i. e., tissue regions. There exist intra- and inter-level coupling based on topological distributions and interactions among the entities. With this motivation, we propose HACT, a HierArchical Cell-to-Tissue (HACT) graph representation to jointly represent low-level CG and high-level TG. Intra-level topology is captured by standalone CG and TG. Inter-level topology is presented by a binary assignment (cell-to-tissue hierarchy) matrix $A_{\text{CG}\to\text{TG}} \in \mathbb{R}^{|V_{\text{CG}}| \times |V_{\text{TG}}|}$ that utilizes the relative spatial distributions of nuclei with respect to tissue regions. For the i^{th} nucleus and j^{th} tissue region, the corresponding assignment is given as,

$$\begin{aligned} A_{\text{CG}\to\text{TG}}[i,j] &= 1, \text{ if } i^{\text{th}} \text{ nucleus centroid} \in j^{\text{th}} \text{ tissue region} \\ A_{\text{CG}\to\text{TG}}[i,j] &= 0, \text{ otherwise} \end{aligned} \tag{4.2}$$

Cell-to-tissue hierarchies for a tissue region are presented in Figure 4.2. Each nucleus is assigned to one and only one tissue region. If a segmented nucleus is at the border of multiple tissue regions, the nucleus is assigned to the tissue region that it has the maximum overlap with. Formally for a given TRoI, a HACT representation is formulated as $G_{\text{HACT}} := \{G_{\text{CG}}, G_{\text{TG}}, A_{\text{CG}\to\text{TG}}\}$.

4.3.4 Graph learning

The HACT graph for a TRoI is processed by a hierarchical GNN to map TRoI composition to TRoI subtype. To this end, we propose HierArchical Cell-to-Tissue Network (HACT-Net), a hierarchical GNN architecture shown in Figure 4.3.

HACT-Net intakes G_{HACT} as input and outputs a graph-level representation $h_{\text{HACT}} \in \mathbb{R}^{d_{\text{HACT}}}$. Subsequently, a multi-layer perceptron (MLP) categorizes h_{HACT}, e. g., to a cancer subtype. Formally, HACT-Net consists of two GNNs, i. e., Cell-GNN (CG-GNN) and Tissue-GNN (TG-GNN), to hierarchically process the HACT graph from fine to coarse level. In this work, we leverage the recent advances in GNNs and model HACT-Net using PNA layers ([Corso et al., 2020]).

First, CG-GNN intakes $G_{\text{CG}} := \{V_{\text{CG}}, E_{\text{CG}}, H_{\text{CG}}\}$, and applies T_{CG} PNA layers to build contextualized cell-node embeddings, inline with Equation 2.4 The node embeddings $h^{(t)}(v), \forall v \in V_{\text{CG}}$ are iteratively updated as,

$$\begin{aligned} a_{\text{CG}}^{(t+1)}(v) &= \bigoplus{}_{u \in \mathcal{N}_{\text{CG}}(v)} M_{\text{CG}}^{(t)}\left(h_{\text{CG}}^{(t)}(v), h_{\text{CG}}^{(t)}(u)\right) \\ h_{\text{CG}}^{(t+1)}(v) &= U_{\text{CG}}^{(t)}\left(h_{\text{CG}}^{(t)}(v), a_{\text{CG}}^{(t+1)}(v)\right) \end{aligned} \tag{4.3}$$

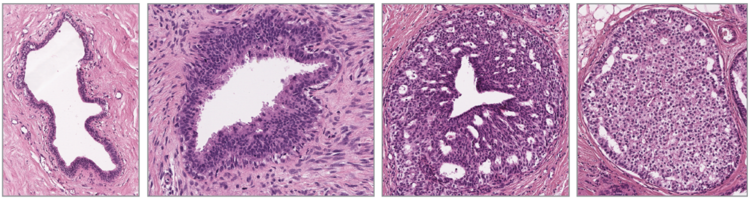
where $t = 0, \ldots, T_{\text{CG}}$ is the iteration index. For a node v, first, the set of neighboring node embeddings $\{h_{CG}^{(t)}(u)\}, \forall u \in \mathcal{N}_{\text{CG}}(v)$ are concatenated with $h_{CG}^{(t)}(v)$, and processed by M_{CG}^{t}, a MLP, to produce a set of neighborhood-aware embeddings. Then, multiple aggregators with degree-scalers denoted by $\bigoplus$ operate on the set of MLP embeddings to extract a set of multivariate information that expresses the neighborhood distribution of node v. Finally, the set of information is concatenated to produce the aggregated message $a_{CG}^{(t+1)}(v)$ for node v. Afterwards, $a_{CG}^{(t+1)}(v)$ and $h_{CG}^{(t)}(v)$ are concatenated and processed by U_{CG}^{t}, a MLP, to update the node embedding, i. e., $h_{CG}^{(t+1)}(v)$. Details of $\bigoplus$ are

$$\begin{aligned} \bigoplus &= \Big[I, \mathcal{S}(D, \alpha=1), \mathcal{S}(D, \alpha=-1)\Big] \bigotimes \Big[\mu, \sigma, \max, \min\Big] \\ \mathcal{S}(D, \alpha) &= \frac{\log\,(D+1)^{\alpha}}{\delta}, \quad \delta = \frac{1}{|V_{train}|} \sum_{i \in V_{train}} \log\,(d_i + 1) \end{aligned} \tag{4.4}$$

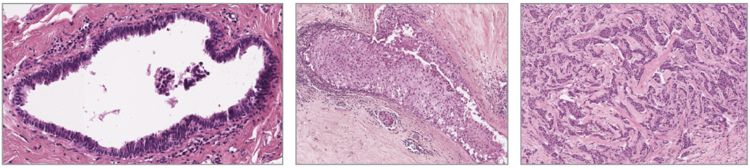
where I is identity matrix, S is degree-scaler matrix, D is node degree matrix, δ is normalization constant, α is scaling variable, and V_{train} is nodes in the training dataset. $[I, \mathcal{S}(D, \alpha=1), \mathcal{S}(D, \alpha=-1)]$ and $\Big[\mu, \sigma, \max, \min\Big]$ denote the list of scalers and the list of aggregators, respectively. The aggregators compute statistics on neighboring multiset of nodes, and the injective scalers discriminate between the multisets of various sizes. $\alpha = \{-1, 0, 1\}$ controls the attenuation, no scaling, or amplification of the scaling, respectively. $\bigotimes$ denotes tensor product between scalers and aggregators, and produces twelve operations that extract the set of multivariate information. The schematic diagram of a PNA layer is shown in Figure 2.4.

After T_{CG} PNA layers, an LSTM-based jumping knowledge technique ([Xu et al., 2018]) is employed to adapt to different CG sub-graph structures, i. e.,

$$h_{\text{CG}}^{(T_{\text{CG}}+1)}(v) = \text{LSTM}\left(\left\{ h_{\text{CG}}^{(t)}(v) \,\middle|\, t = 1, \ldots, T_{\text{CG}} \right\}\right) \tag{4.5}$$



Normal Benign Usual Ductal Hyperplasia Atypical Ductal Hyperplasia



Flat Epithelial Atypia Ductal Carcinoma in Situ Invasive Carcinoma

Figure 4.4: Samples of class-wise tumor regions-of-interest in BRACS dataset.

Following the CG-GNN, the cell-node embeddings, $h_{\mathrm{CG}}^{T_{\mathrm{CG}}+1}(v) \mid v \in V_{\mathrm{CG}}$, and the assignment matrix $A_{\mathrm{CG}\to\mathrm{TG}}$ are used to incorporate hierarchical information and initialize the tissue-node features in the TG, i. e.,

$$h_{\mathrm{TG}}^{(0)}(w) = \mathrm{CONCAT}\left(H_{\mathrm{TG}}(w), \sum_{v \in \mathcal{M}(w)} h_{\mathrm{CG}}^{(T_{\mathrm{CG}}+1)}(v) \right) \tag{4.6}$$

where CONCAT denotes concatenation and $\mathcal{M}(w) := \{ v \in V_{\mathrm{CG}} \mid A_{\mathrm{CG}\to\mathrm{TG}}(v,w) = 1 \}$ is the set of nodes in G_{CG} mapping to a node $w \in V_{\mathrm{TG}}$. Analogous to Equation (4.3), G_{TG} is processed by TG-GNN to compute tissue-node embeddings $h_{\mathrm{TG}}^{(t)}(w)$, $\forall w \in V_{\mathrm{TG}}$. At $t = T_{\mathrm{TG}}$, the embedding of each tissue-node w encodes the cell and tissue information up to T_{TG}-hops from w.

Similar to CG, the tissue-node embeddings in TG are processed via an LSTM-based jumping knowledge technique to combine the intermediate tissue-node embeddings. Finally, the graph-level embedding h_{HACT} is produced by summing all the tissue-node embeddings. An MLP and a softmax operation follows to map h_{HACT} to respective TRoI label. HACT-Net is trained end-to-end by minimizing the cross-entropy loss between the softmax output and the ground-truth TRoI label.

Following [Dwivedi et al., 2020], after each PNA layer we include graph normalization (GraphNorm) followed by a batch normalization (BatchNorm). Graph normalization scales the node features by the number of nodes in the graph. Intuitively, it prevents the node representations from being at different scales, for graphs of different sizes. This normalization helps the network to learn discriminative topological patterns when the number of nodes vary significantly within a class.

4.4 Datasets

BRACS dataset: As part of this work, we introduce a new dataset termed as BReAst Cancer Subtyping (**BRACS**). It contains 4391 TRoIs from 325 H&E breast carcinoma WSIs. The WSIs were selected from the archives of the Department of Pathology at

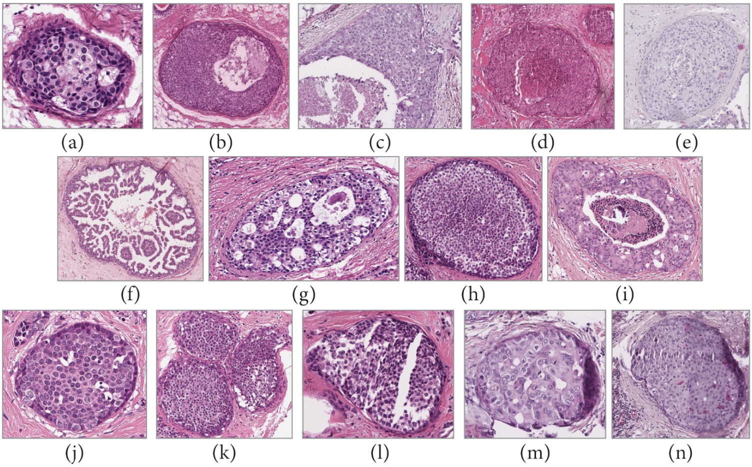


Figure 4.5: Overview of the variability for DCIS category in BRACS. The samples depict variations in, (a, b, c) tumor size, (d, e) staining appearance, sub-patterns: (f) low-grade Papillary, (g) moderate-grade Cribriform, (h, i) high-grade Solid and Comedo, (j, k) number of glandular regions per TRoI, and artifacts due to tissue and slide preparation: (l) tissue-folding or tear, (m) ink stain, (n) blur. Similar variability also persists in other categories in BRACS.

National Cancer Institute- IRCCS-Fondazione Pascale, Naples, Italy. They are scanned with an Aperio AT2 scanner at 0.25 μm/pixel resolution. The TRoIs were selected and annotated using QuPath ([Bankhead et al., 2017]) as being Normal, Benign, Usual ductal hyperplasia (UDH), Atypical Ductal Hyperplasia (ADH), Flat Epithelial Atypia (FEA), Ductal Carcinoma In Situ (DCIS), and Invasive. Figure 4.4 presents sample TRoIs from all cancer subtypes in BRACS. Each TRoI was first annotated independently by three pathologists. TRoIs with disagreement were further discussed and annotated by the consensus. Note that the pathologists used the entire WSI context during annotation. Figure 4.5 presents some DCIS samples in BRACS dataset, and highlight the included appearance variability. Such TRoI variability is typical in practice, and were included in BRACS to mimic the real world diagnosis. It ensures a realistic and representative evaluation set, with results readily applicable in the field.

Table 4.1 presents category-wise statistics of the TRoIs in BRACS. The statistics demonstrate a high variation in TRoI dimensions. We also include the statistics for the CG and TG representations constructed by our framework, which indicate a large variation in the size of the entity-graph representations. For evaluations on BRACS, we partition the TRoIs into train, validation, and test sets at the WSI-level, such that two TRoIs from the same WSI do not fall in different sets. The WSI-level splitting was performed randomly, ensuring a comparable number of TRoIs per cancer subtype. Such partitioning aimed for a fair evaluation of the compared methods.

BACH dataset: We evaluated the proposed methodology also on the publicly available microscopy image dataset, i. e., the Grand Challenge on BreAst Cancer Histology images **BACH** ([Aresta et al., 2019]). It consists of 400 training and 100 test images from four breast cancer subtypes, i. e., Normal, Benign, DCIS, and Invasive. All images are acquired using a Leica DM 2000 LED microscope and a Leica ICC50 HD camera. These images

Table 4.1: Key statistics of the BRACS dataset.

	Metric	Normal	Benign	UDH	ADH	FEA	DCIS	Invasive	Total
Image	Number of images	512	758	471	568	783	749	550	4391
	#pixels (in million)	2.8±2.7	5.7±4.5	2.4±2.9	2.2±2.0	1.2±1.1	5.0±5.0	8.2±5.4	3.9±4.3
	Max/Min pixel ratio	75.3	97.9	180.1	75.3	58.3	128.6	62.4	235.6
CG	Number of nodes	994±732	1826±1547	903±910	863±730	470±352	1723±1598	3609±2393	1468±1642
	Number of edges	3759±2643	6103±5420	3371±3675	3098±2781	1738±1395	5728±5811	12490±10011	5102±6089
	Max/Min node ratio	71.9	126.6	133.3	104.2	45.2	161.3	113.6	256.4
TG	Number of nodes	107±106	217±233	88±93	100±91	45±32	225±217	423±317	172±217
	Number of edges	509±545	1012±1236	393±450	480±474	194±159	1111±1123	2025±1741	815±1125
	Max/Min node ratio	169.5	312.5	125.0	178.6	416.7	312.5	101.0	434.8
Image split	Train	342	586	303	405	599	562	366	3163
	Validation	86	87	88	77	85	97	82	602
	Test	84	85	80	86	99	90	102	626
WSI split	Train	67	86	59	38	37	33	41	198
	Validation	28	24	24	28	17	21	19	68
	Test	15	16	20	17	12	16	16	59

are in RGB TIFF format and have a fixed size of 2048×1536 pixels and a pixel scale of $0.42 \times 0.42\,\mu$m. Notably, BRACS presents three major advantages over BACH:

- **Number of images**: The train and test sets of BRACS are nearly 10 times and 6 times the size of the train and test sets of BACH, respectively. The large test set ensures a robust evaluation of the methods.
- **Diverse subtypes**: BRACS includes diagnostically complex pre-cancerous atypical (ADH and FEA) categories, which represent a major diagnostic dilemma typical in practice due to their high risk of progressing to cancer. The seven cancer subtypes in BRACS represent a broad spectrum of breast cancer in histopathology.
- **Large variability**: The aforementioned high variability in BRACS in terms of TRoI appearances and dimensions is clinically more representative, and corresponds to a more realistic scenario of breast cancer subtyping.

4.5 Results

In this section, we comparatively assess the proposed method for breast cancer subtyping. First, we introduce state-of-the-art CNN and GNN baselines, and their implementation schemes. Second, we conduct ablations on BRACS to examine the impact of various components in our framework. Third, we evaluate the classification performance of our method and compare with the baselines, on BRACS and BACH datasets for different classification settings. Finally, we include a comparison of HACT-Net with three independent expert-pathologists.

4.5.1 Baselines

• **Single-scale CNN** processes TRoIs at a single magnification. A CNN is trained to predict patch-wise cancer subtypes, and we aggregate the patch-wise predictions to produce a TRoI-level prediction. We experiment with images at three magnifications, i. e., 10×, 20×, and 40×, denoted herein as CNN(10×), CNN(20×), and CNN(40×), using the same network architecture and training scheme. For each scale, we extract patches of size 128×128 pixels with a stride of 64 pixels. The CNN follows the single-scale training procedure by [Sirinukunwattana et al., 2018], and patch-wise predictions are aggregated using the Agg-Penultimate strategy by [Mercan et al., 2019a]. We use transfer learning with a ResNet-50 architecture, pre-trained on ImageNet, as the CNN backbone. Following feature extraction by ResNet-50, a two-layer MLP with 128 channels classifies the patches. To improve the classification, the ResNet-50 parameters are fine-tuned. Adam optimizer ([Kingma et al., 2015]) with 10^{-3} learning rate, a batch size of 16, and a dropout of 0.2 is used to optimize the categorical cross-entropy objective.

- **Multi-scale CNN** processes the TRoIs at multiple scales. We extract concentric patches of size 128×128 pixels from multiple magnifications and follow the “Late fusion with single-stream + LSTM" training procedure from [Sirinukunwattana et al., 2018]. We operate at two settings, i. e., (10×+20×) and (10×+20×+40×), and denote by prepending Multi-scale CNN in front of each. The patch-wise predictions are aggregated using the Agg-Penultimate strategy by [Mercan et al., 2019a]. On the concatenated features from the LSTM, we use a two-layer MLP of 128 channels to classify the patches. The training strategy and hyperparameters are the same as Single-scale CNN.

- **CGC-Net** is the Cell Graph Convolutional Network (CGC-Net) proposed by [Zhou et al., 2019a], and it is the state-of-the-art in classifying CG representations for TRoIs. We construct the CG topology for a TRoI using thresholded kNN strategy presented in Section 4.3.3.1. We initialize the CG nodes with hand-crafted features, employ the Adaptive GraphSage-based CGC-Net architecture, and follow the training strategy proposed by [Zhou et al., 2019a].

- **Patch-GNN** implements the method proposed by [Aygüneş et al., 2020], which is the state-of-the-art GNN method for classifying patch-graph representations of TRoIs. It incorporates local inter-patch context through a GNN to construct a graph-level features, which is then processed by an MLP to classify the TRoIs. We experiment with Patch-GNN at three scales, i. e., 10×, 20×, and 40×, denoted herein as Patch−GNN(10×), Patch−GNN(20×), and Patch−GNN(40×). At each magnification, we extract patches of size 128×128 to construct a TRoI-specific patch-graph. We employ the network architecture and training strategy proposed by [Aygüneş et al., 2020].

- **CG-GNN** is provided as a standalone CG-based learning baseline, to compare with our proposed hierarchical learning. CG-GNN uses PNA layers, an LSTM-based jumping knowledge, sum readout, and a two-layer MLP classifier. We follow the CG representation strategy as described in Section 4.3.3.1.

- **TG-GNN** is provided as a standalone TG-based learning baseline, to compare with our proposed hierarchical learning. TG-GNN uses the same architecture as the CG-GNN, with the node features directly initialized by H_{TG} instead of Equation (4.6).

- **CONCAT-GNN** is provided to evaluate the impact of hierarchical graph representation and learning. CONCAT-GNN utilizes standalone CG and TG representations, respectively, as input to standalone CG-GNN and TG-GNN to produce h_{CG} and h_{TG} graph-level embeddings. The TRoI level embedding is constructed by concatenating the graph-level embeddings, i. e., $h_{\text{CONCAT}} = \text{CONCAT}(h_{\text{CG}}, h_{\text{TG}})$. Finally, a two-layer MLP classifies h_{CONCAT} into a cancer subtype.

4.5.2 Implementation

Graph representations: CG representations (Section 4.3.3.1) use, i) patches of size 72×72, and ii) a CNN of ResNet-34 or ResNet-50 to initialize the node features. TG representations (Section 4.3.3.2) use, i) patches of size 144×144, and ii) a CNN of ResNet-34 or ResNet-50 to initialize the node features.

Graph architecture and learning: CG-GNN, TG-GNN, CONCAT-GNN, and HACT-Net all share the same options and hyperparameters below,

- Number of PNA layers in GNN: [3, 4, 5]
- Number of MLP layers in a PNA layer: 2
- Number of channels in a PNA-layer MLP: 64
- Graph-level embedding dimension: 128
- Number of MLP layers in output classifier: 2
- Number of channels in output MLP classifier: 128
- Training parameters: Adam optimizer ([Kingma et al., 2015]) with a learning rate of 10^{-3}, batch size of 16, and a categorical cross-entropy objective.

Evaluation metrics: Considering the imbalanced number of TRoIs per class in train, validation, and test sets (see Table 4.1), we evaluate the classification performance using weighted F1-score, an average weighted by the number of true instances for each class. The best weighted F1-scores on the validation set is used as the model selection criteria during the training of each method. To present any sensitivity to initialization, we report the mean and standard deviation of each model on the test set by training them three times using random weight initialization. Further, we present precision, recall, and confusion matrices to indicate the distribution of class predictions.

Computational resources: All the experiments were conducted using PyTorch ([Paszke et al., 2019]) and Deep Graph Library (DGL) ([Wang et al., 2019b]), on NVIDIA Tesla P100 GPUs and POWER8 processors.

4.5.3 Ablation studies

We conduct ablation to evaluate the impact of three major components of our methodology on TRoI classification performance, i. e., i) node feature initialization, ii) GNN layer type, and iii) jumping knowledge technique. Each component is analyzed individually, while fixing the others. Ablations are performed on BRACS for classifying the TRoIs into 7-classes.

4.5.3.1 Impact of node feature initialization

The performance of GNNs eminently rely on the initial node features ([Kipf et al., 2017]). In our context, we analyze the impact of initial morphological features of the nodes with the following three feature initialization schemes:

Table 4.2: Ablation: Impact of node features. Mean and standard deviation of 7-class weighted F1-scores. Results expressed in %.

	Weighed F1
CG-GNN: No morphological features	45.24±1.5
CG-GNN: Hand-crafted morphological features	48.34±5.2
CG-GNN: CNN morphological features	**55.94±1.0**
TG-GNN: No morphological features	36.81±0.7
TG-GNN: Hand-crafted morphological features	51.62±2.1
TG-GNN: CNN morphological features	**56.62±1.3**
CONCAT-GNN: No morphological features	47.62±1.6
CONCAT-GNN: Hand-crafted morphological features	51.55±1.3
CONCAT-GNN: CNN morphological features	**57.01±2.3**
HACT-Net: No morphological features	48.70±0.2
HACT-Net: Hand-crafted morphological features	52.46±0.2
HACT-Net: CNN morphological features	**61.53±0.9**

Table 4.3: Ablation: Impact of jumping knowledge. Mean and standard deviation of 7-class weighted F1-scores. Results expressed in %.

	Weighed F1
CG-GNN: No aggregator	55.53±0.8
CG-GNN: Concatenation	55.82±1.0
CG-GNN: LSTM	**55.94±1.0**
TG-GNN: No aggregator	55.30±0.8
TG-GNN: Concatenation	56.07±0.8
TG-GNN: LSTM	**56.62±1.3**
CONCAT-GNN: No aggregator	**57.67±4.5**
CONCAT-GNN: Concatenation	56.28±2.7
CONCAT-GNN: LSTM	57.01±2.3
HACT-Net: No aggregator	49.16±1.1
HACT-Net: Concatenation	59.78±1.6
HACT-Net: LSTM	**61.53±0.9**

- **No morphological features:** The nodes of an entity-graph are initialized with only the spatial features. Experiments with this setting demonstrate the impact of standalone graph topology on the classification performance.
- **Hand-crafted morphological features:** The entity-graph nodes are initialized with hand-crafted morphological features as suggested by [Zhou et al., 2019a], i. e., i) *texture features*: difference of average foreground to background; standard deviation, skewness, and mean entropy of intensities; dissimilarity, homogeneity, energy, and angular second moment from Gray-Level Co-occurrence Matrix; and ii) *shape features*: eccentricity, area, maximum and minimum axis lengths, perimeter, solidity, and orientation. Note that, the hand-crafted features for CG and TG are computed, respectively, from the segmented instances of nuclei and tissue regions.
- **CNN morphological features:** The morphological features of the entity-graph nodes are initialized with CNN features (ResNet-34 pre-trained on ImageNet) extracted from patches around the centroids of the nuclei and tissue regions.

Experimental results in Table 4.2 indicate that the standalone CG topology is more discriminative for cancer subtyping than TG topology. The combination of CG and TG topologies further improves discriminative ability. The best performance achieved with the HACT topology confirms the strength of hierarchical representations. Further, including morphological features significantly improves the classification. The superiority of graphs with CNN-based morphological features indicate the richness of morphological information acquired by CNNs, compared to hand-crafted measures.

4.5.3.2 Impact of GNN layer type

We investigate the impact of two state-of-the-art GNN layers, i. e., GIN and PNA on the classification performance. The experiments use CNN-based node feature initialization

Table 4.4: Ablation: Impact of GNN layers. Mean and standard deviation of 7-class weighted F1-scores. Results expressed in %.

	Weighed F1
CG-GNN: GIN	55.70±0.5
CG-GNN: PNA	**55.94±1.0**
TG-GNN: GIN	55.33±1.4
TG-GNN: PNA	**56.62±1.3**
CONCAT-GNN: GIN	56.20±2.1
CONCAT-GNN: PNA	**57.01±2.3**
HACT-Net: GIN	59.73±1.2
HACT-Net: PNA	**61.53±0.9**

and LSTM-based jumping knowledge. Results in Table 4.4 demonstrate that GNNs with PNA layers outperform GNNs with GIN layers, for all the four GNN constructions.

4.5.3.3 Impact of jumping knowledge technique

To investigate the impact of jumping knowledge, we experiment with three settings: no jumping knowledge, CONCAT-based, and LSTM-based. LSTM-based technique follows Equation (4.5). Based on this, CONCAT-based technique replaces the LSTM operation with concatenation. The experiments use CNN-based node feature initialization and PNA layers. Results in Table 4.3 show a generally positive impact of the jumping knowledge. Compared to CONCAT, the LSTM-based technique learns better dependencies between GNN layers, thus generates better graph embeddings.

4.5.3.4 Ablation summary

The ablation experiments conclude the following choice of components for designing our methodology, i) CNN-based initialization of node-level morphological features, ii) use of PNA layers, and iii) an LSTM-based jumping knowledge technique.

4.5.4 Classification results on BRACS dataset

We evaluate our proposed methods, comparatively with CNN and GNN baselines. To analyze the performance for different clinical applications and histopathological needs, we evaluate and report the results separately in the following three settings:

4.5.4.1 Setting 1: 7-class classification

Here, we classify the TRoIs into 7-classes, i. e., Normal, Benign, UDH, ADH, FEA, DCIS, and Invasive, for the differentiation of a large spectrum of breast cancer subtypes. Table 4.5 tabulates the classification performance of the compared methods.

Among single-scale CNNs, CNN(10×) performs the best, indicating the importance of global context information for TRoI classification. Multi-scale CNNs using both global and local context outperform single-scale CNNs. Such benefit from context is significant for ADH, FEA, and DCIS categories, which all require both local and global context for the

Table 4.5: Mean and standard deviation of per-class F1-scores and weighted F1-scores for 7-class classification setting. Results are expressed in %. The best result is in **bold** and the second best is underlined.

	Method	Normal	Benign	UDH	ADH	FEA	DCIS	Invasive	Weighted F1
CNN	CNN (10×)	48.67±1.7	44.33±1.9	45.00±5.0	24.00±2.8	47.00±4.3	53.33±2.6	86.67±2.6	50.85±2.6
	CNN (20×)	42.00±2.2	42.33±3.1	39.33±2.0	22.67±2.5	47.67±1.2	50.33±3.1	77.00±1.4	46.85±2.2
	CNN (40×)	32.33±4.6	39.00±0.8	23.67±1.7	18.00±0.8	37.67±2.9	47.33±2.0	70.67±0.5	39.41±1.9
	Multi-scale CNN (10 × +20×)	48.33±2.0	45.67±0.5	41.67±5.0	32.33±0.9	46.33±1.4	59.33±2.0	85.67±1.9	52.27±1.9
	Multi-scale CNN (10 × +20 × +40×)	50.33±0.9	44.33±1.2	41.33±2.5	31.67±3.3	51.67±3.1	57.33±0.9	86.00±1.4	52.83±1.9
GNN	CGG-Net	30.83±5.3	31.63±4.7	17.33±3.4	24.50±5.2	58.97±3.6	49.36±3.4	75.30±3.2	43.63±0.5
	Patch-GNN (10×)	52.53±3.3	47.57±2.2	23.67±4.6	30.66±1.8	60.73±5.3	58.76±1.1	81.63±2.2	52.10±0.6
	Patch-GNN (20×)	43.86±4.2	43.37±3.2	19.47±2.3	25.73±2.9	55.57±2.1	52.86±1.8	79.20±1.0	47.10±0.7
	Patch-GNN (40×)	41.70±3.1	32.93±1.0	25.07±3.7	25.63±2.0	49.47±3.5	48.60±4.2	71.57±5.1	43.23±0.6
Ours	CG-GNN	58.77±6.8	40.87±3.0	**46.82±1.9**	39.99±3.6	63.75±10.5	53.81±3.9	81.06±3.3	55.94±1.0
	TG-GNN	**63.59±4.9**	**47.73±2.9**	39.41±4.7	28.51±4.3	72.15±1.3	54.57±2.2	82.21±4.0	56.62±1.3
	CONCAT-GNN	60.97±4.5	43.06±2.3	41.96±4.7	26.10±3.7	71.29±2.1	60.83±3.7	85.42±2.7	57.01±2.3
	HACT-Net (Proposed)	61.56±2.1	47.49±2.9	43.60±1.9	**40.42±2.5**	**74.22±1.4**	**66.44±2.6**	**88.40±0.2**	**61.53±0.9**

diagnosis. Multi-scale CNNs also outperform CGC-Net and Patch-GNNs. Interestingly, at each magnification, Patch-GNN outperforms single-scale CNN, which affirms the importance of relational and topological information incorporated in the graphs.

Comparing our proposed GNN solutions, we observe that CG-GNN significantly outperforms CGC-Net, indicating the superiority of CNN-based node feature initialization over handcrafted features, and the significance of GNNs with expressive PNA layers over Adaptive GraphSage in CGC-Net. We notice that CG-GNN and TG-GNN provide comparable performance overall. However, they outperform each other for Normal, Benign, UDH, ADH, and FEA categories, displaying the importance of complementary information captured by standalone TG and CG representations. Further, both HACT-Net and CONCAT-GNN provide overall superior performance compared to all CNN and GNN baselines. HACT-Net significantly outperforms CONCAT-GNN showing the significance of hierarchical modeling and learning. CONCAT-GNN produces overall comparable or superior performance to CG-GNN and TG-GNN, although for individual classes, CONCAT-GNN is rarely better than the two, suggesting that it may be using complementary information from CG and TG. Such complementary information is indeed best utilized by HACT-Net, with high per-class and overall classification performance. Though HACT-Net achieves the third best result for the UDH category, it uses the complementarity of CG and TG to provide better classification than TG-GNN. Moreover, the misclassified UDH samples are predominantly categorized as Benign due to the expected ambiguity between Benign and UDH classes. All the proposed GNNs often outperform all CNN baselines, establishing the potential of entity-based analysis.

Figure 4.6 presents per-class precision and recall for CG-GNN, TG-GNN, CONCAT-GNN, and HACT-Net. HACT-Net produces the highest precision values for most of the classes. The recall ranking between CG-GNN and TG-GNN varies across classes, whereas HACT-Net consistently yields good recall values. Further, standard deviation

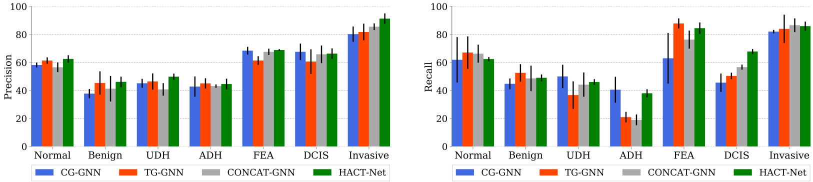


Figure 4.6: Mean and standard deviation of per-class precision and recall for 7-class classification.

of class-wise precision and recall values are the lowest for HACT-Net, for most classes. Figure 4.7 presets row-normalized aggregated 7-class confusion matrix across three runs for HACT-Net. It indicates ambiguities between i) Normal and Benign, ii) UDH and ADH, and iii) ADH and DCIS. Notably, these pair-wise classes bear high pathological ambiguity and are diagnostically very challenging.

4.5.4.2 Setting 2: 4-class classification

This setting categorizes TRoIs into 4-classes as per cancer risk: Normal, Non-cancerous (Benign + UDH), Precancerous (ADH + FEA), and Cancerous (DCIS + Invasive). Classification performance of CNN and GNN baselines, and HACT-Net are presented in Table 4.6. Single scale CNNs exhibit the same behavior as before. However, combining multiple magnifications in multi-scale CNNs does not improve the classification over the single-scales. Among the baselines, CGC-Net and Patch-GNNs perform comparable or inferior to the CNNs, with a low-magnification CNN(10×) outperforming the others. Similarly to the 7-class setting, our proposed methods are superior to the baselines. HACT-Net produces the best overall performance, with the best classification performance for Normal, Precancerous, and Cancerous categories. To highlight, HACT-Net achieves $\approx$ 66% F1-score for the diagnostically challenging Precancerous category.

4.5.4.3 Setting 3: Binary classifications

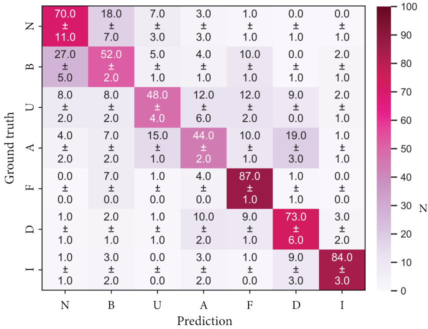
In this setting, we replicate the typical decision process of a pathologist for breast cancer subtyping which follows a diagnostic decision tree as presented in Figure 4.8. It is inspired by the classification scheme presented by [Mercan et al., 2018]. Note that such individual binary decisions are less constrained compared to multi-class classification, thus allows for better discrimination between a selected pair of classes. The binary classifiers can assist pathologists in categorizing ambiguous cases at different bifurcations of the decision tree. Table 4.7 presents the results for six individual binary classifications, at the bifurcations in the decision tree. Results are consistent with the 7-class and 4-class classification settings, with HACT-Net consistently outperforming all baselines and providing the best aggregated score.

4.5.4.4 Domain expert comparison on BRACS dataset

To further benchmark our proposed methodology as well as to assess the quality of the introduced BRACS dataset, we acquired annotations of the BRACS test set from additional independent pathologists. For such comparison with domain experts, we follow

Table 4.6: Mean and standard deviation of per-class F1-scores and weighted F1-scores for 4-class classification setting. Results are expressed in %. The best result is in **bold** and the second best is underlined.

	Method	Normal	Non-cancerous	Precancerous	Cancerous	Weighted F1
CNN	CNN (10×)	54.33±3.7	56.00±0.8	56.33±1.2	83.67±0.9	64.36±1.4
	CNN (20×)	45.33±4.6	55.33±0.4	52.33±1.9	81.67±2.0	61.18±1.9
	CNN (40×)	42.00±4.9	51.00±0.8	47.67±4.1	77.67±2.0	56.99±2.7
	Multi-scale CNN (10 × +20×)	51.67±5.8	55.33±1.2	52.67±2.9	80.67±1.9	61.82±2.5
	Multi-scale CNN (10 × +20 × +40×)	51.33±3.3	56.33±2.0	57.00±1.6	81.33±3.7	63.52±2.6
GNN	CGG-Net	34.53±2.9	47.23±3.7	62.90±2.8	82.20±1.0	59.87±2.3
	Patch-GNN (10×)	53.13±4.4	46.23±2.4	63.96±3.8	77.43±3.2	61.93±2.5
	Patch-GNN (20×)	53.46±1.8	47.16±2.8	63.20±3.8	74.90±3.4	61.26±2.9
	Patch-GNN (40×)	40.90±2.7	38.67±2.8	56.77±3.9	72.20±2.6	54.60±1.9
Ours	CG-GNN	52.95±12.1	56.55±3.7	61.53±3.0	84.47±0.9	66.10±2.6
	TG-GNN	52.96±6.8	56.52±2.8	64.36±1.0	82.21±0.8	66.24±1.1
	CONCAT-GNN	54.54±1.6	**56.63±1.7**	62.58±1.4	81.80±0.8	65.83±0.1
	HACT-Net (Proposed)	**66.08±3.7**	55.28±1.7	**66.21±0.9**	**84.91±0.8**	**69.04±0.5**



Ground truth \ Prediction	N	B	U	A	F	D	I
N	70.0 ± 11.0	18.0 ± 7.0	7.0 ± 3.0	3.0 ± 3.0	1.0 ± 1.0	0.0 ± 1.0	0.0 ± 1.0
B	27.0 ± 5.0	52.0 ± 2.0	5.0 ± 1.0	4.0 ± 1.0	10.0 ± 1.0	0.0 ± 1.0	2.0 ± 1.0
U	8.0 ± 2.0	8.0 ± 2.0	48.0 ± 4.0	12.0 ± 6.0	12.0 ± 2.0	9.0 ± 0.0	2.0 ± 1.0
A	4.0 ± 2.0	7.0 ± 2.0	15.0 ± 1.0	44.0 ± 2.0	10.0 ± 1.0	19.0 ± 3.0	1.0 ± 1.0
F	0.0 ± 0.0	7.0 ± 0.0	1.0 ± 0.0	4.0 ± 0.0	87.0 ± 1.0	1.0 ± 1.0	0.0 ± 0.0
D	1.0 ± 1.0	2.0 ± 1.0	1.0 ± 1.0	10.0 ± 2.0	9.0 ± 1.0	73.0 ± 6.0	3.0 ± 2.0
I	1.0 ± 1.0	3.0 ± 2.0	0.0 ± 0.0	3.0 ± 2.0	1.0 ± 0.0	9.0 ± 3.0	84.0 ± 3.0

Figure 4.7: Mean and standard deviation of row-normalized 7-class confusion matrix.

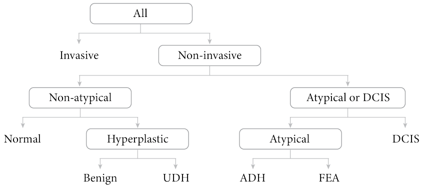


Figure 4.8: Binary decision tree used by pathologists for breast cancer diagnosis.

the evaluation protocol in [Elmore et al., 2015]. We recruited three board-certified pathologists (other than the original three pathologists who provided the initial annotations, namely our ground truth labels), from three different medical centers, to further ensure independence: • National Cancer Institute- IRCCS-Fondazione Pascale, Naples, Italy; • Lausanne University Hospital, CHUV, Lausanne, Switzerland; and • Aurigen, Centre de Pathologie, Lausanne, Switzerland. These experts are specialized in breast pathology and have been in practice for over twenty years. The pathologists independently and remotely annotated BRACS test set TRoIs, without having access to respective WSIs. This protocol ensures equal field-of-view for all the pathologists as well as our methodology.

The independent pathologists' annotations are compared to the ground truth, with the results shown in Table 4.8. We present per-class F1-scores, overall weighted F1-score, and overall weighted accuracy for each pathologist. We also include the aggregated statistics

Table 4.7: Mean and standard deviation of weighted F1-scores for binary classification setting. Further, the aggregated mean and standard deviation for the six binary tasks are reported. Results are expressed in %. The best result is in **bold** and the second best is underlined.

	Method	I vs N+B+A+U+F+D	N+B+U vs A+F+D	N vs B+U	B vs U	A+F vs D	A vs F	Aggregated
CNN	CNN (10×)	95.66±0.5	81.24±0.4	69.83±0.4	76.12±1.1	73.44±2.6	77.59±1.7	78.90±1.4
	CNN (20×)	92.39±0.4	80.84±0.4	66.52±2.1	74.75±1.5	67.87±1.8	71.78±2.5	75.69±1.7
	CNN (40×)	90.74±0.6	79.92±1.7	62.36±2.1	68.13±4.3	64.86±3.0	66.91±1.7	72.15±2.5
	Multi-scale CNN (10 × +20×) (10 × +20×)	94.31±1.3	80.89±1.3	67.99±1.9	75.58±2.1	72.07±1.8	76.91±2.2	77.96±1.8
	Multi-scale CNN (10 × +20 × +40×)	95.12±1.1	82.21±0.3	70.87±2.1	72.89±2.3	72.08±3.2	75.47±3.7	78.11±2.4
GNN	CGG-Net	91.60±2.1	79.73±1.5	63.67±3.1	62.37±3.0	81.56±1.6	73.80±5.4	75.46±3.1
	Patch-GNN (10×)	95.80±0.4	76.53±0.3	72.57±1.1	72.87±3.1	77.17±0.8	78.26±2.6	78.87±1.7
	Patch-GNN (20×)	93.70±0.4	76.63±1.4	70.10±1.9	69.77±3.1	74.10±0.1	81.03±1.8	77.55±1.8
	Patch-GNN (40×)	92.40±0.9	74.43±0.6	71.10±1.7	67.40±2.5	72.97±0.7	76.40±1.9	75.78±1.6
Ours	CG-GNN (Ours)	94.52±0.4	**83.79±0.3**	75.71±1.7	73.15±3.3	77.48±1.7	84.33±0.5	81.50±1.7
	TG-GNN	96.00±0.6	80.38±0.8	69.51±3.1	76.12±1.0	80.67±0.2	84.18±3.6	81.14±2.0
	CONCAT-GNN	95.91±0.6	83.21±0.7	71.84±1.5	75.67±1.8	80.14±2.6	88.88±3.9	82.61±2.1
	HACT-Net (Proposed)	**96.32±0.6**	83.63±0.7	**76.84±0.7**	**77.66±0.4**	**81.11±0.7**	**89.35±0.3**	**84.15±0.6**

of the three pathologists for benchmarking HACT-Net with domain experts. Table 4.8 indicates that HACT-Net outperforms the domain experts in distinguishing TRoIs of diagnostically challenging classes, i. e., atypia and hyperplasia, while yielding comparable performance for the normal and cancerous categories. Per-class standard deviations of pathologists' statistics show the expected high inter-observer variability in breast cancer diagnosis. Compared to the pathologists, HACT-Net yields a superior weighted accuracy and weighted F1 given the ground truth diagnoses for the 7-class classification.

To benchmark the BRACS dataset with respect to the dataset by [Elmore et al., 2015], we compare the aggregated pathologist statistics on both datasets for the same set of classes, i. e., Benign without atypia (Normal + Benign + UDH), Atypia (ADH + FEA), DCIS, and Invasive. Note that the dataset by [Elmore et al., 2015] consists of 240 breast biopsy slides, while BRACS consists of 626 TRoI images. For the dataset by [Elmore et al., 2015], class-wise concordance rates (class-weighted average accuracy of 115 pathologists to a three-expert consensus) are 87%, 48%, 84%, and 96%, respectively for the four aforementioned classes. For BRACS, the similar class-wise concordance rates are 87%, 50%, 72%, and 90%, respectively. The class-wise concordance rates exhibit a similar trend in both datasets. Differences can be attributed to differing fields-of-view, i. e., TRoI vs. WSI, accessible to the pathologist during annotation.

Table 4.9 presents the inter-observer concordance rates for the BRACS test set. We notice significant differences in the concordance rates between pathologists 2 vs. 3 and pathologist 1 vs. the other two. This can be reasoned to differing histopathology practices across different regions.

Table 4.8: Comparison between HACT-Net and domain expert pathologists for 7-class breast cancer subtyping on BRACS dataset. Per-class F1-scores, weighted F1-scores and accuracy for 7-class classification are presented. Results are expressed in %. The best results are in **bold**.

	Normal	Benign	UDH	ADH	FEA	DCIS	Invasive	Weighted F1	Weighted Acc
Pathologist 1	67.53	53.92	41.90	36.00	19.13	71.59	94.00	55.30	56.71
Pathologist 2	47.83	52.94	25.00	35.37	65.22	68.00	94.00	57.07	57.99
Pathologist 3	39.66	49.59	49.43	42.29	54.12	65.19	89.47	56.71	56.55
Pathologist stats	51.57±11.7	**52.15±1.8**	38.78±10.2	37.89±3.1	46.16±19.6	**68.26±2.6**	**92.49±2.1**	56.36±0.8	57.08±0.6
HACT-Net stats	**61.56±2.1**	47.49±2.9	**43.60±1.9**	**40.42±2.5**	**74.22±1.4**	66.44±2.6	88.40±0.2	**61.53±0.9**	**63.21±0.3**

Table 4.9: Concordance among three independent pathologists for annotating BRACS test dataset. Results are expressed in %.

	Pathologist 1	Pathologist 2	Pathologist 3	Ground truth
Pathologist 1	-	47.60	50.96	56.71
Pathologist 2	-	-	64.38	57.99
Pathologist 3	-	-	-	56.55

4.5.4.5 Computational time analysis

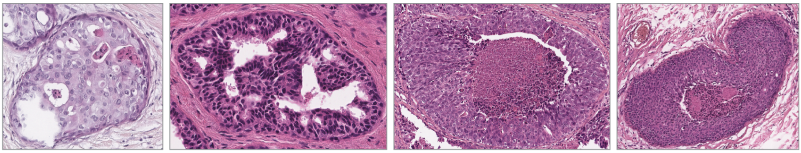
We report computation time for processing a tumor RoI of size 1000 × 1000 pixels on a single-core POWER8 processor combined with an NVIDIA P100 GPU. Stain normalization with the Macenko method takes 0.8 seconds (CPU-only), CG generation 2.51 seconds, and TG generation 4.14 seconds. Thus, the overall computational time for transforming the RoI into HACT representation is 7.92 seconds. The superpixel extraction step can be further optimized by using fast GPU implementations, e. g., as proposed by [Jampani et al., 2018]. Provided the HACT representation, HACT-Net renders near real-time inference by requiring 34.11 milliseconds. Additional run-time analysis is presented by [Jaume et al., 2021a].

4.5.5 Classification results on BACH dataset

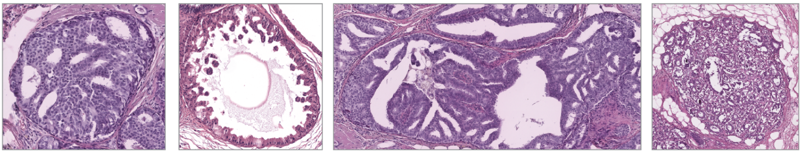
We evaluate the methods on the public BACH dataset. Considering its smaller training set of 400 images, we employ different image augmentation techniques for training HACT-Net. To this end, we use rotation, mirroring, and color augmentations on the training images before extracting HACT graph representations. We do not use other graph augmentation techniques, such as random node and edge dropping, as they may hamper the meaningful topological distribution of the biological entities. The implementation strategies and hyperparameters in Section 4.5.2 are employed for training HACT-Net. Classification performance of HACT-Net and the current state-of-the-art results on the BACH dataset are listed in Table 4.10. Our predictions have been evaluated independently by the organizers of the BACH challenge, ensuring a fair comparison. HACT-Net results in comparable classification accuracy with the state-of-the-art methods. The difference in the accuracies are not significant considering only 100 TRoIs in the test set. Notably, our methodology employs a single, unified network, where the other listed competitors employ an ensemble strategy with multiple networks during inference.

Table 4.10: Accuracy of 4-class breast cancer subtyping in BACH. Results are expressed in %.

	Methods	Accuracy
Ensemble networks ([Aresta et al., 2018] [Aresta et al., 2019])	Wang et al. (2019)	95.00
	[Marami et al., 2018]	94.00
	Yang et al. (2019)	93.00
	[Chennamsetty et al., 2018]	87.00
	Kwok et al. (2018)	87.00
	[Brancati et al., 2018]	86.00
Single network	HACT-Net	91.00



(a) [DCIS, Benign, DCIS] (b) [DCIS, UDH, DCIS] (c) [Invasive, DCIS, DCIS] (d) [ADH, DCIS, DCIS]



(e) [ADH, ADH, DCIS] (f) [Benign, Benign, DCIS] (g) [ADH, ADH, DCIS] (h) [ADH, Invasive, DCIS]

Figure 4.9: Qualitative comparison of CG-GNN, TG-GNN, and HACT-Net for 7-class classification. Predictions by the classifiers are noted below each example. Red and Green denote incorrect and correct classification, respectively. (a,b) TRoIs which TG-GNN misclassifies, while CG-GNN and HACT-Net classify correctly by using the nuclei characteristics. (c,d) TRoIs misclassified by CG-GNN, while correctly classified by TG-GNN and HACT-Net by using context information from necrotic regions. (e,f,g,h) TRoIs which both CG-GNN and TG-GNN misclassify, where HACT-Net classifies correctly by utilizing both cell and tissue microenvironments together.

4.5.6 Qualitative analysis

Qualitative assessment of a few TRoIs from the BRACS dataset using HACT-Net, CG-GNN, and TG-GNN is presented in Figure 4.9. In Figure 4.10, we use GRAPHGRADCAM ([Pope et al., 2019; Jaume et al., 2021b]), a post-hoc gradient based feature attribution technique, to highlight the nuclei and tissue-region nodes in CG and TG, respectively, which HACT-Net focuses on while classifying the TRoIs. Given the DCIS examples in Figures 4.10(a-c, g-i), HACT-Net is seen to focus on the diagnostically relevant tumorous epithelium and necrotic regions in TG, while ignoring the less important stroma and lumen, cf. Figures 4.10(b,h). Further, within the relevant tissue regions, HACT-Net focuses on a subset of tumorous epithelial nuclei in CG, as shown in Figures 4.10(c,i). Interestingly, we observe in Figures 4.10(h,i) that HACT-Net uses complementary information from the necrotic region captured by TG, but not by CG. Similar observations of HACT-Net considering the diagnostically relevant regions can be made for FEA and Benign examples shown in Figures 4.10(d-f, j-l). Noticeably, such feature attribution analysis of GNNs

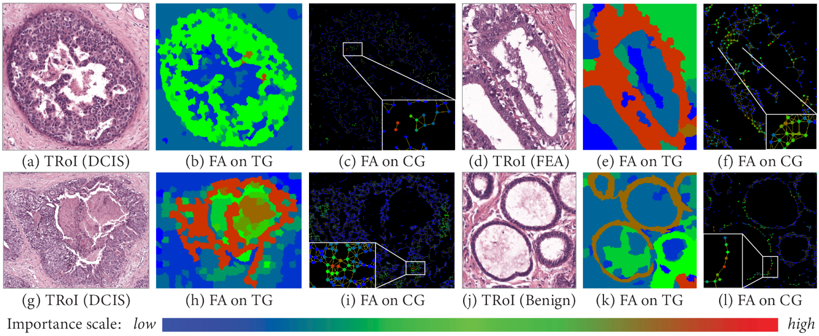


Figure 4.10: Feature attribution (FA) maps of HACT-Net on TG and CG for four sample TRoIs for 7-class classification: Sample TRoIs of (a,g) DCIS, (d) FEA, and (j) Benign classes, with their corresponding feature attribution maps on (b,h,e,k) TG and (c,i,f,l) CG.

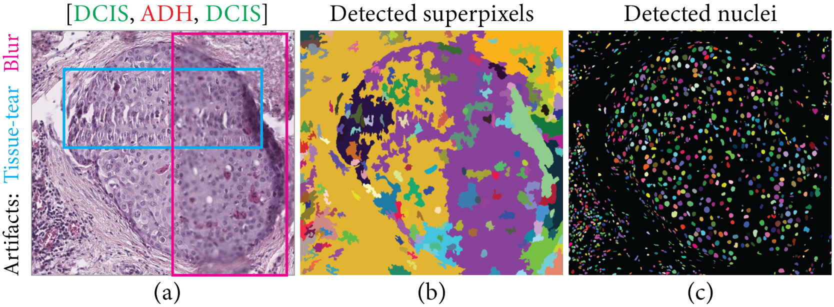


Figure 4.11: (a) A DCIS sample including tissue-tear and blur artifacts. (b) Detected superpixels. (c) Detected nuclei. The classifications by CG-GNN, TG-GNN and HACT-Net are indicated, where Red and Green denote incorrect and correct classification.

localizes and highlights the focus of deep networks in the given entity-paradigm, which is both more interpretable and more explainable compared to feature attribution strategies in a pixel-paradigm ([Jaume et al., 2020; Jaume et al., 2021b]). Interestingly, we also analyze the impact of tissue or slide preparation artifacts on the model performance. In Figure 4.11, we present a DCIS image with tissue-tear and blur artifacts. We observe that the detected superpixels do not aptly depict the tissue in the blur region, and consequently the TG-GNN using standalone TG misclassifies it. However, the nuclei detection is less impacted by the artifact, which allows the CG to appropriately encode the cell microenvironment and correctly classify the sample. To highlight, HACT-Net utilizing the complementary information from both CG and TG compensates for the issue in TG, and correctly identifies the subtype.

4.6 Conclusion

Pixel-based processing of pathology images suffers from the context-resolution trade-off, and misses the notion of biological entity and tissue composition. In this work, we propose an entity-based tissue representation and learning to address these issues. To that end, our two specific contributions are: (i) a hierarchical entity-graph representation of a tissue image by incorporating multisets of pathologically intuitive biological entities, and (ii) a hierarchical graph neural network for sequentially processing the entity-graph representation for mapping tissue compositions to tissue subtypes. Further, we introduce BReAst Cancer Subtyping (BRACS), a large cohort of breast tumor regions-of-interest, annotated with breast cancer subtypes. BRACS encompasses seven breast cancer subtypes to present a realistic breast cancer diagnosis scenario. Using BRACS as well as a public breast cancer subtyping dataset BACH, we demonstrate herein the superior performance of our proposed methodology for classifying breast tumor regions-of-interest into cancer subtypes. Under various experimental settings, our methodology is shown to outperform state-of-the-art pixel-based and entity-graph based classification approaches. Furthermore, we benchmark our methodology on the BRACS dataset by comparing it to three independent pathologists. Notably, our method achieves better performance for per-cancer subtype and overall aggregated classification. Although we have evaluated our method for breast cancer classification, the technology is easily extendable to other tissue types and diseases. Notably, the proposed hierarchical graph methodology can also be adapted to other image modalities, such as natural images, multiplexed images, hyperspectral images, satellite images, and other medical imaging domains, by utilizing domain and task-specific entities.

5

Towards Explainable Graph Representations in Digital Pathology

Explainability of deep learning (DL) techniques in digital pathology (DP) is of great significance to facilitate their wide adoption in clinics. Recently, graph techniques encoding relevant biological entities have been employed to represent and assess DP images. Such paradigm shift from pixel-wise to entity-wise analysis provides more control over concept representation. In this paper, we introduce a post-hoc explainer to derive compact per-instance explanations emphasizing diagnostically important entities in the graph. Although we focus our analyses to cells and cellular interactions in breast cancer subtyping, the proposed explainer is generic enough to be extended to other topological representations in DP. Qualitative and quantitative analyses demonstrate the efficacy of the explainer in generating comprehensive and compact explanations.

5.1 Introduction

Convolutional Neural Networks (CNNs), so far the most successful DL method in image analysis, have been widely adopted to assess DP images to improve diagnosis and patient outcome. However, concept representations of CNNs remain unexplained in DP and thus hinder their adoption in typical workflows. Therefore, explainable DL technologies in DP have become of paramount interest to build trust and promote the employment of DL in clinical settings [Holzinger et al., 2017].

Typically CNNs process complex and large DP images in a patch-wise manner, followed by aggregating the patch-wise learning to address downstream DP tasks. Recently, several research works have been devoted to demystify the concept representations of CNNs in automated diagnosis. Patch-level explainable methods [Graziani et al., 2018; Hägele et al., 2020; Korbar et al., 2017; Mobadersany et al., 2018; Cruz-Roa et al., 2013; Xu et al., 2017b] build patch-level *heatmaps*, where an importance score is computed per pixel to identify the regions of importance. For instance, [Hägele et al., 2020] use layer-wise relevance propagation [Bach et al., 2015] to generate positive scores for pixels that are positively correlated with the class label and negative scores otherwise. Such approaches have several limitations. First, pixel-level heatmaps fail to capture the spatial organization and

---

This chapter has been published as: Guillaume Jaume†, Pushpak Pati†, Antonio Foncubierta-Rodríguez, Jean-Philippe Thiran, Orcun Goksel, Maria Gabrani, "Towards Explainable Graph Representations in Digital Pathology", In: International Conference on Machine Learning Workshop on Computational Biology (CompBio), Vienna, Austria, Jul 2020, [arXiv]

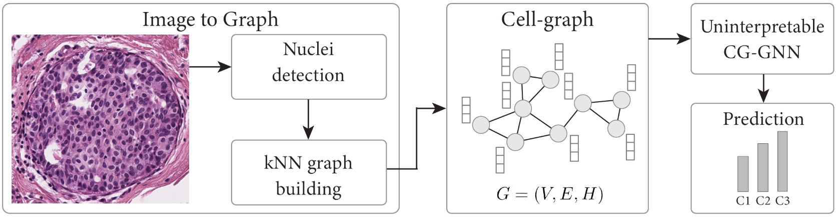


Figure 5.1: A TRoI is transformed into a CG, and is processed by a CG-GNN to predict the cancer subtype.

interactions of relevant biological entities. Second, the pixel-level analysis is completely detached from any biological reasoning that pathology guidelines recommend for decision making. Third, pixel-level explanation are common in the form of blurry heatmaps, which then do not allow to discriminate the relevance of nearby entities and their interactions.

Recently, graph techniques have been adopted to map DP images to graph representations and process such graphs for pathology tasks [Demir et al., 2004; Zhou et al., 2019a; Sharma et al., 2016; Anand et al., 2019; Wang et al., 2019a; Pati et al., 2020]. Graph representations embed biological entities and their interactions. To the best of our knowledge, explainability of *graph-based* approaches for DP has not been addressed yet. In this paper, a major step towards explainability in DP is presented based on two proposals: First, we advocate for shifting the analysis from a pixel-space to a relevant histological entity-space. The learning can then be regulated to specific entities and interactions, aligned with the prior pathological knowledge. Second, we propose to adopt an instance-level post-hoc explainability method that extracts a relevant subset of entities and interactions from the input graph. We define this subset as the explanation of our original entity-graph representation. We hypothesize that the explanation will be deemed useful if and when the subset aligns with prior pathological knowledge.

In this paper, we map DP images to cell-graphs [Demir et al., 2004], where cells and cellular interactions are represented as nodes and edges of the graph, and focus on the interpretability of cell-graphs towards cancer subtyping.

5.2 Methodology

In this section, we first present the extraction of graph representations from DP images, and further present the Graph Neural Network (GNN) framework for processing the representations. Second, we introduce the explainability module to acquire comprehensive explanations.

5.2.1 Cell-graph representation and learning

The DP images are transformed into cell-graph (CG) representations. Formally, we define a CG, $G_{CG} = (V, E, H)$ as an undirected graph composed of a vertices V and edges E. Each vertex is described by an embedding $h \in \mathbb{R}^d$, or equivalently expressed in its matrix form as $H \in \mathbb{R}^{|V| \times d}$. The graph topology is described by a symmetric adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$, where $A_{u,v} = 1$ if an edge exists between vertices u and v.

To build CG, we detect nuclei at 40× resolution using Hover-Net [Graham et al., 2019a], a state-of-the-art nuclei segmentation algorithm pre-trained on MoNuSeg dataset [Kumar et al., 2020]. We extract 16 hand-crafted features incorporating shape, texture and color attributes to represent each nucleus as in [Zhou et al., 2019a]. We include centroid location normalised by the image size to spatially encode the nucleus. The detected nuclei and their 18-dimensional embeddings serve as the node and initial node embeddings of our CG. The CG topology assumes that spatially close cells encode biological interactions and consequently should form an edge. We use the k-Nearest Neighbors (kNN) algorithm, i. e., for each node u, we build edges e_{uv} to the k closest vertices v. As isolated cells have weak cellular interaction with other cells, they ought to stay detached. Thus, we threshold the kNN graph by removing edges that are longer than a specified distance. We set $k = 5$ and the distance threshold to 50 pixels in our modeling.

For the downstream DP task, we determine the breast cancer subtypes of regions-of-interest (TRoIs). For a dataset with N TRoIs, we create $\mathcal{D} = \{G_{\text{CG},i}, l_i\}_{i=\{1,\ldots,N\}}$ consisting of N CGs and corresponding labels l_i. A GNN [Defferrard et al., 2016; Kipf et al., 2017; Veličković et al., 2018; Xu et al., 2019b], denoted as CG-GNN, is employed to build fixed-size graph embeddings from the CGs. These embeddings are fed to a MLP to predict the cancer subtype. In particular, we use the Graph Isomorphism Network (GIN) [Xu et al., 2019b], an instance of message passing neural network [Gilmer et al., 2017]. A block diagram with the main steps is presented in Figure 5.1.

5.2.2 Cell-graph explainer

We propose a cell-graph explainer (CGExplainer) inspired by the GnnExplainer [Ying et al., 2019], a post-hoc interpretability method based on a graph pruning optimization. Considering the large number of cells in a TRoI, we hypothetize that many of them will provide little information in the decision making, whereas others will be responsible for class specific patterns that would allow better understanding of the disease. Thus, we prune the redundant and uninformative graph components, and define the resulting sub-graph as the *explanation*.

Formally, let us consider a trained GNN model $\mathcal{M}$, and a sample $\{G_{\text{CG}}, l\}$ from $\mathcal{D}$ predicted as $\hat{y} = \mathcal{M}(G_{\text{CG}})$. We aim to find a sub-graph $G_s = (V_s, E_s, H_s) \subset G_{\text{CG}}$ such that the mutual information between the original prediction and the sub-graph is maximized, i. e.,

$$\max_{G_s} \text{MI}(\hat{Y}, G_s) = \mathcal{H}(\hat{Y}) - \mathcal{H}(\hat{Y}|G_{\text{CG}} = G_s) \tag{5.1}$$

which is equivalent to minimizing the conditional entropy:

$$\mathcal{H}(\hat{Y}|G_{\text{CG}} = G_s) = -\,\mathbb{E}_{\hat{Y}|G_s}[\log(P_{\mathcal{M}}(\hat{Y}|G_s))] \tag{5.2}$$

Intuitively, G_s maximizes the probability of $\hat{y}$. Direct optimization of Equation (5.2) is intractable due the combinatorial nature of graphs. Therefore, CGExplainer proposes to learn a mask that activates or deactivates parts of the graph. Considering the coherent pathological explainability of cells compared to cellular interactions, we focus on

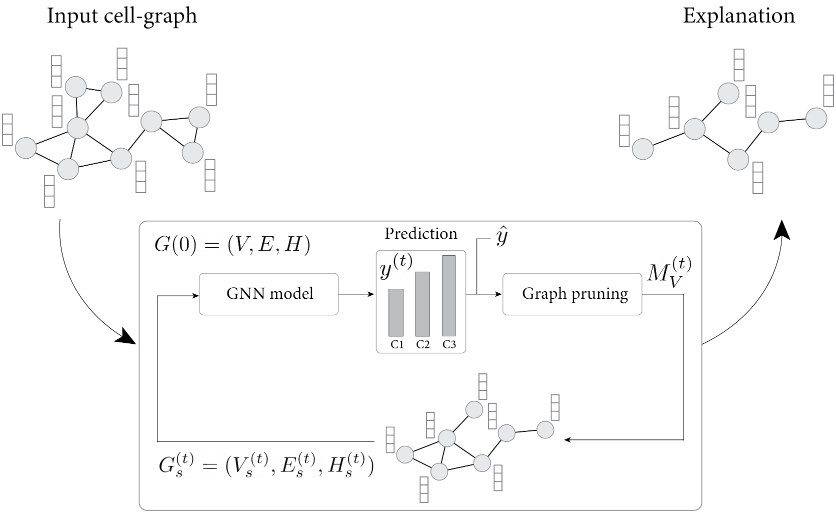


Figure 5.2: Overview of CGExplainer. The original input CG is iteratively pruned until convergence of the optimization.

interpreting the *cells* in this work. Thus, we aim at learning a mask M_V at *node-level* that satisfies:

$$\min_{M_V} - \sum_{c=1}^{C} \mathbb{1}_{[y=c]} \log(P_{\mathcal{M}}(\hat{Y}|G_{\text{CG}}, \sigma(\text{diag}(M_V))H))) \tag{5.3}$$

where C denotes the number of classes, σ is the sigmoid activation, and $\text{diag} : \mathbb{R}^{|V|} \to \mathbb{R}^{|V| \times |V|}$ is the diagonal matrix of the weight vector M_V. We intend the explanations to be as compact as possible, ideally with binarized weights, while providing the same prediction as the original graph. Heuristically, we enforce these constraints by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{KD}}(\hat{y}, y^{(t)}) + \alpha_{M_V} \sum_{i}^{|V|} \sigma(M_{V_i}^{(t)}) + \alpha_{\mathcal{H}} \mathcal{H}^e(\sigma(M_V^{(t)})) \tag{5.4}$$

where, t is the optimization step. First term is the knowledge-distillation loss $\mathcal{L}_{\text{KD}}$ between the new logits $y^{(t)}$ and the original prediction $\hat{y}$. Second term ensures the compactness of M_V. Third term binarizes M_V by minimizing its element-wise entropy $\mathcal{H}^e$. Following [Hinton et al., 2015], $\mathcal{L}_{\text{KD}}$ is a combination of distillation and cross-entropy loss:

$$\mathcal{L}_{\text{KD}} = \lambda \mathcal{L}_{\text{CE}} + (1 - \lambda)\mathcal{L}_{\text{DIST}} \text{ where } \lambda = \frac{\mathcal{H}^e(y^{(t)})}{\mathcal{H}^e(\hat{y})} \tag{5.5}$$

As the element-wise entropy $\mathcal{H}^e(y^{(t)})$ increases, $\mathcal{L}_{\text{CE}}$ gains importance and avoids a change in predicted label. M_V, produced by optimizing Equation (5.4), identifies important nodes with a weight factor. An overview of the explainer module is shown in Figure 5.2.

5.3 Dataset

We evaluate CGExplainer on BRACS dataset, an in-house collection of BReAst Carcinoma Subtyping[1] images. The dataset consists of 2080 TRoIs acquired from 106 H&E stained breast carcinoma whole-slide-images (WSIs). The TRoIs are extracted at 40× magnification producing images of various sizes and appearances. The TRoIs are annotated by the consensus of three pathologists as: normal (N), benign[2] (B), atypical[3] (A), ductal carcinoma in situ (D), and invasive (I) (a 5-class problem). We also study two simplified scenarios: (1) a 2-class problem: benign (N+B) and malignant (D+I) categories, and (2) a 3-class problem: benign (N+B), atypical (A), and malignant (D+I) categories. These scenarios allow us to study the relation between the task complexity and the generated explanations. Non-overlapping train, validation and test splits are created at WSI-level consisting of 1356, 365, and 359 TRoIs respectively.

5.4 Results

5.4.1 Implementation

The experiments were conducted using PyTorch [Paszke et al., 2019] and the DGL library [Wang et al., 2019a]. The CG-GNN consisted of three GIN layers with a hidden dimension of 32. Each GIN layer used a 2-layer MLP with ReLU activation. The classifier consisted of a 2-layer MLP with 64 hidden neurons that mapped the hidden dimensions to the number of classes. The model was trained using the Adam optimizer with an initial learning rate of 10^{-3} and a weight decay of 5×10^{-4}. The batch size was set to 16.

The explanation module used the trained CG-GNN. The mask M_V was learned by using the Adam optimizer with a learning rate of 0.01. The size constraint and the entropy constraint contributed to the loss by weighting factors $\alpha_{M_V} = 0.005$ and $\alpha_{\mathcal{H}} = 0.1$, respectively. The weights were adjusted such that the individual losses have comparable range. An early stopping mechanism was triggered, if G_s predicted a different label before reaching convergence. This ensured that the graph and its explanation always had the same prediction.

5.4.2 Quantitative and qualitative analyses

We conducted absolute and comparative analyses between CGExplainer and a random-explainer (RGExplainer). RGExplainer generates a random explanation from an original CG for a TRoI by retaining equal number of nodes and edges as retained by CGExplainer. We quantitatively and qualitatively evaluated the explainers under 2-, 3-, and 5-class scenarios, and assessed them using surrogate metrics in the absence of any ground truth explanations. Table 5.1 presents the weighted F1-scores for CG-GNNs, the average node and edge reduction in the CGExplainer explanations, and cross-entropy (CE) loss of CG-GNN for processing the original CG, CGExplainer-based CG, and RGExplainer-based CG. The cross-entropy was computed between the predicted class probabilities and the ground-truth labels of the TRoIs.

---

[1] BRACS dataset for breast cancer subtyping: https://www.bracs.icar.cnr.it

[2] includes benign and usual ductal hyperplasia

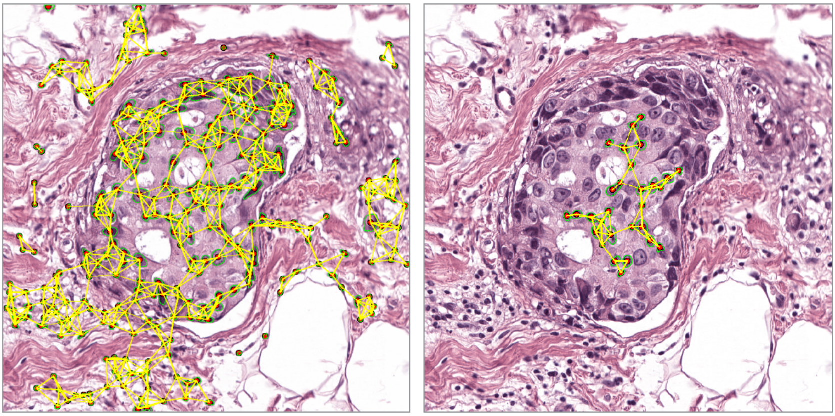
[3] includes flat epithelial atypia and atypical ductal hyperplasia

Table 5.1: Quantitative results for CG-GNN, CGExplainer compactness, CGExplainer and RGExplainer performances. ↑ and ↓ indicate higher and lower values are better, respectively.

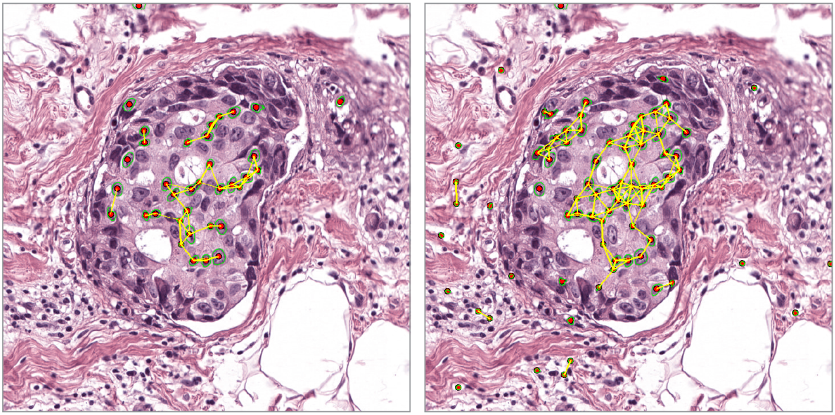
Metric/Scenario	**2-class scenario**			**3-class scenario**				**5-class scenario**					
	N+B	D+I	All	N+B	A	D+I	All	N	B	A	D	I	All
Weighted F1-score (↑)	0.97	0.97	0.97	0.95	0.35	0.80	0.77	0.56	0.74	0.37	0.62	0.77	0.61
Node reduction (%) (↑)	97.7	91.6	94.6	89.5	92.4	85.6	88.5	92.3	93.8	75.8	63.3	59.00	76.9
Edge reduction (%) (↑)	99.2	93.8	96.4	94.3	98.7	90.5	93.5	97.1	97.0	90.6	74.1	62.8	84.2
Original CE (↓)	0.21	0.21	0.21	0.45	2.05	0.38	0.72	2.65	0.59	2.22	0.72	0.48	1.21
Explanation CE (↓)	0.10	0.21	0.16	0.44	1.41	0.55	0.67	1.65	0.73	1.61	2.57	0.67	1.41
Random CE (↓)	0.02	3.14	1.61	1.00	0.38	1.75	1.20	0.62	0.93	1.52	11.4	2.85	3.55

The CGExplainer removes a large percentage of nodes and edges to generate compact explanations for 2-, 3-, and 5-class scenarios, while preserving the TRoI predictions. The decrease in the percentage of node-reduction with the increase in the number of classes per-task indicates that with the increment of task complexity, the explainer exploits more nodes to extract valuable information. A similar pattern is observed for the percentage of edge-reduction. Further, the reduction percentage within a task decreases with the increase in the malignancy of the TRoI. It indicates that the explainer discards the abundantly available but less relevant benign epithelial, stromal, and lymphocytes, while retains the relevant tumor and atypical nuclei. Combining the CG explanations in Figure 5.3 and the nuclei types annotation in Figure 5.4, we infer that the explanations retain relevant tumor epithelial nuclei for DCIS diagnosis. For the 2-class scenario, the CG includes tumor nuclei in the central region of the gland. In this case, a few tumor epithelial nuclei are sufficient to differentiate (D) from (N+B). For the 3-class scenario, the CG includes more tumor epithelial nuclei in the central and the periphery regions of the gland, and does not consider the atypical nuclei. This pattern differentiates (D) from (A). For the 5-class scenario, the CG includes more tumor nuclei distributed within and around the gland, and some lymphocytes around the gland. The CG also includes more cellular interactions to identify a large cluster of tumor nuclei. Pathologically this behavior differentiates (D) from (I) which has small clusters of tumor nuclei scattered throughout the TRoI. Additionally, the retained tumor nuclei and their interactions are consistent across the considered classification scenarios.

Further, we compared the class-wise predicted probabilities with the ground truth labels for the original, CGExplainer-based CG, and RGExplainer-based CG via a cross-entropy (CE) function. Table 5.1 presents the class-wise CE and average CE across all the classes. The CGExplainer-based CG and the original CG have comparable class-wise CE and average CE across all scenarios. We observe that in each scenario, the RGExplainer-based CG is biased towards one class. For instance, in the 2-class scenario, RGExplainer frequently predicts the class (N+B) leading to a per-class CE smaller than CGExplainer. Further, on average across all the classes, the RGExplainer CE is consistently higher than the CGExplainer. It conveys that the RGExplainer removes relevant entities from CGs,



(a) Original CG (b) 2-class explanation CG



(c) 3-class explanation CG (d) 5-class explanation CG

Figure 5.3: Qualitative comparison of original CG and CGExplainer generated explanations for 2-, 3-, and 5-class scenarios on a DCIS TRoI.

thereby increasing the loss. These qualitative and quantitative analyses conclude that the CGExplainer generates meaningful and consistent explanations.

5.5 Conclusion

We believe that our work, though preliminary, is a step in the right direction towards better representations and interpretability in DP. We have herein focused on the methodological introduction and cell-level analyses. In future work, we plan to extend our approach to other biological entities and further to pathological assessment. Ultimately, our goal is

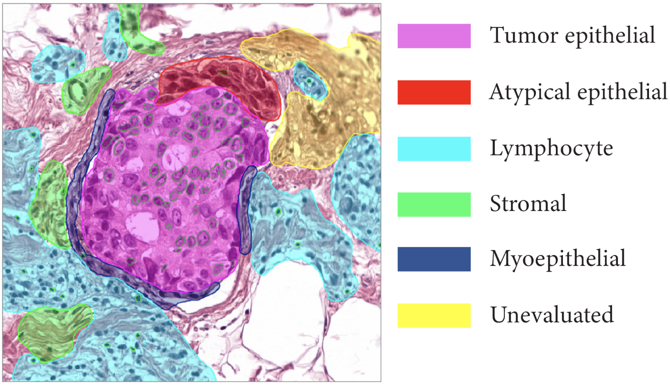


Figure 5.4: Nuclei types annotation. Overlaid segmentation masks of nuclei from 5-class explanation in green.

to understand any information additional to a DL model prediction that one needs to provide to a user, to build trust and to facilitate adoption and deployment of such DL technologies in clinical setting.

6

Quantifying Explainers of Graph Neural Networks in Digital Pathology

Explainability of deep learning methods is imperative to facilitate their clinical adoption in digital pathology. However, popular deep learning methods and explainability techniques (explainers) based on pixel-wise processing disregard biological entities' notion, thus complicating comprehension by pathologists. In this work, we address this by adopting biological entity-based graph processing and graph explainers enabling explanations accessible to pathologists. In this context, a major challenge becomes to discern meaningful explainers, particularly in a standardized and quantifiable fashion. To this end, we propose herein a set of novel quantitative metrics based on statistics of class separability using pathologically measurable concepts to characterize graph explainers. We employ the proposed metrics to evaluate three types of graph explainers, namely the layer-wise relevance propagation, gradient-based saliency, and graph pruning approaches, to explain Cell-Graph representations for Breast Cancer Subtyping. The proposed metrics are also applicable in other domains by using domain-specific intuitive concepts. We validate the qualitative and quantitative findings on the BRACS dataset, a large cohort of breast cancer RoIs, by expert pathologists. The code, data, and models can be accessed at `https://github.com/histocartography/patho-quant-explainer`.

6.1 Introduction

Histopathological image understanding has been revolutionized by recent machine learning advancements, especially deep learning (DL) [Bera et al., 2019; Serag et al., 2019]. DL has catered to increasing diagnostic throughput as well as a need for high predictive performance, reproducibility and objectivity. However, such advantages come at the cost of a reduced transparency in decision-making processes [Holzinger et al., 2017; Tizhoosh et al., 2018; Hägele et al., 2020]. Considering the need for reasoning any clinical decision, it is imperative to enable the explainability of DL decisions to pathologists.

This chapter has been published as: Guillaume Jaume†, Pushpak Pati†, Behzad Bozorgtabar, Antonio Foncubierta-Rodriguez, Anna Maria Anniciello, Florinda Feroce, Tilman Rau, Jean-Philippe Thiran, Maria Gabrani, Orcun Goksel "Quantifying Explainers of Graph Neural Networks in Computational Pathology", In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8106-8116, Nashville, USA, Jun 2021, doi: 10.1109/CVPR46437.2021.00801, [arXiv]

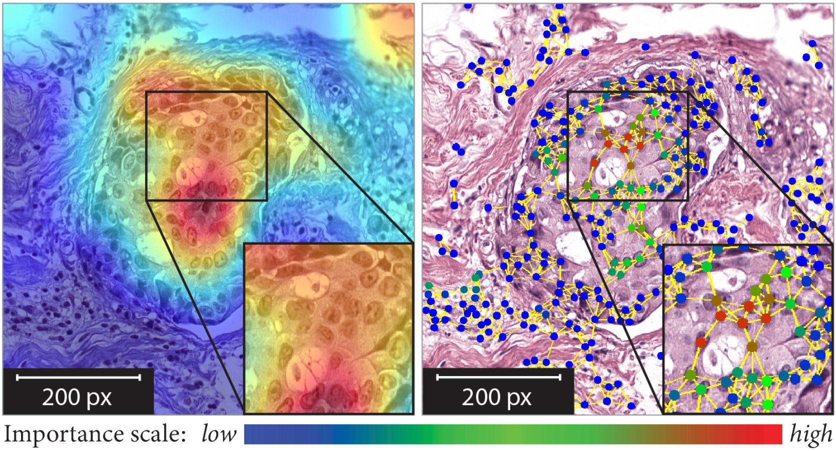


Figure 6.1: Sample explanations produced by pixel- and entity-based explainability techniques for a ductal carcinoma *in situ* RoI.

Inspired by the explainability techniques (explainers) for DL model decisions on natural images [Simonyan et al., 2013; Zeiler et al., 2014; Yosinski et al., 2015; Bach et al., 2015; Montavon et al., 2015; Selvaraju et al., 2017; Kindermans et al., 2015; Zintgraf et al., 2017; Chattopadhay et al., 2018; Kim et al., 2018], several explainers have been implemented in digital pathology, such as feature attribution [Korbar et al., 2017; Binder et al., 2018; Hägele et al., 2020], concept attribution [Graziani et al., 2020], and attention-based learning [Lu et al., 2021]. However, pixel-level explanations, exemplified in Figure 6.1, pose several notable issues, including: (1) a pixel-wise analysis disregards the notion of biological tissue entities, their topological distribution, and inter-entity interactions; (2) a typical patch-based DL processing and explainer fail to accommodate complete tumor macro-environment information; and (3) pixel-wise visual explanations tend to be blurry. Explainability in entity space is thus a natural choice to address the above issues. To that end, an entity-graph representation is built for a histology image, where nodes and edges denote biological entities and inter-entity interactions followed by a Graph Neural Network (GNN) [Kipf et al., 2017; Xu et al., 2019b]. The choice of entities, such as cells [Demir et al., 2004; Zhou et al., 2019a; Pati et al., 2022], tissues [Pati et al., 2022] or others, can be task-dependent. Subsequently, explainers for graph-structured data [Baldassarre et al., 2019; Pope et al., 2019; Ying et al., 2019] applied to the entity-graphs highlight responsible entities for the concluded diagnosis, thereby generating intuitive explanations for pathologists.

In the presence of various graph explainers producing distinct explanations for an input, it is crucial to discern the explainer that best fits the explainability definition [Arrieta et al., 2020]. In the context of computational pathology, explainability is defined as making the DL decisions understandable to pathologists [Holzinger et al., 2017]. To this end, the qualitative evaluation of explainers' explanations by pathologists is the candid measure. However, it requires evaluations by task-specific expert pathologists, which is subjective, time-consuming, cumbersome, and expensive. Additionally, though the explanations are intuitive, they do not relate to pathologist-understandable terminologies, e. g., “How big are the important nuclei?", and “How irregular are their shape?", which toughens the comprehensive analysis. These bottlenecks undermine not only any qualitative assessment but also quantitative metrics requiring user interactions [Mohseni et al., 2021]. Furthermore, expressing the quantitative metrics in user-understandable terminologies [Arrieta et al., 2020] is fundamental to achieve interpretability [Doshi-Velez et al., 2017; Nguyen et al., 2020]. To this end, the most popular quantitative metric, explainer *fidelity* [Ribeiro et al., 2016; Dhurandhar et al., 2017; Samek et al., 2017; Hoffman et al., 2018; Mohseni et al., 2021; Pope et al., 2019], is not satisfactory. Moreover, explainers intrinsically maintain high-*fidelity*, e. g., GNNEXPLAINER [Ying et al., 2019] produces an explanation to match the GNN’s prediction on the original graph.

Thus, we propose a set of novel user-independent quantitative metrics expressing pathologically-understandable *concepts*. The proposed metrics are based on class separability statistics using such *concepts*, and they are applicable in other domains by incorporating domain-specific *concepts*. We use the proposed metrics to evaluate three types of graph-explainers, (1) graph pruning: GNNEXPLAINER [Ying et al., 2019; Jaume et al., 2020], (2) gradient-based saliency: GRAPHGRAD-CAM [Selvaraju et al., 2017; Pope et al., 2019], GRAPHGRAD-CAM++ [Chattopadhay et al., 2018], (3) layer-wise relevance propagation: GRAPHLRP [Bach et al., 2015; Montavon et al., 2015; Schwarzenberg et al., 2019], for explaining Cell-Graphs [Demir et al., 2004] in Breast Cancer Subtyping as shown in Figure 6.1. Our specific contributions in this work are:

- A set of novel quantitative metrics based on the statistics of class separability using domain-specific *concepts* to characterize graph explainability techniques. To the best of our knowledge, our metrics are the first of their kind to quantify explainability based on domain-understandable terminologies;
- Explainability in computational pathology using pathologically intuitive entity-graphs;
- Extensive qualitative and quantitative assessment of various graph explainability techniques in computational pathology, with a validation of the findings by expert pathologists.

6.2 Related work

Graphs in digital pathology: Graph-based tissue image analysis effectively describes a tissue environment by incorporating morphology, topology, and tissue components interactions. To this end, cell-graph (CG) is the most popular graph representation, where nodes and edges depict cells and cellular interactions [Demir et al., 2004]. Cell morphology is embedded in the nodes via hand-crafted features [Demir et al., 2004; Zhou et al., 2019a; Pati et al., 2020] or DL features [Chen et al., 2020a; Pati et al., 2022]. The graph topology is heuristically defined using k-Nearest Neighbors, probabilistic modeling, Waxman model etc. [Sharma et al., 2015]. Subsequently, the CGs are processed by classical machine learning [Sharma et al., 2015; Sharma et al., 2016; Sharma et al., 2017a] or GNN [Zhou et al., 2019a; Chen et al., 2020a; Anand et al., 2019; Pati et al., 2020] to map the tissue structure to function relationship. Recently, improved graph-representations using patches [Aygüneş et al., 2020], tissue components [Pati et al., 2022], and hierarchical cell-to-tissue relations [Pati et al., 2022] are proposed to enhance the structure-function

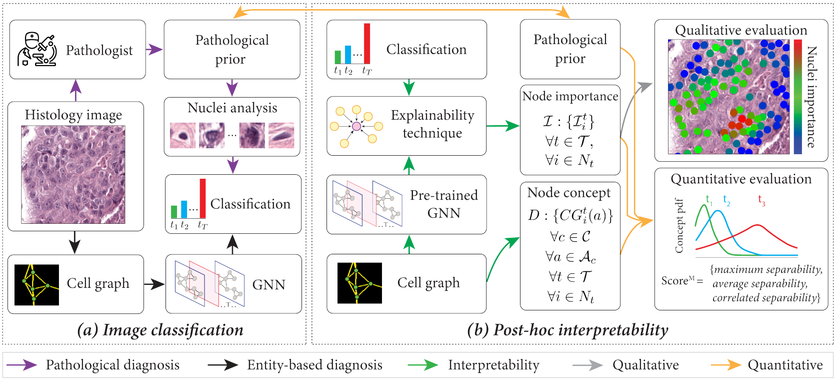


Figure 6.2: Overview of the proposed framework. (a) presents pathologist, and entity-based (cell-graph + GNN) diagnosis of a histology image. (b) presents nuclei-level pathologically relevant *concept* measure D, a post-hoc graph explainability technique to derive nuclei-level importance $\mathcal{I}$ for *concepts* $\mathcal{C}$, measurable *attributes* $\mathcal{A}_c$, and classes $\mathcal{T}$. D, $\mathcal{I}$ and prior pathological knowledge defining *concepts'* relevance are utilized to propose a novel set of quantitative metrics to evaluate the explainer quality in pathologist-understandable terms.

mapping. Other graph-based applications in computational pathology include cellular community detection [Javed et al., 2020], whole-slide image classification [Zhao et al., 2020b; Adnan et al., 2020] etc. Intuitively, a graph representation utilizes pathologically relevant entities to represent a tissue specimen, which allows pathologists to readily relate with the input, also enabling them to include any task-specific prior knowledge.

Explainability in digital pathology: Explainability is an integral part of pathological diagnosis. Though DL solutions have achieved remarkable diagnostic performance, their lack of explainability is unacceptable in the medical community [Tizhoosh et al., 2018]. Recent studies have proposed visual explanations [Hägele et al., 2020] and salient regions [Korbar et al., 2017; Hägele et al., 2020] using feature-attribution techniques [Selvaraju et al., 2017; Chattopadhay et al., 2018]. Differently, concept-attribution technique [Graziani et al., 2020] evaluates the sensitivity of network output w.r.t. quantifiable image-level pathological *concepts* in patches. Although such explanations are pathologist-friendly, image-level *concepts* are neither fit nor meaningful for real-world large histology images that contain many localized concepts. Furthermore, attention-based learning [Lu et al., 2021], and multimodal mapping between image and diagnostic report [Zhang et al., 2019] are devised to localize network attention. However, the pixel-wise and patch-based processing in all the aforementioned techniques ignore biological entities' notion; thus, they are not easily understood by pathologists. Separately, the earlier stated entity-graph-based processing provides an intuitive platform for pathologists. However, research on explainability and visualization using entity-graphs has been scarce: CGC-Net [Zhou et al., 2019a] analyzes cluster assignment of nodes in CG to group them according to their appearance and tissue types. CGExplainer [Jaume et al., 2020] introduces a post-hoc graph-pruning explainer to identify decisive cells and interactions. Robust spatial filtering [Sureka et al., 2020] utilizes an attention-based GNN and node occlusion to highlight cell contributions. No previous work has comprehensively analyzed and quantified graph explainers in computational pathology while expressing explanations in a pathologist-understandable form to the best of our knowledge. This gap between the existing and desired explainability of DL outputs in digital pathology motivates our work herein.

6.3 Methodology

In this section, we present entity-graph processing, explainability methods, and our proposed evaluation metrics. First, we transform a histology region-of-interest (RoI) into a *biological entity-graph*. Second, we introduce a "black-box" GNN that maps the *entity-graph* to a corresponding class label. Third, we employ a post-hoc graph explainer to generate explanations. Finally, we perform a qualitative and quantitative assessments of the generated explanations. An overview of the methodology is shown in Figure 6.2.

6.3.1 Entity-graph notations

We define an attributed undirected entity-graph $G := (V, E, H)$ as a set of nodes V, edges E, and node attributes $H \in \mathbb{R}^{|V| \times d}$. d denotes the number of attributes per node, and $|.|$ denotes set cardinality. The graph topology is defined by a symmetric graph adjacency, $A \in \mathbb{R}^{|V| \times |V|}$, where $A_{u,v} = 1$ if $e_{uv} \in E$. We denote the neighborhood of a node $v \in V$ as $\mathcal{N}(v) := \{u \in V \mid v \in V,\ e_{uv} \in E\ \}$. We denote a set of graphs as $\mathcal{G}$.

6.3.2 Entity-graph construction

Our methodology begins with transforming RoIs into entity-graphs. It ensures the method's inputs are pathologically interpretable, as the inputs consist of biologically-defined objects that pathologists can directly *relate-to* and *reason-with*. Thus, image-to-graph conversion moves from *uninterpretable* to *interpretable* input space. In this work, we consider cells as entities, thereby transforming RoIs into cell-graphs (CGs). A CG nodes and edges capture the morphology of cells and cellular interactions. A CG topology acquires both tissue micro and macro-environment, which is crucial for characterizing cancer subtypes.

First, we detect nuclei in a RoI at 40× magnification using Hover-Net [Graham et al., 2019a], a nuclei segmentation algorithm pre-trained on MoNuSeg [Kumar et al., 2017]. We process patches of size 72×72 around the nuclei by ResNet34 [He et al., 2016] pre-trained on ImageNet [Deng et al., 2009] to produce nuclei visual attributes. We further concatenate nuclei spatial attributes, *i.e.*, nuclei centroids min-max normalized by RoI dimension. The nuclei and their attributes (visual and spatial) define the nodes and node attributes of the CG, respectively. Following prior work [Pati et al., 2022], we construct the CG topology by employing thresholded k-Nearest Neighbors algorithm. We set $k = 5$, and prune the edges longer than 50 pixels (12.5 μm). The CG-topology encodes how likely two nearby nuclei will interact [Francis et al., 1997]. A CG example is presented in Figure 6.1.

6.3.3 Entity-graph learning

Given $\mathcal{G}$, the set of CGs, the aim is to infer the corresponding cancer subtypes. We use GNNs [Scarselli et al., 2009; Defferrard et al., 2016; Kipf et al., 2017; Hamilton et al., 2017; Veličković et al., 2018; Ying et al., 2018; Gilmer et al., 2017], the conceptual analogous

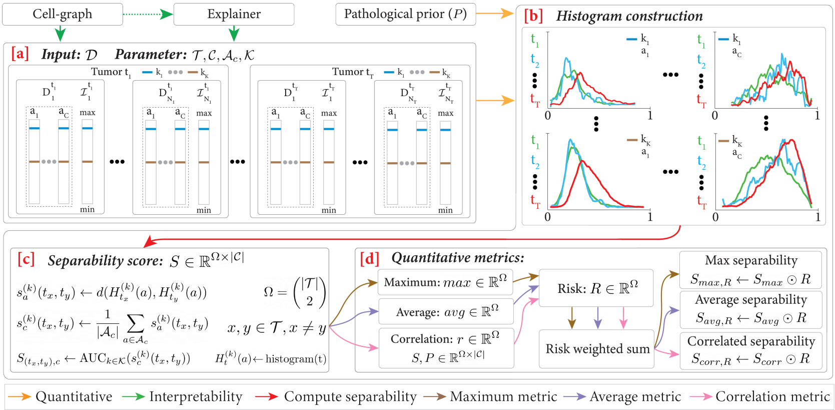


Figure 6.3: Overview of proposed quantitative assessment. (a) presents input dataset $\mathcal{D}$, and parameters *concepts* $\mathcal{C}$, measurable *attributes* $\mathcal{A}_c$, classes $\mathcal{T}$, and importance thresholds $\mathcal{K}$. For simplicity $|\mathcal{A}_c| = 1, \forall c \in \mathcal{C}$ in this figure. (b) shows histogram probability densities for $\forall a \in \mathcal{A}_c, \forall k \in \mathcal{K}, \forall t \in \mathcal{T}$. (c) displays the algorithm for computing class separability score S. (d) presents the algorithm for computing the proposed class separability-based risk-weighted quantitative metrics.

of 2D convolution for graph-structured data, to classify the CGs. A GNN layer follows two steps: for each node $v \in V$, (1) *aggregation step*: the states of neighboring nodes, $\mathcal{N}(v)$, are aggregated via a differentiable and permutation-invariant operator to produce $a(v) \in \mathbb{R}^d$, then, (2) *update step*: the state of v is updated by combining the current node state $h(v) \in \mathbb{R}^d$ and the aggregated message $a(v)$ via another differentiable operator. After L iterations, i. e., the number of GNN layers, a *readout step* is employed to merge all the node states via a differentiable and permutation-invariant function to result in a fixed-size graph embedding. Finally, the graph embeddings are processed by a classifier to predict the class label.

In this work, we use a flavor of Graph Isomorphism Network (GIN) [Xu et al., 2019b], that uses *mean* and a *multi-layer perceptron* (MLP) in the *aggregation* and *update* step respectively. Formally, we define a layer as,

$$h(v)^{(l+1)} = \text{MLP}^{(l)}\left(h(v)^{(l)} + \frac{1}{|\mathcal{N}(v)|}\sum_{u \in \mathcal{N}(v)} h(u)^{(l)}\right) \tag{6.1}$$

where $h(v)$ denotes features of node v, and $l \in \{1, ..., L\}$. Our GNN consists of 3-GIN layers, with each layer including a 2-layer MLP. The dimension of latent node embeddings is fixed to 64 for all layers. We use *mean* operation in *readout step*, and feed the graph embedding to a 2-layer MLP classifier. The GNN is trained end-to-end by minimizing cross-entropy loss between predicted logits and target cancer subtypes. We emphasize that the entity-based processing follows a pathologist's diagnostic procedure that identifies diagnostically relevant nuclei and analyzes cellular morphology and topology in a RoI, as shown in Figure 6.2.

6.3.4 Post-hoc graph explainer

We generate an explanation per entity-graph by employing post-hoc graph explainers. The explanations allow to evaluate the pathological relevance of black-box neural network reasoning. Specifically, we aim to evaluate the agreement between the pathologically relevant set of nuclei in a RoI, and the explainer identified set of important nuclei, i. e., nuclei driving the prediction, in a CG. In this work, we consider three types of graph explainers for explaining CGs, which follow similar operational setting, i. e., (1) input data are attributed graphs, (2) a GNN is trained *a priori* to classify the input data, and (3) each data point can be inferred independently to produce an explanation. Here, we briefly present the graph explainers. The detailed formulations are described in Section 2.2.4.

GraphLRP: Layerwise relevance propagation (LRP) [Bach et al., 2015] propagates the output logits backward in the network using a set of propagation rules to quantify the positive contribution of input pixels for a certain prediction. Specifically, LRP assigns an importance score to each neuron such that the output logit relevance is preserved across layers. While initially developed for explaining fully-connected layers, LRP can be extended to GNN by treating the GNN *aggregation step* as a fully connected layer that projects the graph adjacency matrix on the node attributes as in [Schwarzenberg et al., 2019]. LRP outputs per-node importance.

GraphGrad-CAM: GRAD-CAM [Selvaraju et al., 2017] is a feature attribution approach designed for explaining CNNs operating on images. It produces class activation explanation following two steps. First, it assigns weights to each channel of a convolutional layer l by computing the gradient of the targeted output logit w.r.to each channel in layer l. Second, importance of the input elements are computed by the weighted combination of the forward activations at each channel in layer l. The extension to GNN is straightforward [Pope et al., 2019], and only requires to compute the gradient of the predicted logits w.r.to a GNN layer. Following prior work [Pope et al., 2019], we take the average of node-level importance-maps obtained from all the GNN layers $l \in \{1, ..., L\}$ to produce smooth per-node importance.

GraphGrad-CAM++: GRAD-CAM++ [Chattopadhay et al., 2018] is an increment on GRAD-CAM by including spatial contributions into the channel-wise weight computation of a convolution layer. The extension allows weighting the contribution by each spatial location at a layer for improved spatial localization. The spatial locations in a convolutional layer are analogous to the size of the graph in a GNN layer. With this additional consideration, we propose an extension of GRAD-CAM++ to graph-structured data.

GnnExplainer: GNNEXPLAINER [Ying et al., 2019; Jaume et al., 2020] is a graph pruning approach that aims to find a compact sub-graph $G_s \subset G$ such that mutual information between G_s and GNN prediction of G is maximized. Sub-graph G_s is regarded as the explanation for the input graph G. GNNEXPLAINER can be seen as a feature attribution technique with binarized node importances. To address the combinatorial nature of finding G_s, GNNEXPLAINER formulates it as an optimization problem that learns a mask to activate or deactivate parts of the graph. [Jaume et al., 2020] reformulates the initial approach in [Ying et al., 2019] to learn a mask over the nodes instead of edges. The approach in [Jaume et al., 2020] is better suited for pathology as the nodes, i. e., biological entities, are more intuitive and substantial for disease diagnosis than heuristically-defined edges. The optimization for an entity-graph results in per-node importance.

6.3.5 Quantitative metrics for graph explainability

In the presence of several graph explainers producing distinct explanations for an input, it is imperative to discern the explainer that produces the most pathologically-aligned explanation. Considering the limitations of existing qualitative and quantitative measures presented in Section 6.1, we propose a novel set of quantitative metrics based on class separability statistics using pathologically relevant *concepts*. Intuitively, a good explainer should emphasize the relevant *concepts* that maximize the class separation. Details of the metric evaluations are presented as follows.

Input: A graph explainer outputs an explanation, i. e., node-level *importance* $\mathcal{I}$, for an input CG. To quantify a *concept* $c \in \mathcal{C}$, $\mathcal{C}$ denoting the set of *concepts*, we measure nuclear *attributes* $a \in \mathcal{A}_c$ for each nucleus in CG, e. g., for $c =$ *nuclear shape*, we measure $\mathcal{A}_c =$ {*perimeter, roughness, eccentricity, circularity*}. We create a dataset $\mathcal{D} = \bigcup_{t \in \mathcal{T}} \mathcal{D}_t$, $\mathcal{T}$ denoting the set of cancer subtypes. We define $\mathcal{D}_t := \{(D_i^t, \mathcal{I}_i^t) | i = 1, \ldots, N_t\}$ $\forall t \in \mathcal{T}$, where N_t is the number of CGs for tumor type t. $\mathcal{I}_i^t$ and D_i^t are, respectively, the sorted importance matrix for a CG indexed by i and corresponding node-level attribute matrix. To perform inter-concept comparisons, we conduct *attribute*-wise normalization across all D_i^t $\forall t, i$. In order to compare different explainers, we conduct CG-wise normalization of $\mathcal{I}$. The structure of input dataset $\mathcal{D}$ is presented in Figure 6.3(a).

Note that the notion of important nuclei vary (1) per-CG since the number of nodes vary across CGs, and (2) per-explainer. Hence, selecting a *fixed* number of important nuclei per-CG and per-explainer is not meaningful. To overcome this issue, we assess different number of important nuclei $k \in \mathcal{K}$, selected based on node importances, per-CG and per-explainer. In the following sections we will show how to aggregate the results for a given explainer.

Histogram construction: Given the input dataset $\mathcal{D}$, and parameters $\mathcal{K}, \mathcal{C}, \mathcal{A}_c, \mathcal{T}$, we apply threshold $k \in \mathcal{K}$ on $\mathcal{I}_i^t, \forall t \in \mathcal{T}, \forall i \in N_t$ to select CG-wise most important nuclei. The cancer subtype-wise selected set of nuclei data from $\mathcal{D}$ are used to construct histograms $H_t^{(k)}(a), \forall a \in \mathcal{A}_c$, $\forall c \in \mathcal{C}$ and $\forall t \in \mathcal{T}$. For histogram $H_t^{(k)}(a)$, bin-edges are decided by quantizing the complete range of *attribute* a, i. e., $\mathcal{D}(a)$, by a fixed step size. We convert each $H_t^{(k)}(a)$ into a probability density function. Similarly, sets of histograms are constructed by applying different thresholds $k \in \mathcal{K}$. Sample histograms are shown in Figure 6.3(b).

Separability Score (S): Given two classes $t_x, t_y \in \mathcal{T}$ and corresponding probability density functions $H_{t_x}^{(k)}(a)$ and $H_{t_y}^{(k)}(a)$, we compute *class separability* $s_a^{(k)}(t_x, t_y)$ based on optimal transport as the Wasserstein distance between the two density functions. We average $s_a^{(k)}(t_x, t_y)$ over all $a \in \mathcal{A}_c$ to obtain a score $s_c^{(k)}(t_x, t_y)$ for *concept* c and threshold k. Finally, we compute the area-under-the-curve (AUC) over the threshold range $\mathcal{K}$ to get the aggregated class separability $S_{(t_x,t_y),c}$ for a *concept* c. The class separability score indicates the significance of *concept* c for the purpose of separating t_x and t_y. Thus, separability scores can be used to compare different *concepts* and to identify relevant ones for differentiating t_x and t_y. A pseudo-algorithm is presented in Algorithm 1, and illustrated in Figure 6.3(c). A separability matrix $S \in \mathbb{R}^{\Omega \times |\mathcal{C}|}$ is built by computing class separability scores for all pair-wise classes, i. e., $\forall \ (t_x, t_y) \in \Omega := \binom{|\mathcal{T}|}{2}$ and $\forall c \in \mathcal{C}$.

Statistics of Separability Score: Since explainability is not uniquely defined, we include multiple metrics highlighting different facets. We compute three separability statistics $\forall(t_x, t_y) \in \Omega$ using S as given in Equation (6.2), i. e., (1) *maximum*: the utmost separability, (2) *average*: the expected separability. These two metrics encode (model+explainer)'s focus, i. e., "how much the black-box model implicitly uses the *concepts* for class separability?", (3) *correlation*: encodes the agreement between (model+explainer)'s focus and pathological prior P. $P \in \mathbb{R}^{\Omega\times|\mathcal{C}|}$ signifies the relevance $\forall c \in \mathcal{C}$ for differentiating $(t_x, t_y) \in \Omega$, e. g., nuclear *size* is highly relevant for classifying benign and malignant tumor as important nuclei in malignant are larger than important nuclei in benign. Formally:

$$\begin{aligned} s_{\max}(t_x, t_y) &= \max_{c\in\mathcal{C}} S_{(t_x,t_y),c} \\ s_{\text{avg}}(t_x, t_y) &= \frac{1}{|C|}\sum_{c\in\mathcal{C}} S_{(t_x,t_y),c} \\ s_{\text{corr}}(t_x, t_y) &= \rho(S_{(t_x,t_y),c=1,..,|\mathcal{C}|}, P_{(t_x,t_y),c=1,..,|\mathcal{C}|}) \end{aligned} \tag{6.2}$$

where ρ denotes Pearson correlation. $s_{\max}, s_{avg} \in [0,\infty)$ show separation between unnormalized class-histograms; and $s_{corr} \in [-1, 1]$ shows agreement between S and P. We build $S_{\max}$, S_{avg} and S_{corr} by computing Equation (6.2) $\forall(t_x, t_y) \in \Omega$. These metrics' complementary nature may lead to relevant *concepts* different to pathological understanding.

Risk: We *conceptually* introduce the notion of risk as a weight to indicate the cost of misclassifying a sample of class t_x, erroneously as class t_y [Thai-Nghe et al., 2010; He et al., 2013]. Indeed, misclassifying a malignant tumor as a benign tumor is riskier than misclassifying it as an atypical tumor. Thus, we construct a risk vector $R \in \mathbb{R}^{\Omega}$. In this work, each entry in R defines the symmetric risk of differentiating t_x from t_y measured as the number of class-hops needed to evolve from t_x to t_y.

Metrics: Finally, we propose three quantitative metrics based on class separability to assess an explainer quality. The metrics are computed as the risk weighted sum of the statistics of separability scores, i. e.,, (1) *maximum separability* $S_{\max,R} := S_{\max} \odot R$, (2) *average separability* $S_{\text{avg},R} := S_{\text{avg}} \odot R$, (3) *correlated separability* $S_{\text{corr},R} := S_{\text{corr}} \odot R$, where $\odot$ defines the Hadamard product. The first two metrics are pathologist-independent, and the third metric requires expert pathologists to impart the domain knowledge in the form of pathological prior P. Such prior can be defined individually by a pathologist or collectively by consensus of several pathologists, and it is independent of the algorithm generated explanations.

6.4 Dataset

We experiment on BReAst Cancer Subtyping (BRACS), a large collection of breast tumor RoIs [Pati et al., 2022]. BRACS consists of 4391 RoIs at 40× resolution from 325 H&E stained breast carcinoma whole-slides. The RoIs are annotated by the consensus of three pathologists as, (1) Benign (B): normal, benign and usual ductal hyperplasia, (2) Atypical (A): flat epithelial atypia and atypical ductal hyperplasia, and (3) Malignant (M): ductal carcinoma *in situ* and invasive. The RoIs consist of an average #pixels=3.9 ± 4.3 million, and average #nuclei=1468 ± 1642, and are stain normalized using [Stanisavljevic et al., 2018]. The train, validation, and test splits are created at the whole-slide level, including 3163, 602, and 626 RoIs.

Algorithm 1 Class separability computation.

```
Input: $\mathcal{D} = \{(D_i^t, \mathcal{I}_i^t)\}, t \in \mathcal{T}, i \in N_t$ Parameters: $\mathcal{T}, \mathcal{C}, \mathcal{A}_c, \mathcal{K}$
Result: $S \in \mathbb{R}^{\binom{|\mathcal{T}|}{2} \times |\mathcal{C}|}$
  for c in $\mathcal{C}$ do                                                    ▷ go over concepts
      for k in $\mathcal{K}$ do                                                ▷ go over nuclei thresholds
          for a in $\mathcal{A}_c$ do                                          ▷ go over attributes
              for t in $\mathcal{T}$ do                                        ▷ go over classes
                  var $\leftarrow D_i^t(a)[: k]$                               ▷ sorted $I_i^t$
                  $H_t^{(k)}(a) \leftarrow$ histogram(var)
              end for
              for $(t_x, t_y)$ in $\binom{|\mathcal{T}|}{2}$ do                ▷ go over class pairs
                  $s_a^{(k)}(t_x, t_y) \leftarrow d(H_{t_x}^{(k)}(a), H_{t_y}^{(k)}(a))$
              end for
          end for
          $s_c^{(k)}(t_x, t_y) \leftarrow \frac{1}{|\mathcal{A}_c|} \sum_{a \in \mathcal{A}_c} s_a^{(k)}(t_x, t_y)$
      end for
      $S_{(t_x,t_y),c} \leftarrow \text{AUC}_{k \in \mathcal{K}}(s_c^{(k)}(t_x, t_y))$
  end for
```

6.5 Results

This section describes the analysis of CG explainability for breast cancer subtyping. We evaluate three types of graph explainers and quantitatively analyze the explainer quality using the proposed class separability metrics.

6.5.1 Implementation

We conducted our experiments using PyTorch [Paszke et al., 2019] and the Deep Graph Library (DGL) [Wang et al., 2019b]. The GNN architecture for CG classification is presented in Section 6.3.3. The CG classifier was trained for 100 epochs using Adam optimizer [Kingma et al., 2015], 10^{-3} learning rate and 16 batch size. The best CG-classifier achieved 74.2% weighted F1-score on the test set for the three-class classification. Average time for processing a 1K×1K RoI on a NVIDIA P100 GPU is 2s for CG generation and 0.01s for GNN inference.

6.5.2 Qualitative assessment

Figure 6.4 presents explanations, i. e., nuclei importance maps, from four studied graph explainers. We observe that GRAPHGRAD-CAM and GRAPHGRAD-CAM++ produce similar importance maps. The GNNEXPLAINER generates almost binarized nuclei importances. Interestingly, the gradient and pruning-based techniques consistently highlight similar regions. Indeed, the approaches focus on relevant epithelial region and unfocus on stromal nuclei and lymphocytes outside the glands. Differently, GRAPHLRP produces less interpretable maps through high spatial localization Figure 6.4(d) or less spatial localization Figure 6.4(h,l). Qualitative visual assessment of Figure 6.4 conclude that, (1) *fidelity* preserving explainers result differently based on the underlying mechanism, (2) high *fidelity* does not guarantee straightforward pathologist-understandable explanations, (3)

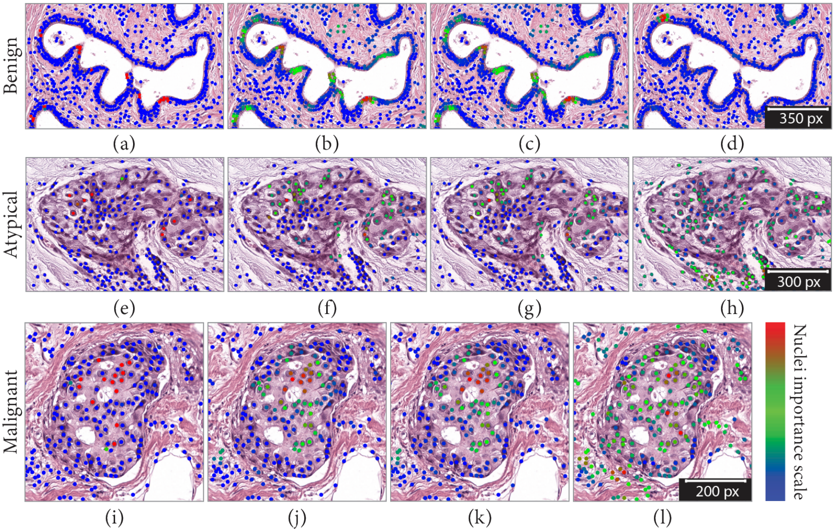


Figure 6.4: Qualitative results. The rows represent the cancer subtypes, i. e., Benign, Atypical and Malignant, and the columns represent the graph explainability techniques, i. e., GNNEXPLAINER, GRAPHGRAD-CAM, GRAPHGRAD-CAM++, and GRAPHLRP. Nuclei-level importance ranges from blue (the least important) to red (the most important).

qualitative assessment cannot rigorously compare explainers' quality, and (4) large-scale tedious pathological evaluation is inevitable to rank the explainers.

6.5.3 Quantitative assessment

For cancer subtyping, relevant *concepts* are nuclear morphology and topology [Rajbongshi et al., 2018; Kashyap et al., 2018; Nguyen et al., 2017; Allison et al., 2016]. Here, we focus on nuclear morphology, i. e., $\mathcal{C} = \{size, shape, shape\ variation, density, chromaticity\}$. Table 6.2 lists the *attributes* $\mathcal{A}_c, \forall c \in \mathcal{C}$. In our experiments, we select $\mathcal{K} = \{5, 10, ..., 50\}$ nuclei per CG. We further introduce a RANDOM explainer via *random* nuclei selection strategy per CG to assess a lower bound per quantitative metric. Table 6.1 presents the statistics of pair-wise class separability and aggregated separability w/ and w/o risk to assess the studied explainers quantitatively. Also, for each class pair (t_x, t_y), we compute classification accuracy by using the CGs of type t_x, t_y.

Noticeably, GNNEXPLAINER achieves the best *maximum* and *average separability* for majority of pair-wise classes. GRAPHGRAD-CAM++ and GRAPHGRAD-CAM followed GNNEXPLAINER except for (B vs. A), where GRAPHLRP outperforms them. All explainers outperform RANDOM which conveys that the quality of the explainers' explanations are better than random. Notably, GRAPHGRAD-CAM and GRAPHGRAD-CAM++ quantitatively perform very similarly, which is consistent with our qualitative analysis in Figure 6.4. Interestingly, a positive correlation is observed between pair-wise class accuracies and *average separability* for the explainers, i. e., better classification leads to better *concept* separability, and thus produces better explanations. Further, the observation does not hold for RANDOM generated explanations, which possesses undifferentiable *average concept* separability.

Table 6.1: Quantitative assessment of graph explainers: GnnExplainer, GraphGrad-CAM, GraphGrad-CAM++ and GraphLRP, using proposed *maximum*, *average*, and *correlated separability* metrics. Results are provided for each pair-wise breast subtyping tasks, and are aggregated w/o and w/ risk weighting, i. e., $S_{\max}$ and $S_{\max,R}$. The first and second best values are indicated in **bold** and underline.

Tasks (Ω)		B vs. A	B vs. M	A vs. M	B vs. A vs. M			
Accuracy (in %)		77.19	90.29	80.42	74.92			
Explainer	Metric	$\forall\,(t_x, t_y) \in \Omega$ (↑)			Agg. Metric w/o Risk (↑)		Agg. Metric w/ Risk (↑)	
GnnExplainer	$s_{\max}(t_x, t_y)$	**3.26**	**6.24**	**3.48**	$S_{\max}$	**12.98**	$S_{\max,R}$	**19.22**
GraphGrad-CAM		1.24	4.41	3.36		9.01		13.42
GraphGrad-CAM++		1.27	4.42	3.40		9.09		13.51
GraphLRP		2.33	2.46	1.28		6.07		8.53
Random		1.02	1.26	1.11		3.39		4.65
GnnExplainer	$s_{\text{avg}}(t_x, t_y)$	**1.54**	**2.78**	1.93	S_{avg}	**6.25**	$S_{\text{avg},R}$	**9.03**
GraphGrad-CAM		1.15	2.57	2.08		5.80		8.37
GraphGrad-CAM++		1.18	2.58	**2.09**		5.85		8.43
GraphLRP		1.38	1.59	1.47		4.44		6.03
Random		1.05	1.00	0.95		3.00		4.00
GnnExplainer	$s_{\text{corr}}(t_x, t_y)$	−0.02	0.36	0.38	S_{corr}	0.72	$S_{\text{corr},R}$	1.08
GraphGrad-CAM		−0.01	0.57	0.58		1.14		1.71
GraphGrad-CAM++		**−0.01**	**0.58**	**0.59**		**1.16**		**1.74**
GraphLRP		−0.15	−0.49	−0.23		−0.87		−1.36
Random		−0.37	−0.31	−0.18		−0.86		−1.17

Table 6.2: Quantification of *concepts* for pair-wise and aggregated class separability in GnnExplainer. The first and second best values are indicated in **bold** and underline. The per-*concept attributes* are presented in the first column.

Concept (Attributes) / Tasks (Ω)	B vs. A	B vs. M	A vs. M	w/o risk (↑)	w/ risk (↑)
Size (area)	**3.26**	**6.24**	**3.48**	**12.98**	**19.22**
Shape (perimeter, roughness, eccentricity, circularity)	1.27	2.23	1.60	5.10	7.34
Shape variation (shape factor)	0.69	2.30	1.99	4.97	7.28
Density (mean density, std density)	1.01	0.80	0.52	2.33	3.14
Chromaticity (GLCM contrast, homogeneity, ASM, entropy, variance)	1.44	2.31	2.07	5.82	8.13
Average separability (↑)	1.54	2.78	1.93	6.25	9.03

To obtain pathological prior to compute *correlation separability*, we consulted three pathologists to rank the *concepts* in order of their relevance for discriminating each pair of classes. For instance, given an atypical RoI, we asked how important is nuclear *shape* to classify the RoI as *not* benign and *not* malignant. Acquired *concept* ranks for each class pair are *min-max* normalized to output prior matrix P. We observe that GnnExplainer, GraphGrad-CAM, and GraphGrad-CAM++ have positive *correlated separability* for (B vs. M), (A vs. M), and nearly zero values for (B vs. A). It shows that the explanations for (B vs. M) and (A vs. M) bear similar order of *concepts* as the pathologists, and focus on a different order of *concepts* for (B vs. A). GraphGrad-CAM++ has the best overall agreement at the *concept*-level with the pathologists, followed by GraphGrad-CAM and GnnExplainer. Random agrees significantly worse than the three explainers, and GraphLRP has the least agreement. Table 6.2 provides more insights by highlighting the per-*concept* metrics of GnnExplainer. Nuclei *size* is the most relevant *concept*, followed by *chromaticity* and *shape variation*. Comparatively nuclear *density* is the least relevant.

6.6 Conclusion

In this work, we presented an approach for explaining black-box DL solutions in computational pathology. We advocated for biological entity-based analysis instead of conventional pixel-wise analysis, thus providing an intuitive space for pathological understanding. We employed four graph explainability techniques, i. e., graph pruning (GnnExplainer), gradient-based saliency (GraphGrad-CAM, GraphGrad-CAM++) and layerwise relevance propagation (GraphLRP), to explain "black-box" GNNs processing the entity-graphs. We proposed a novel set of user-independent quantitative metrics expressing pathologically-understandable *concepts* to evaluate the graph explainers, which relaxes the exhaustive qualitative assessment by expert pathologists. Our analysis concludes that the explainer bearing the best class separability in terms of *concepts* is GnnExplainer, followed by GraphGrad-CAM++ and GraphGrad-CAM. GraphLRP is the worst explainer in this category while outperforming a randomly created explanation. We observed that the explainer quality is directly proportional to the GNN's classification performance for a pair of classes. Furthermore, GraphGrad-CAM++ produces explanations that best agrees with the pathologists in terms of *concept* relevance, and objectively highlights the relevant set of *concepts*. Considering the expansion of entity-graph-based processing in the domains of radiology, computation biology, satellite and natural images etc., graph explainability and their quantitative evaluation is crucial. The proposed method encompassing domain-specific user-understandable terminologies can potentially be of great use in this direction. It is a meta-method that is applicable to other domains and tasks by incorporating relevant entities and corresponding *concepts*. For instance, with entity-graph nodes denoting body parts of cars in Stanford Cars [Krause et al., 2013]/ Human poses [Andriluka et al., 2014], and expert knowledge available on car-model/ human activity, our method can infer relevant entities by quantifying their agreement with experts.

6.7 Appendices

6.7.1 BRACS dataset

In this paper, the BRACS dataset is used to analyze CG explainability for breast cancer subtyping. The pixel-level and entity-level statistics of the dataset are presented in Table 6.3. Training, validation, and test splits are created at the whole-slide level for conducting the experiments. The details of the class-wise distribution of images in each split are presented in Table 6.3.

6.7.2 Concepts and Attributes

In this paper, we focus on pathologically-understandable nuclear *concepts* $\mathcal{C}$ pertaining to nuclear morphology for breast cancer subtyping. To quantify each $c \in \mathcal{C}$, we use several measurable *attributes* $\mathcal{A}_c$. Table 6.4 presents the list of *concepts* and corresponding *attributes* used to perform the proposed quantitative analysis in this work. Also, Table 6.4 includes the class-wise expected criteria for each *concept*.

Table 6.3: Statistics of BRACS dataset.

	Metric	Benign	Atypical	Malignant	Total
Image	Number of images	1741	1351	1299	4391
	Number of pixels (in million)	3.9±3.5	1.62±1.5	6.35±5.2	3.9±4.3
	Max/Min pixel ratio	180.1	75.3	128.6	235.6
CG	Number of nodes	1331±1134	635±510	2521±1934	1468±1642
	Number of edges	4674±4131	2309±2110	8591±7646	5102±6089
	Max/Min node ratio	312.5	416.7	312.5	434.8
Image split	Train	1231	1008	928	3163
	Validation	261	162	179	602
	Test	249	185	192	626

The *attributes* of the nuclei in a TRoI are computed as proposed in [Parvatikar et al., 2020]. It uses the TRoI and corresponding nuclei segmentation map, denoted as I_{seg}. Area of a nucleus x, denoted as $A(x)$, is defined as the number of pixels belonging to x in I_{seg}. $P(x)$, the perimeter of x, is measured as the contour length of x in I_{seg}. $P_{\text{ConvHull}}(x)$, the convex hull perimeter of x, is the contour length of convex hull induced by x in I_{seg}. The major and minor axis of x, noted as $a_{\text{major}}(x)$ and $a_{\text{minor}}(x)$, are the longest diameter of x and the longest line segment perpendicular to $a_{\text{major}}(x)$, respectively. The chromatin *attributes* are computed from the normalized gray level co-occurrence matrix (GLCM) [Haralick et al., 1973], which captures the probability distribution of co-occurring gray values in x.

6.7.3 Quantitative assessment

In this section, we analyze two key components of the proposed quantitative metrics: the histogram construction and class separability scores for threshold set $\mathcal{K}$. Further, we relate the analysis to class-wise expected criteria for each *concept*, as shown in Table 6.4.

Histogram analysis: Histogram construction is a key component in the proposed quantitative metrics. Figure 6.5 presents per-class histograms for each explainer and the best *attribute* per *concept*. We set the importance threshold to $k = 25$, i. e., for each TRoI, we select 25 nuclei with the highest node importance. The best *attribute* for a *concept* is the one with the highest average pair-wise class separability.

The row-wise observation exhibits that GnnExplainer and GraphLRP provide, respectively, the maximum and the minimum pair-wise class separability. The histograms for a *concept* and for an explainer can be analyzed to assess the agreement between the selected important nuclei *concept*, and the expected *concept* behavior as presented in Table 6.4, for all the classes. For instance, nuclear *area* is expected to be higher for malignant TRoIs than benign ones. The *area* histograms for GnnExplainer, GraphGrad-CAM and GraphGrad-CAM++ indicate that the important nuclei set in malignant TRoIs includes nuclei with higher area compared to benign TRoIs. Similarly, the important nuclei in malignant TRoIs are expected to be vesicular, i. e., high texture entropy, compared to

Table 6.4: Pathologically-understandable nuclear *concepts*, corresponding measurable *attributes*, and computations are shown in Columns 1, 2, 3, respectively. The expected *concept* behavior for three breast cancer subtypes is shown in Columns 4, 5, 6, respectively.

Concept ($\mathcal{C}$)	Attribute ($\mathcal{A}$)	Computation	Benign	Atypical	Malignant
Size	Area	$A(x)$	Small	Small-Medium	Medium-Large
Shape	Perimeter	$P(x)$	Smooth	Mild irregular	Irregular
	Roughness	$\frac{P_{\text{ConvHull}}(x)}{P(x)}$			
	Eccentricity	$\frac{a_{\text{minor}}(x)}{a_{\text{major}}(x)}$			
	Circularity	$\frac{4\pi A(x)}{P(x)^2}$			
Shape variation	Shape factor	$\frac{4\pi A(x)}{P^2_{\text{ConvHull}}}$	Monomorphic	Monomorphic	Pleomorphic
Spacing	Mean spacing	$\text{mean}(d_y \mid y \in \text{kNN(x)})$	Evenly crowded	Evenly spaced	Variable
	Std spacing	$\text{std}(d_y \mid y \in \text{kNN(x)})$			
Chromatin	GLCM dissimilarity	$\sum_i \sum_j \lvert i-j \rvert p(i,j)$	Light euchromatic	Hyperchromatic	Vesicular
	GLCM contrast	$\sum_i \sum_j (i-j)^2 p(i,j)$			
	GLCM homogenity	$\sum_i \sum_j \frac{p(i,j)}{1+(i-j)^2}$			
	GLCM ASM	$\sum_i \sum_j p(i,j)^2$			
	GLCM entropy	$-\sum_i \sum_j p(i,j) \log(p(i,j))$			
	GLCM variance	$\sum_i \sum_j (i-\mu_i)^2 p(i,j)$ with $\mu_i = \sum_i \sum_j i p(i,j)$			

light euchromatic, i. e., moderate texture entropy, in benign TRoIs. The *chromaticity* histograms for GNNEXPLAINER, GRAPHGRAD-CAM and GRAPHGRAD-CAM++ display this behavior. Additionally, the histogram analysis can reveal the important *concepts* and important *attributes*. For instance, nuclear *density* proves to be the least important *concept* for differentiating the classes.

Separability score for threshold set $\mathcal{K}$: Multiple importance thresholds $\mathcal{K}$ are required to address the varying notion of important nuclei across different cell graphs and different explainers. Figure 6.6 presents the behavior of pair-wise class separability for using various $k \in \mathcal{K} = \{5, 10, ..., 50\}$. For simplicity, we present the behavior for the best *attribute* per *concept*. In general, the pair-wise class separability is observed to decrease with decreasing k. Intuitively, decreasing k results in including more unimportant nuclei into the evaluation, thereby gradually decreasing the class separability.

The degree of agreement between the difference in the expected behavior per *concept* and the pair-wise class separability in Figure 6.6, for all pair-wise classifications and various $k \in \mathcal{K}$ can be used to assess the explainer's quality. For instance, according to Table 6.4, the difference in the expected nuclear *size* can be considered as benign–atypical $<$ benign–malignant, and atypical–malignant $<$ benign–malignant. GNNEXPLAINER, GRAPHGRAD-CAM and GRAPHGRAD-CAM++ display these behaviors $\forall k \in \mathcal{K}$. GNNEXPLAINER provides the highest class separability in each pair-wise classification, thus proving to be the best explainer pertaining to *size concept*. Detailed inspection of Figure 6.6 shows that all the differences in the expected behavior, per *concept* for all pair-wise classifications, is inline with the *concept*-wise expected behavior in Table 6.4,

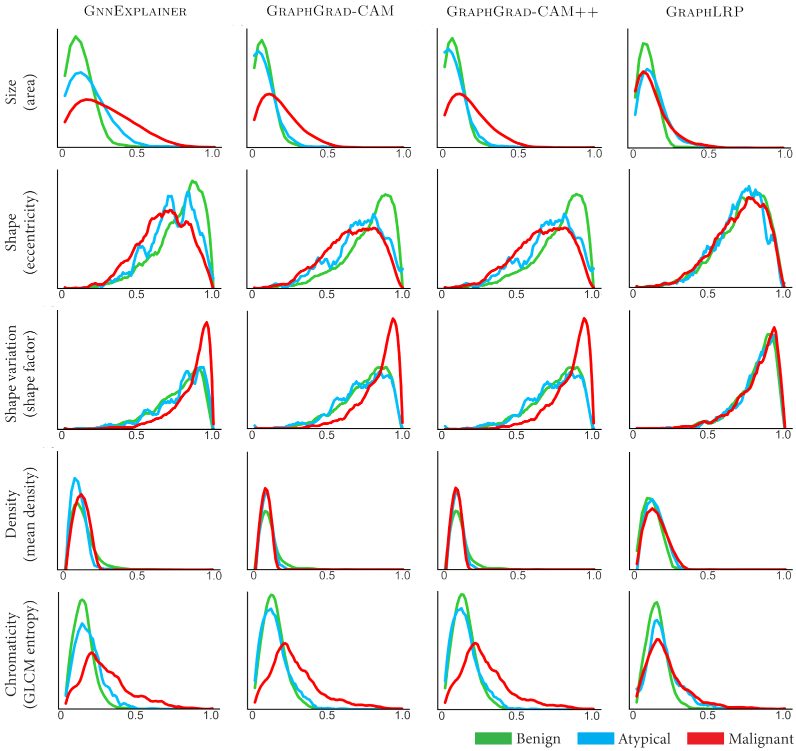


Figure 6.5: Per-class histograms for different *concepts* across different graph explainers. For simplicity, histograms are shown for the best *attribute* per *concept* at importance threshold $k = 25$.

$\forall c \in \mathcal{C}$ and $\forall k \in \mathcal{K}$. Overall, GnnExplainer is seen to be the best explainer as it agrees to the majority of the expected differences $\forall c \in \mathcal{C}$ for all pair-wise classifications, while providing high-class separability. Furthermore, *size* proves to be the most important *concept* that provides the maximum class separability across all pair-wise classifications.

6.7.4 Qualitative assessment

Figure 6.7 and Figure 6.8 present CG explanations produced by GnnExplainer, GraphGrad-CAM, GraphGrad-CAM++, and GraphLRP for TRoIs across benign, atypical and malignant breast tumors. It can be observed that GnnExplainer learns to binarize the explanations, thereby producing the most compact explanations by retaining the most important nuclei set of nuclei with high importance. However, GraphGrad-CAM and GraphGrad-CAM++ produce explanations with more distributed nuclei importance than GnnExplainer. GraphLRP produces the largest explanations by retaining most of the nuclei in the CGs.

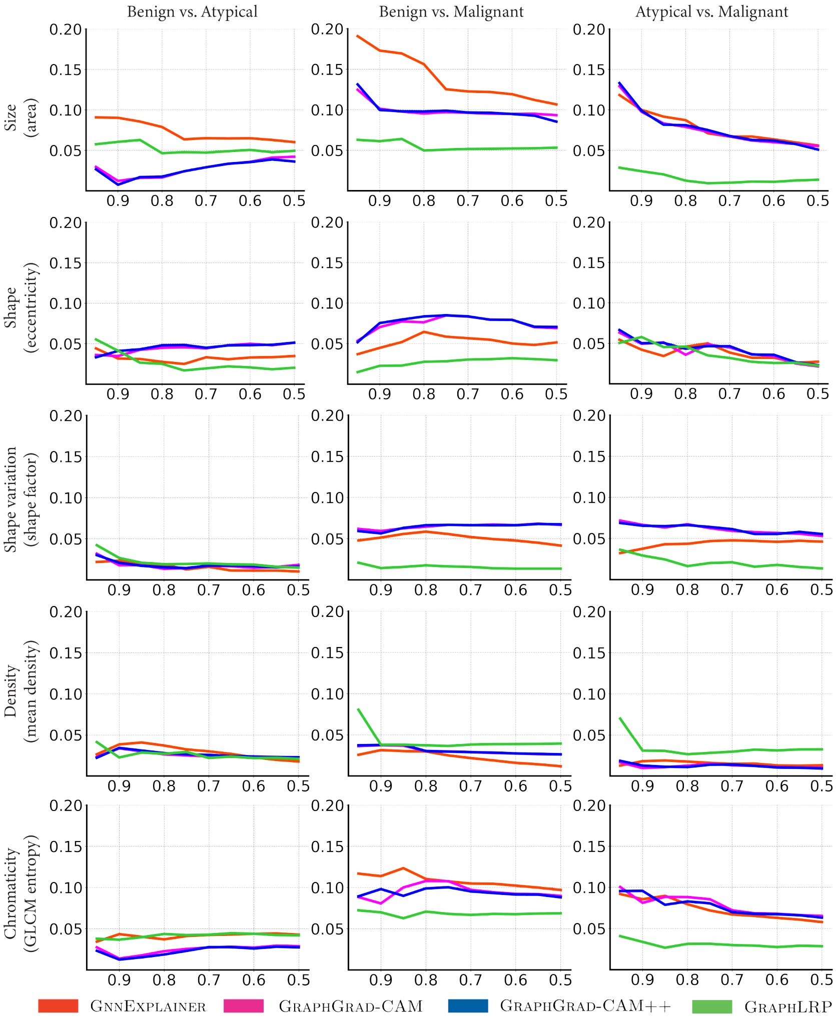


Figure 6.6: Visualizing the variation of pair-wise class separability score (Y-axis) w.r.t. several nuclei importance thresholds in $\mathcal{K}$ (X-axis). The analysis is provided for different graph explainers, and for the best *attribute* per *concept*.

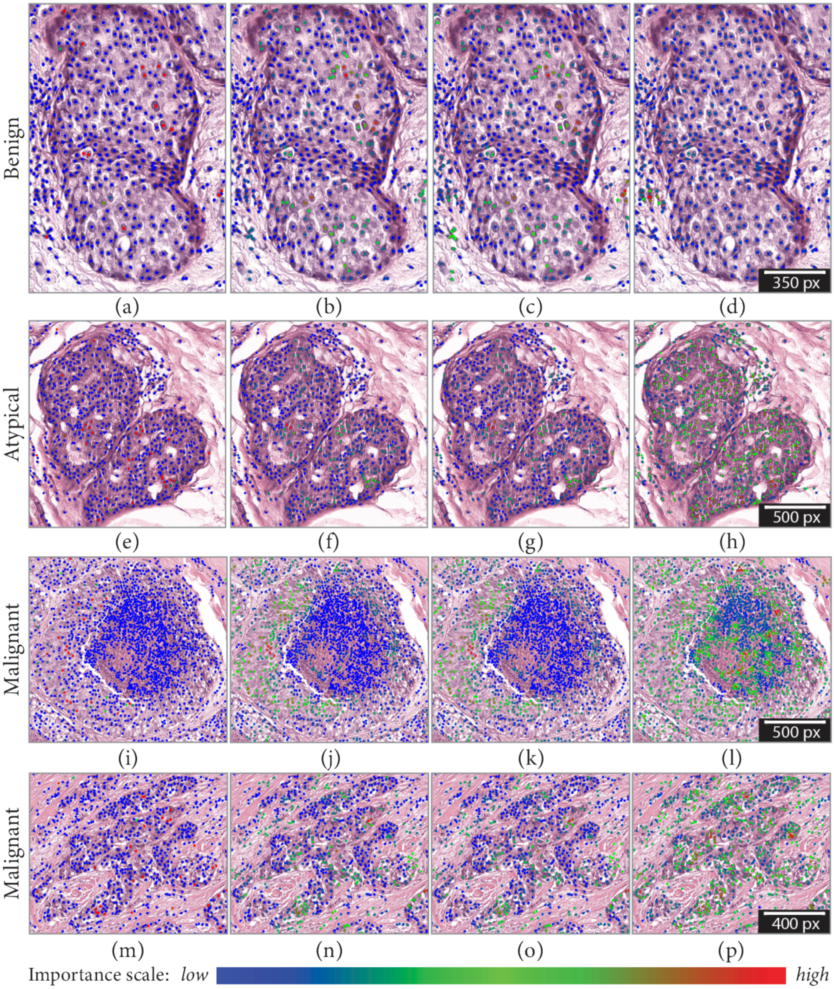


Figure 6.7: Qualitative results. The rows represent breast cancer subtypes, and columns represent graph explainers, i. e., GNNEXPLAINER, GRAPHGRAD-CAM, GRAPHGRAD-CAM++, and GRAPHLRP. Nuclei level importance ranges from blue (the least important) to red (the highest important).

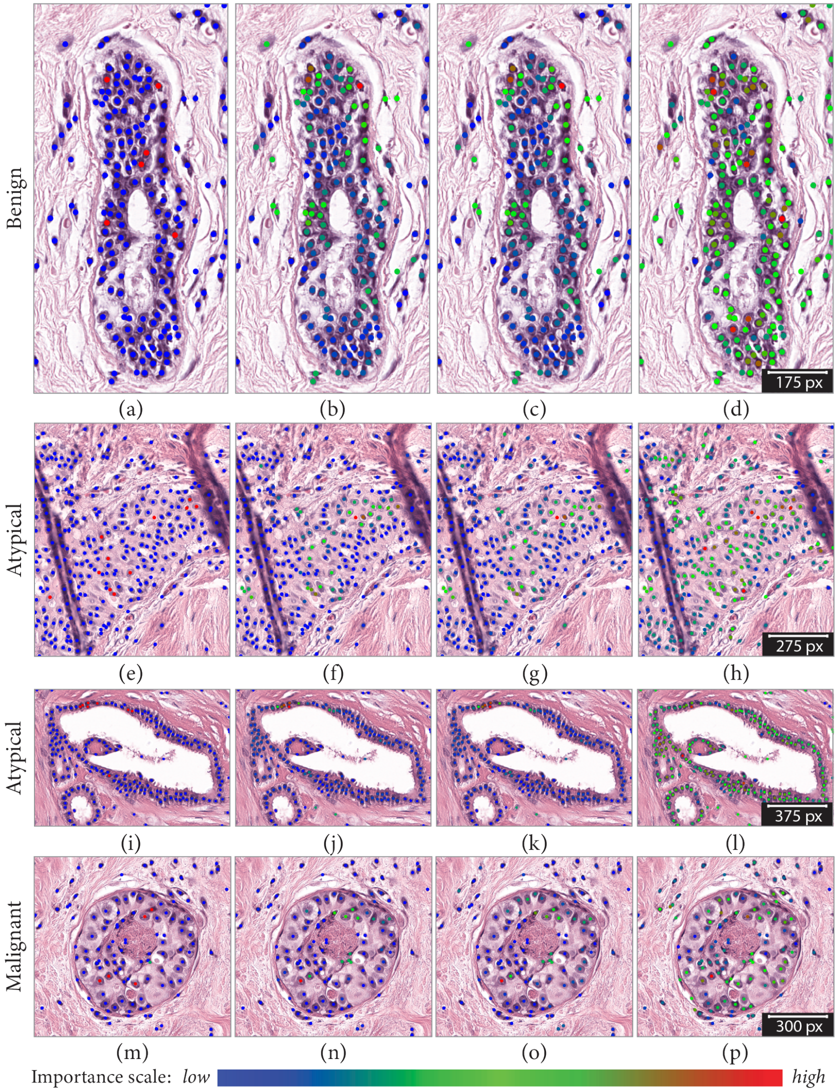


Figure 6.8: Qualitative results. The rows represent breast cancer subtypes, and columns represent graph explainers, i. e., GNNEXPLAINER, GRAPHGRAD-CAM, GRAPHGRAD-CAM++, and GRAPHLRP. Nuclei level importance ranges from blue (the least important) to red (the highest important).

7

Reducing Annotation Effort in Digital Pathology: Co-Representation Learning Framework for Classification Tasks

Classification of digital pathology images is imperative in cancer diagnosis and prognosis. Recent advancements in deep learning and computer vision have greatly benefited the pathology workflow by developing automated solutions for classification tasks. However, the cost and time for acquiring high quality task-specific large annotated training data are subject to intra- and inter-observer variability, thus challenging the adoption of such tools. To address these challenges, we propose a classification framework via co-representation learning to maximize the learning capability of deep neural networks while using a reduced amount of training data. The framework captures the class-label information and the local spatial distribution information by jointly optimizing a categorical cross-entropy objective and a deep metric learning objective, respectively. A deep metric learning objective boosts the classification, especially in the low training data regime. Further, a neighborhood-aware multiple similarity sampling strategy, and a soft-multi-pair objective that optimizes interactions between multiple informative sample pairs, is proposed to accelerate deep metric learning. We evaluate the proposed framework on five benchmark datasets from three digital pathology tasks, i. e., nuclei classification, mitosis detection, and tissue type classification. For all the datasets, our framework achieves state-of-the-art performance by using approximately 50% of the training data. On using complete training data, the proposed framework outperforms the state-of-the-art on all the five datasets.

7.1 Introduction

Histopathological analysis is a common clinical procedure for diagnosing the presence, type, and progression of diseases such as cancer. Pathologists manually identify and examine the nuclear phenotype, tissue topology, and cytology among several other factors for the staging and grading of cancer. In clinical practice, an inspection of tissue slides under a microscope is tedious, time-consuming, and subject to inter- and intra-observer variability, e. g., concordance rates can be as low as 48% for some cases of breast cancer atypia ([Elmore et al., 2015]). As a consequence of the recently gained efficiency in

---

This chapter has been published as: Pushpak Pati, Antonio Foncubierta-Rodriguez, Orcun Goksel, Maria Gabrani "Reducing Annotation Effort in Digital Pathology: A Co-Representation Learning Framework for Classification Tasks", In: Medical Image Analysis, vol. 67, pp. 101859, 2021, doi: 10.1016/j.media.2020.101859

scanning techniques that digitize glass slides into impressive resolution images, digital pathology (DP) has profoundly transformed the daily practice of pathologists. DP facilitates remote diagnostic work, teleconsultation, workload efficiency, collaborations, central clinical review, and virtual education ([Wilbur et al., 2009; Hamilton et al., 2012; Sagun et al., 2018; Nauhria et al., 2019; Pantanowitz et al., 2018; Hanna et al., 2019]). Additionally, DP promotes innovative research opportunities in image analysis and computing to automate cancer diagnosis ([Litjens et al., 2017]).

Recent advancements in computer vision and deep learning (DL) have enabled to learn sub-visual image features that may not be easily discernible by the human eye. DL algorithms learn representations directly from images and have demonstrated superior performance compared to handcrafted feature-based methods [Litjens et al., 2017]. When applied to medical images, such as DP images, these algorithms offer the opportunity for better quantitative modeling of disease appearance, and hence improved diagnosis and patient outcome. DL methods have successfully addressed several tasks in DP by substantially reducing the laborious and tedious nature of providing accurate quantification, and by reducing the observer variability.

However, these compelling opportunities in DP come with their own set of challenges ([Tizhoosh et al., 2018]). Most DL algorithms require large sets of high quality labeled data, and they do not generalize to data deviating from the training dataset. Further, variations in tissue acquisition, staining procedure, scanning technology, and a high degree of polymorphism in tissue structures across organs in DP hinder the generalization of DL. To address these difficulties, large datasets need to be acquired encompassing all potential variations. Further, DL algorithms belong to the class of narrow AI algorithms. They are designed to perform only one task, thus requiring task-specific large annotated data. Though several multi-task learning algorithms have been proposed in the literature ([Graham et al., 2019a; Yan et al., 2019]), still they are limited to be useful for very closely related tasks. Annotating such large datasets by experts is time-consuming and expensive. Annotation procedure becomes highly complicated in presence of low-resolution images, artifacts, or ambiguous features. Alternative solutions such as crowd-sourcing may be cheaper and quicker but have the potential to introduce noise ([Alialy et al., 2018]).

The aforementioned challenges underline the need for learning strategies to maximally utilize the dataset, especially for scenarios with scarce labeled data. Several techniques in machine learning tackle the scarce data challenge, such as data augmentation, data generation, transfer learning, semi-supervised learning, and active learning. Generative adversarial networks (GANs, [Goodfellow et al., 2014]) have been a potential solution to synthesize labeled data. However, generating high-resolution images incorporating complex medical structures, training instability, and non-convergence inhibit GAN's application to medical imaging ([Yi et al., 2019]). GANs also require a substantial amount of data to train, and visual inspection by experts for model selection ([Yu et al., 2017]). Unified generator for multiple classes results in inferior performance compared to class-specific generators [Frid-Adar et al., 2018]. Furthermore, [Finlayson et al., 2018] argue that images generated from GANs may serve as an effective augmentation in the medium-data regime, but may not help in high or low-data regime. Transfer learning from natural images has become a de-facto method for DL applications to medical imaging for small labeled data. Evaluation of the effect of transfer concludes that transfer offers little benefit to performance. Some differences from transfer learning are due to the over-parameterization of standard models rather than sophisticated feature reuse. [Raghu et al., 2019] present these observations and emphasize on more efficient model exploration. Active learning is more tightly coupled with human-in-the-loop learning that requires an annotator to annotate unlabeled data during the training phase as presented by [Budd et al., 2021] and [Lutnick et al., 2019]. Further, [Schaumberg et al., 2016] introduce a non-intrusive approach to approximate a pathologist's routine clinical work and generate large annotated dataset in a non-algorithmic manner. However, the availability of an expert in the training loop or mimicking a pathologist for each task is expensive and may not be feasible. Semi-supervised learning iteratively augments the training data by annotating the unlabeled data during the training phase. The performance of these algorithms rely on the base classification performance for label assignment.

In this paper, we propose a co-representation learning (CoReL) framework for classification tasks to extract maximum information from the training data via multiple learning strategies. CoReL aims at enhancing the classification performance, especially for low training data regime. CoReL is a generic framework that can be employed in generative modeling, semi-supervised learning or active learning setting to further boost the performance of individual settings. Additionally, CoReL can be extended to a wide-range of classification tasks ranging from nuclei classification to whole-slide-image classification. In regards, the major contributions in this paper are,

- We propose a CoReL for classification framework that leverages class label information via optimizing categorical cross-entropy, and spatial distribution information of samples in the embedding space via optimizing a deep metric learning objective. The framework improves the classification compared to a standalone cross-entropy based classifier without additional training parameters and and inference time.
- We propose a neighborhood-aware multiple similarity sampling, a novel pair mining strategy that utilizes context information of samples and pair-wise similarity measures to identify informative pairs. Further, a soft-multi-pair objective is proposed that jointly optimizes the interactions of multiple positive and negative pairs in a triplet setting to accelerate deep metric learning.
- We evaluate the CoReL framework on five benchmark datasets across three DP classification tasks, i. e., nuclei classification, mitosis detection, and tissue type classification. The proposed deep metric learning strategy along with the joint learning technique improve the performance of the CoReL framework. The improvement is significant especially in the low training data regime. CoReL achieves the state-of-the-art performances on all datasets by using substantially reduced training data. For using the complete training data, it outperforms the current state-of-the-art approaches on all five datasets.

7.2 Related work

In the past few years, DL algorithms have addressed various classification tasks in DP. The applications include nuclei classification ([Sirinukunwattana et al., 2016; Graham et al., 2019a; Pati et al., 2020]), nuclei detection ([Sirinukunwattana et al., 2016; Ciresan et al., 2013]), tissue classification ([Kather et al., 2019; Xu et al., 2019b]) tumor staging ([Spanhol et al., 2016b; Aresta et al., 2019; Pati et al., 2020]), tumor grading ([Veta et al., 2019; Tellez et al., 2021]), tumor detection ([Liu et al., 2017b; Pati et al., 2018]), outcome prediction ([Kather et al., 2019; Bychkov et al., 2018]) etc. DL algorithms outperform earlier handcrafted feature-based approaches by automatically learning robust representations. In this work, we review nuclei classification, mitosis detection and tissue classification tasks.

[Sirinukunwattana et al., 2016] builds a spatially constrained CNN and a neighboring ensemble predictor coupled with a CNN to respectively detect and classify nuclei. [Shao et al., 2018] uses deep active learning with pairwise constraints to actively query the most valuable nuclei for annotation by an expert and subsequently updates a CNN by incorporating the newly annotated samples. [Hamad et al., 2018] presents a two-stage pipeline by combining a fully convolutional regression network with a CNN for nuclei localization and classification. [Li et al., 2019c] uses a position of interest network with a cascade residual fusion block to localize nuclei and categorizes the nuclei by a multi-cropping network. [Zhou et al., 2018] presents a sibling fully convolutional network with prior objectness interaction to simultaneously detect and classify nuclei. [Graham et al., 2019a] proposes a CNN, leveraging information encoded within the vertical and horizontal distances of nuclear pixels for simultaneous nuclei segmentation and classification.

Several contests, including ICPR12 ([Roux et al., 2013]), AMIDA13 ([Veta et al., 2015]), MITOS-ATYPIA-14 ([Roux, 2014]) and TUPAC16 ([Veta et al., 2019]), have promoted remarkable advances in the area of automatic mitosis detection. [Ciresan et al., 2013] proposes a computationally expensive sliding-window-based detection method. [Chen et al., 2016] uses a fully convolutional neural network for a coarse retrieval of mitosis candidates, and a fine discrimination model utilizing knowledge transferred from cross-domain to identify mitoses from hard mimics. [Li et al., 2018a] applies a proposal-based deep detection network for mitosis detection and a patch-based deep verification network to improve the predictions. [Li et al., 2019a] proposes a concentric loss based semantic segmentation approach to identify mitoses. [Paeng et al., 2016] trains a large-view CNN using mitotic and non-mitotic patches, and proposes a fully convolutional network-based inference to predict on a large image with a single forward pass.

[Kather et al., 2019] utilizes CNN for classifying tissue patches from colorectal cancer into nine tissue categories. [Xu et al., 2019b] integrates a CNN with a focal loss to identify tissue composition in colorectal cancer.

Several research works have improved various DP classification tasks using large annotated data. However, only a few efforts based on active learning ([Carse et al., 2019; Budd et al., 2021; Lutnick et al., 2019]), semi-supervised learning ([Akram et al., 2018; Peikari et al., 2018]) and deep metric learning ([Teh et al., 2020; Han et al., 2017]) have been dedicated to tackle the issue of limited training data. Though transfer learning-based approaches have been adopted in this regard, [Raghu et al., 2019] discards the benefit of transfer and emphasizes on efficient model exploration. [Tizhoosh et al., 2018] lists a set of challenges in annotating large datasets from a pathology perspective, and mentions its impact on learning algorithms from an engineering perspective. Annotating large datasets may incorporate human error, such as data processing error, and incomplete annotation error. For instance, annotation errors are identified in the popular CAMELYON16 lymph node metastasis detection challenge by [Liu et al., 2017b]. Additionally, acquiring a large labeled data with balanced classes is very difficult for DP tasks which can adversely impact the learning of classifiers ([Johnson et al., 2019]).

Recently, a few frameworks have been proposed that jointly optimize a classification objective via categorical cross-entropy (CCE) and a similarity constraint via a deep metric learning (DML) objective for better classification performance on natural images. Both objectives constrain the embedding space independently, thus providing complementary information. [Zhang et al., 2016b] uses multi-level relevance through quadruplets to optimize a multi-task learning framework to effectively learn fine-grained feature representation. [Liu et al., 2017a] proposes an (N+M)-tuplet cluster loss and identity-based sampling to improve DML, and combines DML with softmax loss in a unified two fully connected layer branched framework to improve facial expression recognition. [Li et al., 2018d] proposes a DL framework to jointly optimize softmax loss and a pair loss objective. These frameworks include dataset-specific heuristically tailored sampling strategies and similarity constraints for improving DML performance. Our proposed CoReL framework consists of a generic sampling strategy and a generic deep metric objective. Through ablation studies on uncorrelated classification tasks, we demonstrate the generic learning ability of our framework. We further extend the CoReL framework to address limited training data scenario.

Recent advancements in DML ([Kaya et al., 2019]) have paved way for learning discriminatory networks using limited training data. DML measures the similarity among samples in the embedding space while using an optimal distance metric for learning tasks. The success of DML relies on the capacity of the embedding network to understand the similarity relationship among samples, an informative sample selection strategy and an appropriate distance metric objective. [Hadsell et al., 2006] proposes a Siamese network optimizing a contrastive loss that encourages pair of samples from the same class (positive pair) to be as close as possible, while pushing away pair of samples from different classes (negative pair) beyond a certain margin. It fixes a constant margin for all pairs of negative samples, thus embeds visually diverse classes and visually similar classes in the same small embedding space without allowing for distortions. [Schroff et al., 2015] introduces a triplet loss that aims to keep all positives closer to any negatives for each example. This formulation allows the embedding space to be arbitrarily distorted and does not impose a constant margin. However, triplet loss fails to utilize the full mini-batch information as it uses separate pairs consisting of one positive and one negative sample per anchor. [Oh Song et al., 2016] proposes a lifted structure loss to improve a mini-batch optimization using all pairs available in the batch. However, the lifted structure loss mines an equal number of positive and negative pairs randomly, thus discards a large number of informative negative pairs arbitrarily. [Sohn, 2016] proposes multi-class N-pair loss, similar to lifted structure loss, in the sense that it recruits multiple negative samples in a given mini-batch to compute the loss term. [Wang et al., 2019b] proposes a multi-similarity loss to utilize meaningful pair-wise relations by jointly considering the self-similarity and the relative similarities among pairs. In this work, we propose a soft-multi-pair loss that jointly optimizes multiple positive and negative pair interactions to accelerate and improve DML. Similarly, various sampling strategies have been proposed to accelerate DML convergence. [Chang et al., 2017] introduces an active bias learning to emphasize on high variance samples. Self-paced learning [Kumar et al., 2010], which pays more attention on samples with a higher confidence, is explored to design noise-robust algorithms by [Jiang

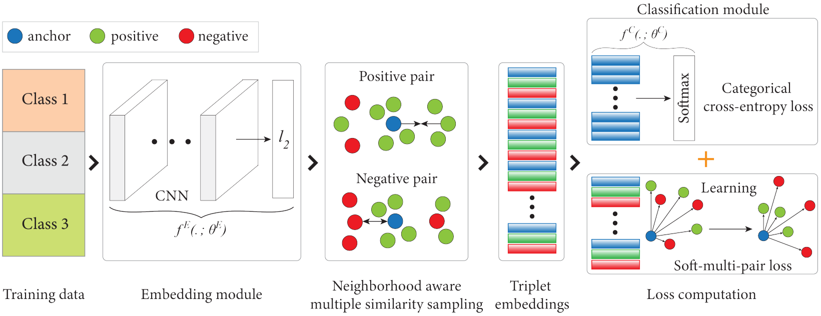


Figure 7.1: Overview of the proposed co-representation learning for classification framework that jointly optimizes categorical cross-entropy objective, computed using the softmax output, and a deep metric objective, computed using a proposed neighborhood-aware multiple similarity sampling with proposed soft-multi-pair loss.

et al., 2014]. [Schroff et al., 2015] proposes to use a semi-hard negative mining to yield negative samples that are fairly hard but not too hard. [Hermans et al., 2017] proposes to sample the hardest positive and the hardest negative samples within a mini-batch to form triplets. However, mining hard negative samples often leads to collapsed models. [Wu et al., 2017] exploits the distribution of pair-wise distances on an unit sphere to weight positive and negative pairs. This scheme only leverages the self-similarity of pairs. [Wang et al., 2019b] uses both self-similarity and relative similarities of pairs to sample informative pairs. We propose a neighborhood-aware multiple-similarity sampling strategy that weights and mines a pair of samples by using the neighborhood of the samples within a local data distribution, the self-similarity of the samples, and the relative similarities of the pair to equivalent pairs in the embedding space.

7.3 Background

In this work, we propose a CoReL framework to capture class label information and local spatial distribution information of samples via jointly optimizing a CCE and a DML objective. In this section, we explain the building blocks of our framework. We begin with the notations and terminologies, and then introduce the CCE and DML objectives. Additionally, we present the notion of complementary information captured by both the objectives from an information theory perspective.

7.3.1 Preliminaries

Let $\mathbf{X} = \{x_i\}_{i=1}^{M}$ is a data matrix of M sample images, and $\mathbf{Y} = \{y_i\}_{i=1}^{M}$ denotes the corresponding ground-truth vector. $x_i \in \mathbb{R}^{H \times W \times 3}$ is the i-th RGB image with height H and width W. $y_i \in C$ is the class label for x_i, where C is the set of classes. x_i is projected onto a unit sphere in a D-dimensional space by an embedding module, $f^E(.;\theta^E) : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^D$. Further, a classification module $f^C(.;\theta^C) : \mathbb{R}^D \rightarrow \mathbb{R}^C$, operates on $f^E(x_i;\theta^E)$ to predict $p(\hat{y}_i = c|x_i), \forall c \in C$, where $\hat{y}_i$ is the predicted class label for x_i. f^E and f^C are neural networks parameterized by θ^E and θ^C, respectively. For simplicity, we denote $f^E(.;\theta^E)$ and $f^C(.;\theta^C)$ as $f^E(.)$ and $f^C(.)$, respectively. The feature representations $f^E(x_i), \forall x_i \in \mathbf{X}$, are normalized to have unit length for training stability ([Schroff et al., 2015]).

7.3.2 Categorical cross-entropy

CCE is widely used in classification tasks. $f^C(f^E(.))$ transforms an input into class-wise probabilities through a series of convolutions, non-linear activations and a softmax function. $f^C(.)$ learns $p(\hat{y}_i = c \mid x_i), \forall c \in C$ vector such that the input image has the highest compatibility with its ground-truth $p(y_i = c \mid x_i), \forall c \in C$ vector. $p(y_i = c \mid x_i)$ is 1 for only one class and 0 for others. During training, $f^C(.)$ aims at jointly maximizing $p(\hat{y}_i = c \mid x_i), \forall x_i \in \mathbf{X}$ so that all inputs are correctly classified. The identical form for this aim is to minimize CCE, given by,

$$\begin{aligned} \arg\max_{\theta^E, \theta^C} \prod_{i=1}^{M} p(\hat{y}_i = c \mid x_i) &= \arg\min_{\theta^E, \theta^C} -\frac{1}{M} \sum_{i=1}^{M} \log p(\hat{y}_i = c|x_i) \\ = \arg\min_{\theta^E, \theta^C} -\frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{C} p(y_i = c|x_i). \log p(\hat{y}_i = c|x_i) &= \arg\min_{\theta^E, \theta^C} \mathcal{L}_{cce}(\mathbf{Y}, \hat{\mathbf{Y}} \mid \mathbf{X}) \end{aligned} \tag{7.1}$$

where, $\hat{\mathbf{Y}}$ denotes the predicted label vector for $\mathbf{X}$ and $\mathcal{L}_{cce}$ denotes the CCE loss. Optimizing CCE aims at tuning the network parameters to maximize the mutual information between the ground-truth class probability distribution and the predicted class probability distribution for all samples.

7.3.3 Deep metric learning

DML utilizes the local spatial distribution of samples in the non-linearly projected embedding space to bring similar samples (samples from the same class) closer, while pushing dissimilar samples (samples from different classes) apart. Similarity between a pair of samples (x_i, x_j) is measured by the euclidean distance, denoted by $d_{ij} := \|f^E(x_i) - f^E(x_j)\|^2$. d_{ij} is bounded by $[0, 2]$ as the embeddings lie on an unit sphere. In this work, we use triplets to capture the spatial distribution of samples. A triplet consists of an anchor, a positive and a negative identity, where the anchor and the positive identity are similar, and the anchor and the negative identity are dissimilar. For simplicity, we denote an anchor-positive identity pair as **positive-pair**, and an anchor-negative identity pair as a **negative-pair**. Following [Schroff et al., 2015], triplets are formed in an online manner using the samples from a mini-batch. Let $\mathbf{B} = \{B_c\}_{c=1}^{C}$ be a mini-batch, and B_c be the number of samples per class in $\mathbf{B}$. Let $\mathbf{P}_i$ and $\mathbf{N}_i$ denote the set of positive identities and negative identities respectively for an anchor $x_i \in \mathbf{B}$. A triplet $t_l : (t_l^a, t_l^p, t_l^n) = (x_i, x_j, x_k)$ is formed such that $i \neq j \neq k$, $y_i = y_j$, $y_i \neq y_k$, and $(x_i, x_j, x_k) \in \mathbf{B}$. t_l^a, t_l^p and t_l^n are anchor, positive and negative identities respectively. Let $\mathcal{T} = \{(t_l^a, t_l^p, t_l^n)\}_{l=1}^{T}$ be the set of triplets created from $\mathbf{B}$ with cardinality T. Thus, DML aims at,

$$d(f^E(t_l^a), f^E(t_l^p)) + m < d(f^E(t_l^a), f^E(t_l^n)), \forall l \in T \tag{7.2}$$

where, m is a margin factor that controls the distance between a positive-pair and a negative-pair. Several structured loss functions have been proposed by [Schroff et al., 2015; Sohn, 2016; Oh Song et al., 2016; Wu et al., 2017; Wang et al., 2019b] to optimize the DML objective. For instance, [Schroff et al., 2015] propose a triplet loss to optimize Equation 7.2, given by,

$$\mathcal{L}_{triplet} = \sum_{l \in T} [d(f^E(t_l^a), f^E(t_l^p)) - d(f^E(t_l^a), f^E(t_l^n)) + m]_+ \tag{7.3}$$

Similarly, [Sohn, 2016] propose a multi-class N-pair loss that employs one positive identity t_l^p, and N-1 negative identities $\{(t_{l1}^n, t_{l2}^n, ..., t_{lN-1}^n)\}$ per anchor to facilitate interaction with multiple negative classes in each update. N-pair loss is given as,

$$\mathcal{L}_{npair} = \sum_{l \in T} \log(1 + \sum_{k=1}^{N-1} \exp(f^E(t_l^a)^T . f^E(t_{lk}^n) - f^E(t_l^a)^T . f^E(t_l^p))) \tag{7.4}$$

Furthermore, the number of all possible triplets in **B** has a cubic complexity. It introduces a large number of uninformative triplets, i. e., triplets that do not violate equation 7.2. Also, it is time exhaustive and computationally infeasible to deal with all triplets. Therefore, several mining strategies are proposed by [Schroff et al., 2015; Hermans et al., 2017; Wu et al., 2017; Wang et al., 2019b] to sample informative triplets.

[Tschannen et al., 2019] shows that optimizing DML objective is equivalent to maximizing the mutual information between an anchor, and its corresponding distribution of positive and negative identities in the embedding space. Unlike CCE, DML leverages the local spatial context of an anchor to define its embedding. Therefore, jointly optimizing CCE and DML for an anchor can maximize the mutual information between the anchor and its context, and the ground-truth class probability distribution and the predicted class probability distribution for the anchor. The context information from DML can support CCE, thereby improving the classification performance.

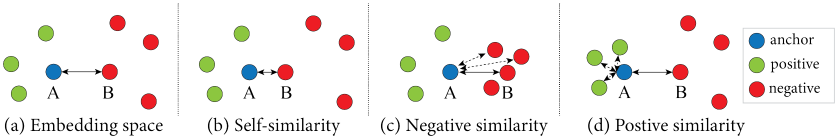
7.4 Methodology

In this section, we first propose the CoReL framework that learns class-label information and local spatial distribution information of samples. We describe the strengths of CoReL, and compare with standalone CCE and DML classifiers under limited training data. Then, we propose a neighborhood-aware multiple similarity sampling strategy to mine informative pair of samples, and a soft-multi-pair objective to boost the DML performance. Further, the proposed DML methodology is integrated with CoReL to enhance its performance. Figure 7.1 presents the overview of the CoReL framework and its components.

7.4.1 Co-representation learning framework for classification

CCE and DML optimize different objectives for classification. Optimizing CCE minimizes the gap between the predicted class probabilities and the ground-truth class distribution for a sample. Whereas, DML utilizes local spatial distribution of samples to learn robust embeddings for class delineation. [Horiguchi et al., 2020] thoroughly compare CCE and DML objectives to demonstrate their individual strengths and weaknesses. The objective of the CoReL framework aims at exploiting their complementarity:

$$\mathcal{L}_{\text{CoReL}}(\mathcal{T}) = \mathcal{L}_{cce}(\mathcal{T}_a) + \alpha\, \mathcal{L}_{dml}(\mathcal{T}_{apn}) + \lambda\, ||W||^2 \tag{7.5}$$



(a) Embedding space (b) Self-similarity (c) Negative similarity (d) Postive similarity

Figure 7.2: Illustration of multiple similarities for a negative pair considering the spatial distribution of samples in the embedding spaces. With respect to a given pair AB in the base embedding space (a), increased self-similarity of AB in (b), reduced negative relative similarity of AB in (c), and reduced positive relative similarity of AB in (d), are demonstrated.

where, $\mathcal{L}_{\text{CoReL}}$ is determined by $\mathcal{L}_{cce}$, CCE loss for the anchor identities in $\mathcal{T}$, $\mathcal{L}_{dml}$, DML loss for the triplets in $\mathcal{T}$, and L2 regularizer operating on network parameters $W = \{\theta^E, \theta^C\}$. $\alpha > 0$ is the loss trade-off parameter that adjusts the contribution by the losses. λ is the regularization coefficient. The CoReL framework possesses several advantages over standalone CCE and DML based classifications.

- CoReL can improve the classification by using the complementary information from CCE and DML objectives. The framework learns discriminative features, preserves intra-class variance and increases inter-class separability, without sacrificing the classification accuracy.
- In CoReL, the CCE objective can facilitate the mining of informative samples. Also, the faster convergence of CCE can benefit the DML convergence.
- The joint optimization prevents the framework from focusing too much on the class label information. The DML objective can be viewed as an indirect regularization of the framework. DML regularizes the feature representation using local distribution information, consequently preventing the network from overfitting.
- For limited training data, class labels provide limited information. However, the similarity constraints by DML can provide additional information about the spatial distribution, leading to a better classification. Also, the regularization provided by DML prevents the CCE objective to overfit to small training data. Both the measures boost the efficacy of CoReL using limited training data.

7.4.2 Improving deep metric learning

In addition to CoReL, we propose a methodology to improve DML classification performance, thus boosting the joint learning. Mining informative positive and negative pairs, and optimizing appropriate distance metric impact the quality of the learned embeddings, the convergence speed and the performance of DML. We hypothesize that the informativeness of a pair relies on the local spatial distribution of the constituting samples, the similarity between the constituting samples, and the relative similarity of the pair to its equivalent pairs in the embedding space. To this end, we introduce three types of similarity measures and a neighborhood-aware weighting scheme to define the informativeness of a pair. Finally, we propose a neighborhood-aware multiple similarity sampling (NAMSS) strategy for mining informative positive and negative pairs. Similarly, we propose a soft-multi-pair (SoMP) objective that considers to jointly optimize the interactions between multiple positive and negative pairs for an anchor to accelerate DML, and improve class separability.

7.4.2.1 Multiple similarities

We define three types of similarities for a pair, i. e., self-similarity, positive similarity and negative similarity, inspired by [Wang et al., 2019b]. For simplicity, we take a negative pair AB in Figure 7.2 across different embedding spaces to describe the similarities. A positive pair can be analyzed similarly.

• **Self-similarity:** Self-similarity is computed from a pair itself. The self-similarity between A and B is increased when the euclidean distance between A and B is decreased from Figure 7.2(a) to Figure 7.2(b). A high self-similarity between the samples of a negative pair indicates the difficulty of distinguishing the two paired samples from different classes. Such pairs are more informative to learn discriminative features. However, self-similarity does not fully describe the sample distribution in the embedding space. Also, it does not capture any correlation to other negative pairs, which can make a significant impact on similarity measurement.

• **Negative relative similarity:** Negative relative similarity is estimated by considering the correlation from neighboring negative pairs. It captures the significance of a pair in comparison to other equivalent pairs. Compared to Figure 7.2(a), the negative samples neighboring to B move closer to the anchor A in 7.2(c). This increases the self-similarities of the negative pairs neighboring to AB. Thus, the negative relative similarity of AB is reduced even when its self-similarity is unchanged.

• **Positive relative similarity:** A positive relative similarity captures the distribution of positive pairs with a same anchor to influence the informativeness of a negative pair. Compared to Figure 7.2(a), the positive samples are closer to anchor A in Figure 7.2(d). The AB pair is informative in 7.2(a) as it violates the triplet objective from Equation 7.2. But, in Figure 7.2(d) the AB pair satisfies Equation 7.2 and is less informative. Thus, the positive relative similarity of AB is reduced even when its self-similarity is unchanged.

• **Multiple similarity sampling (MSS):** We incorporate the three similarities to formulate a weighting scheme. To sample a negative pair, first, positive relative similarity is used to discard the less informative pairs. Second, self-similarity and negative relative similarity are used to weight the mined pairs. Third, informative pairs are uniformly sampled from the weight distribution.

Specifically, first, a negative pair is compared to the hardest positive pair. Formally, for an anchor x_i, a negative pair (x_i, x_k) is selected if d_{ik} satisfies,

$$d_{ik}^{-} < \max_{y_i = y_l} d_{il} + m, \text{ where } x_l \in \mathbf{P}_i \tag{7.6}$$

The index set of selected negative pairs are denoted as $\mathcal{N}_i \in \mathbf{N}_i$. Second, the weight $[w_{ik}^{-}]_{MS}$ for a mined negative pair $(x_i, x_k) \in \mathcal{N}_i$ is computed as,

$$[w_{ik}^{-}]_{MS} = \frac{\exp(\beta\, d_{ik})}{\sum\limits_{l \in \mathcal{N}_i} \exp(\beta\, d_{il})} \tag{7.7}$$

where, β is a fixed hyperparameter to scale the importance of hard-negatives. The numerator includes the self-similarity d_{ik}, and the denominator includes equivalent negative pairs to compute relative similarity. Finally, a negative pair is uniformly sampled

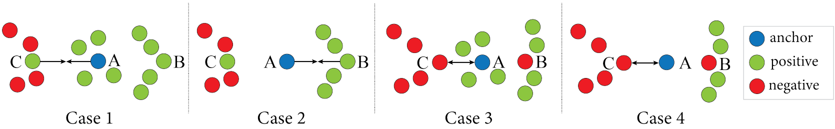


Figure 7.3: Illustrating the contribution of neighborhood awareness for sampling informative positive and negative pairs under neighborhood-aware multiple similarity sampling blanket. Samples B and C are equidistant from anchor A in the presented scenarios, thus AB and AC pairs are equally weighted according to multiple similarity measures. Evaluated scenarios are given as, (Case 1) positive identity mining for an easy anchor, (Case 2) positive identity mining for a difficult anchor, (Case 3) negative identity mining for an easy anchor, and (Case 4) negative identity mining for a difficult anchor.

from the weight distributions. The three steps are iteratively used to sample negative pairs for each anchor.

Similarly, a positive pair is mined using negative relative similarity, weighted using self-similarity and positive relative similarity, and sampled uniformly from the weight distribution. Formally, for an anchor x_i, a positive pair (x_i, x_j) is mined if,

$$d^+_{ij} > \min_{y_i \neq y_l} d_{il} - m, \text{ where } x_l \in \mathbf{N}_i \tag{7.8}$$

The index set of mined positive pairs are denoted as $\mathcal{P}_i \in \mathbf{P}_i$. Weight $[w^+_{ij}]_{MS}$ for a positive pair $(x_i, x_j) \in \mathcal{P}_i$ is computed as,

$$[w^+_{ij}]_{MS} = \frac{\exp\left(-\gamma\, d_{ij}\right)}{\sum\limits_{l \in \mathcal{P}_i} \exp\left(-\gamma\, d_{il}\right)} \tag{7.9}$$

where, γ is a fixed hyperparameter to scale the importance of hard-positives.

7.4.2.2 Neighborhood awareness

Investigating the neighborhood distributions of samples in the embedding space characterize the spatial pattern. The relative spatial position of a sample with respect to similar samples and dissimilar samples in the embedding space indicate the degree of disorganization. A disorganized sample, i. e., a sample surrounded by dissimilar samples or located away from similar samples, is more informative as it contributes the maximum to the DML objective. The neighborhood analysis identifies such samples, thus is vital for DML performance.

To characterize the notion of neighborhood awareness, we perform class-wise neighborhood analysis and assign a neighborhood-aware weight to each sample. Class-wise weighting of samples depict the relative degree of organization of similar samples. For a sample, we compute its average distance to K-nearest neighbors from the same class in the embedding space. Next, a class-wise weight normalization is used to compute the neighborhood-aware weights. A small and a large weight indicate relative organization and disorganization of the samples in the embedding space, respectively. A threshold value δ on the degree of organization delineates the samples into two categories, i. e., easy-neighborhood-samples and difficult-neighborhood-samples. Formally, the neighborhood-aware weight for x_i is defined as,

$$
\begin{gathered}
v_{i1} = \underset{j \in B_c, y_i = y_j = c}{\arg\min} \; d_{ij}; \; v_{ik} = \underset{j \in B_c, y_i = y_j = c, v_{i1}, \dots v_{ik-1}}{\arg\min} \; d_{ij} \\
w_i = \frac{1}{K} \sum_{k \in K} d_{i v_{ik}}, \; [w_i]_{NA} = \frac{w_i}{\sum\limits_{j \in B_c} w_j} \\
x_i = \begin{cases} \text{easy-neighborhood-sample; if } [w_i]_{NA} \leq \delta \\ \text{difficult-neighborhood-sample; otherwise} \end{cases}
\end{gathered}
\tag{7.10}
$$

where, $\{v_{i1}, ..., v_{ik}\}$ denote the index of K-nearest neighbors of x_i, and $[w_i]_{NA}$ denotes the neighborhood-aware weight for x_i. We follow the aforementioned scheme to categorize the anchor, positive and negative identities into easy or difficult categories.

7.4.2.3 Neighborhood-aware multiple similarity sampling

The neighborhood-awareness of the samples and similarity measures for a pair capture the structural information of the spatial distribution of samples in the embedding space to quantify the informativeness of the pair. In the context of mining informative triplets for DML, we define the sampling strategies for the following four scenarios,

- **Case-1: positive mining for easy-anchor:** Figure 7.3(a) presents an easy-anchor A surrounded by similar samples. Given B and C are equidistant from A, MSS assigns equal and the highest weights among all the positive identities. Easy-positive B is already organized appropriately in the respective class-cluster. Hence, pulling the difficult-positive C towards the respective class-cluster is more informative. Thus, difficult-positives are weighted more compared to easy-positives when mining positive identity for an easy-anchor.

- **Case-2: positive mining for difficult-anchor:** Figure 7.3(b) shows a difficult-anchor A distant from the similar samples. MSS assigns the highest and equal weights to AB and AC pairs, given B and C are equidistant from A. Disorganized A should be pulled towards the stable positive cluster instead of getting pulled towards the disorganized difficult-positive C. Thus, easy-positives are weighted more than difficult-positives when the positive pair involves a difficult-anchor.

- **Case-3: negative mining for easy-anchor:** Easy-anchor A in Figure 7.3(c) should push away from the cluster of easy-negatives for better class separation. Thus, given equidistant samples B and C from A, easy-negative pair AC is weighted higher.

- **Case-4: negative mining for difficult-anchor:** The difficult-anchor A in Figure 7.3(d) should push away from the cluster of easy-negatives for better class separability. Pushing A from a disorganized difficult-negative B may not lead the anchor towards a stable positive cluster. Thus, easy-negatives are assigned higher weights compared to difficult-negatives when the negative pair involves a difficult-anchor.

Considering the aforementioned scenarios, the weighting of a positive pair (x_i, x_j) and a negative pair (x_i, x_k) based on neighborhood awareness can be given as,

$$\begin{aligned} [w_{ij}^{+}]_{NA} &= \begin{cases} \frac{\exp([w_j]_{NA})}{\sum\limits_{l \in \mathbf{P}_i} \exp([w_l]_{NA})}; \text{ if } [w_i]_{NA} \leq \delta \\ \frac{\exp(-[w_j]_{NA})}{\sum\limits_{l \in \mathbf{P}_i} \exp(-[w_l]_{NA})}; \text{ otherwise} \end{cases} \\ [w_{ik}^{-}]_{NA} &= \frac{\exp(-[w_k]_{NA})}{\sum\limits_{l \in \mathbf{N}_i} \exp(-[w_l]_{NA})} \end{aligned} \tag{7.11}$$

For a given anchor, we combine the neighborhood awareness and the multiple similarities to define the NAMSS pair weighting scheme. Informative pairs are sampled uniformly from the normalized weight distribution given as follows,

$$\begin{aligned} w_{ij}^{+} &= [w_{ij}^{+}]_{NA} . [w_{ij}^{+}]_{MS}, \ [w_{ij}^{+}]_{NAMSS} = \frac{w_{ij}^{+}}{\sum\limits_{l \in \mathbf{P}_i} w_{il}^{+}} \\ w_{ik}^{-} &= [w_{ik}^{-}]_{NA} . [w_{ik}^{-}]_{MS}, \ [w_{ik}^{-}]_{NAMSS} = \frac{w_{ik}^{-}}{\sum\limits_{l \in \mathbf{N}_i} w_{il}^{-}} \end{aligned} \tag{7.12}$$

7.4.2.4 Soft-multi-pair loss

DP images contain high intra-class variability and high inter-class ambiguity. Similarity analysis via DML on such data by considering only one positive and only one negative identity per anchor exhibits slow convergence in learning the data variability. Further, considering the difference in the convergence speed between CCE and DML, CoReL may be unable to utilize the information from DML. Thus, We propose a triplet based soft-multi-pair (SoMP) loss that jointly optimizes the interactions between multiple positive pairs and multiple negative pairs per anchor to accelerate DML. Unlike multi-class N-pair loss ([Sohn, 2016]), SoMP considers multiple positives and multiple negatives per anchor to additionally constrain the embedding space.

Formally, let $\mathcal{T}$ be the set of triplets with cardinality T. The l-th triplet in $\mathcal{T}$ is presented by, $(t_l^a, t_{l1}^p, t_{l2}^p, ..., t_{lP}^p, t_{l1}^n, t_{l2}^n, ..., t_{lN}^n)$. t_l^a denotes the anchor identity, $(t_{l1}^p, t_{l2}^p, ..., t_{lP}^p)$ denotes the set of P unique positive identities, and $(t_{l1}^n, t_{l2}^n, ..., t_{lN}^n)$ denotes the set of N unique negative identities. The SoMP objective over $\mathcal{T}$ is defined as,

$$\begin{aligned} \mathcal{L}_{SoMP} = \sum_{l=1}^{T} \log(1 + \sum_{j=1}^{P} \sum_{k=1}^{N} \exp[(d(f^E(t_l^a), f^E(t_{lj}^p)) \\ - d(f^E(t_l^a), f^E(t_{lk}^n)) + m)]_+) \end{aligned} \tag{7.13}$$

where, m denotes the acceptable margin between a positive pair and a negative pair. The margin m enforces a distance between class clusters, and relaxes the objective by avoiding to correct "already correct" triplets. This ensures the optimization to focus on informative triplets that violate in Equation 7.2. For a triplet, SoMP jointly optimizes the interactions between multiple positive pairs and multiple negative pairs per anchor, thus eliminating the need of optimizing multiple triplets per anchor.

7.4.3 Co-representation learning training strategy

We perform online triplet mining from a mini-batch during training. Within a mini-batch, each sample is considered once as an anchor identity. A set of P positive identities and a set of N negative identities are sampled using NAMSS strategy to form triplets for each anchor. P and N are considered as design choices based on the size of the training data and the computational resources. SoMP loss is discounted by loss trade-off parameter α, and is jointly optimized with CCE loss as per Equation 7.5. CCE loss is optimized only for the anchor identities of the triplets, and the SoMP loss is optimized for the set of constructed triplets. Since each sample is once employed as an anchor, optimizing CCE for positive or negative identities is redundant. Optimizing CCE for the anchors ensures that complete class-label information is utilized during joint learning.

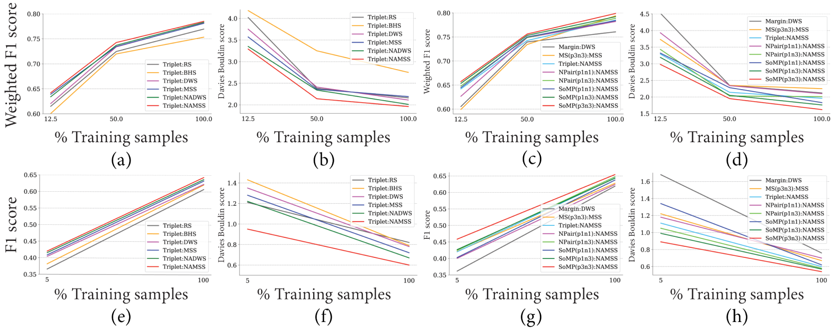
7.5 Results

In this section, we evaluate the CoReL framework on three classification tasks in DP: nuclei classification, mitosis detection and tissue type classification. We consider CRCHistoPhenotypes ([Sirinukunwattana et al., 2016]), CoNSeP ([Graham et al., 2019a]), ICPR12 ([Roux et al., 2013]), AMIDA13 ([Veta et al., 2015]) and Kather Multiclass ([Kather et al., 2019]) datasets in this regard. First, we perform ablation studies on CRCHistoPhenotypes and AMIDA13 to demonstrate the performance of the proposed DML methodology. Due to computationally prohibitive runtimes, not all datasets could be included in an extensive ablation study. Considering CRCHistoPhenotypes and ConSeP both having the same multi-class nuclei classification objective and similar dataset statistics, the well-established former dataset is selected for ablation experiments. Similarly, between AMIDA13 and ICPR12 having the same mitosis detection objective, the comparatively larger AMIDA13 dataset is used to obtain more reliable ablation statistics. Kather Multiclass presents an easier task and a relatively large dataset, thus, making it less suitable for extensive experiments. Second, we evaluate the performance of CoReL, incorporating the DML methodology, on all datasets with different percentages of training data.

7.5.1 Implementation

The CoReL framework is experimented separately for individual dataset, for different sampling strategies, different DML objectives and incremental subsets of training data. For each dataset, hyperparameters are tuned by training CoReL on complete training data, which are then applied to training with other subsets of the training data. In this section, we specify the implementation details of CoReL that are common to all datasets. Dataset-specific implementations and results are provided in the respective subsections.

First, we select the base deep network for the embedding module. Three networks, ResNet ([He et al., 2016]), Wide Residual Network (WRN) ([Zagoruyko et al., 2016]) and DenseNet ([Huang et al., 2017]), are trained to optimize CCE on the complete training data. The network with the highest classification performance on the validation set is selected as the base network. The output from the penultimate layer of the base network passes through D 1×1 convolutions, a global average pooling and a $l2$-normalization to produce the D-dimensional embedding. A classification module, consisting of a fully-connected layer and a softmax layer, processes the embedding to produce softmax output. DML and CCE objectives are computed using the D-dimensional embedding and the softmax output, respectively, and combined as per Equation 7.5 to define the CoReL objective. For



(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 7.4: Classification scores (higher is better) and DB scores (lower is better) on CRCHistoPhenoTypes (subfigures a to d) and AMIDA13 (subfigures e to h) datasets: (a) and (b) show the weighted F1 scores and DB scores for various triplet sampling strategies, (c) and (d) show the weighted F1 scores and DB scores for various DML losses. Similarly, (e) and (f) show the F1 scores and DB scores for various triplet sampling strategies, (g) and (h) show the F1 scores and DB scores for various DML losses.

the baselines, standalone classifiers are trained with individual classification strategies. For instance, a DML classifier optimizes DML objective and fits a K-nearest neighbor classifier to obtain class predictions.

Upon evaluation, we select a basic-wide WRN for AMIDA13 and ICPR12 datasets, and a DenseNet for CRCHistoPhenotypes, CoNSeP and Kather Multiclass datasets. The WRN has 16 base filters, a widening factor of 2, a depth of 40 layers and no average pooling. The DenseNet has 3 dense blocks with 12 layers per dense block, 48 base filters, 0.5 reduction rate and a depth of 40 layers. We use Adam optimizer ([Kingma et al., 2015]) with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1e^{-8}$. The network weights are initialized with He normal ([He et al., 2015]). The hyperparameters β, γ, δ and λ are set to 2, 50, 0.5 and 10^{-4}, respectively. Experimented hyperparameters and their values are given as, embedding size ($D \in [32, 64, 128, 256, 512]$), base batch size ($b \in [2, 4, 8, 16, 32]$), learning rate ($lr \in [0.001, 0.01, 0.1]$), loss trade-off parameter ($\alpha \in [0.001, 0.01, 0.1, 0.5, 1]$) and margin ($m \in [0.5, 1.0, 1.5]$). The learning rate is reduced by 0.5 when the evaluation metric on the validation data does not improve for 5 consecutive epochs. Early stopping with a patience of 20 epochs is used to prevent overfitting. Each training mini-batch is constructed by mimicking the per-class distribution of samples in the training data. The experiments are performed using an NVIDIA Tesla P100 GPU with POWER8 processors. All networks are implemented in Keras with TensorFlow 1.8 backend.

7.5.2 Ablation framework

We perform ablation studies to evaluate the impact of four factors on DML classification, i. e., (1) triplet sampling strategy, (2) DML loss function, (3) embedding size, and (4) batch size. Each factor is analyzed individually, while fixing the others.

We compare six triplet sampling strategies. For a triplet (x^a, x^p, x^n), and corresponding positive set **P** and negative set **N** in a mini-batch, the sampling strategies are given as,

Random sampling (RS): x^p and x^n are sampled independently of x^a by randomly choosing from **P** and **N**, respectively.

Batch-hard sampling (BHS) [Hermans et al., 2017]: The hardest positive $x^p \in \mathbf{P}$ (the farthest from x^a) and the hardest negative $x^n \in \mathbf{N}$ (the closest to x^a) are sampled for x^a.

Distance weighted sampling (DWS) [Wu et al., 2017]: All negative pairs for x^a are weighted in inverse proportion, and all positive pairs for x^a are weighted in direct proportion to the distribution of pair-wise distances on an unit sphere. x^p and x^n are uniformly sampled from the weighted distribution.

Multiple similarity sampling (MSS): x^p and x^n are uniformly sampled from the weighted probability distributions specified in Equation 7.9 and Equation 7.7, respectively.

Neighborhood-aware distance weighted sampling (NADWS): All positive and negative pairs for x^a are weighted by combining the neighborhood awareness distribution, Equation 7.10, and the pair-wise distance weighted distribution, [Wu et al., 2017]. x^p and x^n are sampled uniformly from the joint probability distribution.

Neighborhood-aware multiple similarity sampling (NAMSS): All positive and negative pairs for x^a are weighted by combining the neighborhood awareness and multiple similarity measures as per Equation 7.12. x^p and x^n are uniformly sampled from these joint probability distributions.

Further, we compare five DML objectives, i. e., triplet loss ([Schroff et al., 2015]), multi-class n-pair loss ([Sohn, 2016]), margin loss ([Wu et al., 2017]), multiple similarity loss ([Wang et al., 2019b]) and the proposed SoMP loss. We consider three settings to evaluate the impact of optimizing multiple pair interactions in DML, i. e., $(p1, n1)$, $(p1, n3)$ and $(p3, n3)$, where the numbers next to p and n denote the number of positive and negative identities per anchor in a triplet. We perform ablation studies for five embedding sizes ($D \in [32, 64, 128, 256, 512]$) and four base batch sizes ($b \in [4, 8, 16, 32]$).

Weighted F1 score and F1 score on test set measure the classification performance on CRCHistoPhenotypes and AMIDA13, respectively. Weights for weighted F1 score on CRCHistoPhenotypes are given in Section 7.5.3. We analyze the quality of the embeddings using Davies-Bouldin (DB) score [Davies et al., 1979] that measures the appropriateness of data partitions in the embedding space. DB score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Clusters that are farther apart and less dispersed results in a better (i. e., lower) score.

7.5.3 CRCHistoPhenotypes

CRCHistoPhenotypes ([Sirinukunwattana et al., 2016]) dataset consists of 100 H&E stained histology images of size 500×500 pixels from 10 whole-slides across 9 colorectal adenocarcinoma patients. The images are captured at 20× magnification by Omnyx VL120 scanner from normal and malignant regions. These images include a total of 22444 nuclei that are classified into four categories: 7722 epithelial, 5712 fibroblast, 6971 inflammatory, and 2039 miscellaneous nuclei. Inflammatory category includes lymphocyte plasma, neutrophil and eosinophil, and miscellaneous category includes all the remaining types of nuclei, such as adipocyte, endothelium, mitotic figure, and nucleus of necrotic. We employ two-fold cross-validation evaluation, same as [Sirinukunwattana et al., 2016],

Table 7.1: Effect of embedding size on classification performance using various subsets of training data on CRCHistoPhenotypes and AMIDA13.

Embedding size	CRCHistoPhenotypes		AMIDA13	
%Training Data	12.5%	100%	5%	100%
32	0.625	0.779	0.364	0.624
64	0.645	0.784	0.415	0.633
128	0.649	0.791	0.426	0.646
256	**0.657**	**0.799**	**0.458**	**0.654**
512	0.648	0.794	0.451	0.653

Table 7.2: Effect of base batch size on classification performance using various subsets of training data on CRCHistoPhenotypes and AMIDA13.

Batch size	CRCHistoPhenotypes		AMIDA13	
%Training Data	12.5%	100%	5%	100%
4	0.654	0.793	0.448	0.633
8	**0.657**	**0.799**	**0.458**	**0.654**
16	0.645	0.794	0.449	0.632
32	0.641	0.786	0.445	0.621

and split the dataset into 40, 10 and 50 images for train, validation and test respectively. Further, to experiment with limited training data constraints, we prepare six incremental subsets of the training data. We split the training set into sets of [1, 5, 10, 20, 30, 40] images corresponding to [2.5%, 12.5%, 25%, 50%, 75%, 100%] of training data. Patches of size 36×36 pixels are extracted from the images centered around the annotated nucleus centroids. Patches are randomly augmented using rotation (90^0, 180^0, 270^0), translation (±4 pixels), mirroring along horizontal and vertical axes and color augmentation in HSV space. In the HSV space, the hue (H), saturation (S), and value (V) variables are separately multiplied by random numbers $r_H \in [0.95, 1.05]$, and r_S, $r_V \in [0.9, 1.1]$. For a test nuclei, class-wise mean of the predicted probabilities for 5 augmentations of the nuclei are used to assign the final class label. We compute the F1 score for each nuclei class and compute their weighted average by the number of nucleus samples to summarize the overall classification performance.

7.5.3.1 Ablation study

- **Impact of embedding size:** We study DML performance for varying embedding sizes on two training subsets, i. e., 12.5% and 100%. DML classifiers are trained using NAMSS strategy, SoMP$(p3, n3)$ loss with $m = 0.5$, $b = 8$ and $lr = 0.01$. Table 7.1 indicates consistent performance gain for both the subsets with increasing the embedding size up till 256.

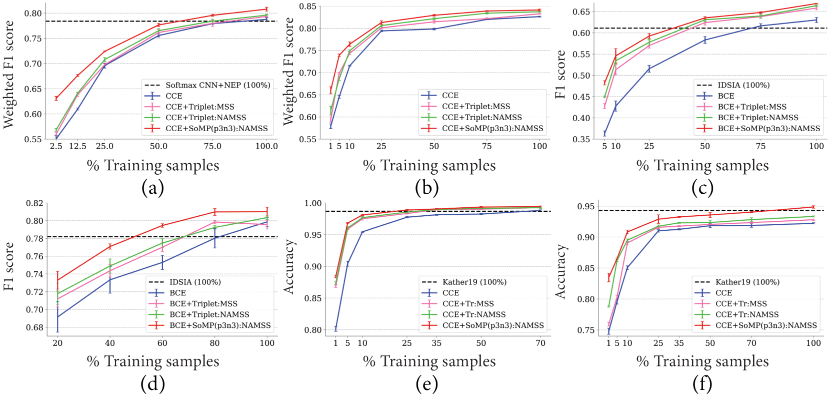


Figure 7.5: Mean and standard deviation of classification performances for various learning objectives, including standalone categorical cross entropy, and the proposed CoReL framework with various DML losses and sampling strategies. All frameworks using incremental subsets of training data are trained three times with different weight initializations on (a) CRCHistophenotypes, (b) CoNSeP, (c) AMIDA13, (d) ICPR12, (e) Kather Multiclass Internal, and (f) Kather Multiclass External datasets. Also, a reference to the state-of-the-art classification approaches for each dataset trained using complete training data is given by a dashed line.

- **Impact of base batch size:** Similar studies are performed with $D = 256$ and varying base batch sizes. Effective batch size of a mini-batch corresponding to $b = [4, 8, 16, 32]$ are $[25, 52, 104, 208]$. Table 7.2 shows increase in classification performance for increasing b till 8, and decreases with further increment in b. CRCHistoPhenotypes dataset is a fine-grained dataset with small inter-class variations, making it easy to collect hard negative pairs with small batch size.

- **Impact of triplet sampling techniques:** Experiments are conducted by optimizing triplet loss with $m = 0.5$, $D = 256$, $b = 8$ and $lr = 0.01$ on three subsets of training data, i. e., 12.5%, 50%, and 100%. Figure 7.4(a) and Figure 7.4(b) present the weighted F1 scores and DB scores for experiments with different sampling strategies. Selecting only hard-negatives in BHS provides too hard-negatives causing the gradient to have high variance, thus reducing the performance. All other sampling strategies perform better than RS, signifying the importance of sampling in DML. MSS performs better than DWS indicating the positive impact of relative similarity measures. Both NADWS and NAMSS further improve the performances of DWS and MSS, depicting the importance of neighborhood awareness in pair mining. The DB score evaluating the cluster quality in the embedding space confirms the above observations. The proposed NAMSS strategy provides the best classification performance among the sampling strategies.

- **Impact of DML objective:** We analyze the impact of DML objectives for NAMSS strategy. Figure 7.4(c) and Figure 7.4(d) present the weighted F1 scores and DB scores for the experiments. Unlike N-pair loss, triplet loss includes a margin to not penalize already correct triplets. The better performance of triplet loss than N-pair $(p1, n1)$ indicates the positive impact of margin relaxation. However, N-pair $(p1, n3)$ outperforms

Table 7.3: Results on CRCHistoPhenotypes dataset. Weighted F1 scores for various learning strategies, including categorical cross entropy alone, and CoReL with various deep metric losses and sampling strategies. The best weighted F1 scores are shown in bold.

%Training Data	2.5%	12.5%	25%	50%	75%	100%
Softmax CNN + NEP ([Sirinukunwattana et al., 2016])	-	-	-	-	-	0.784
Softmax CNN + SSPP ([Sirinukunwattana et al., 2016])	-	-	-	-	-	0.748
CCE	0.571	0.639	0.701	0.758	0.780	0.791
CCE + Triplet: MSS	0.565	0.642	0.703	0.764	0.785	0.798
CCE + Triplet: NAMSS	0.572	0.643	0.712	0.772	0.790	0.799
CCE + SoMP(p1,n1): NAMSS	0.591	0.656	0.712	0.764	0.789	0.802
CCE + SoMP(p1,n3): NAMSS	0.602	0.663	0.721	0.769	0.792	0.807
CCE + SoMP(p3,n3): NAMSS	**0.637**	**0.678**	**0.726**	**0.781**	**0.798**	**0.813**

Table 7.4: Results on AMIDA13 dataset. F1 scores for various learning strategies, including binary cross entropy alone, and CoReL with various deep metric losses and sampling strategies. The best F1 scores are shown in bold.

%Training Data	5%	10%	25%	50%	75%	100%
SegMitos ([Li et al., 2019a])	-	-	-	-	-	**0.673**
IDSIA ([Veta et al., 2015])	-	-	-	-	-	0.611
BCE	0.371	0.442	0.511	0.577	0.618	0.644
BCE + Triplet: MSS	0.435	0.530	0.578	0.631	0.642	0.664
BCE + Triplet: NAMSS	0.451	0.540	0.587	0.637	0.645	0.667
BCE + SoMP(p3,n3): NAMSS	**0.489**	**0.565**	**0.598**	**0.639**	**0.650**	0.671

triplet loss indicating the importance of jointly optimizing multiple negative interactions. SoMP performance increases with increasing the number of interactions as shown by SoMP$(p3, n3) >$ SoMP$(p1, n3) >$ SoMP$(p1, n1)$. SoMP $(p3, n3)$ outperforms N-pair $(p1, n3)$ signifying the effect of margin relaxation and the optimization of multiple positive interactions. Further, we compare SoMP$(p3, n3)$ with two recent DML frameworks proposed by [Wu et al., 2017] and [Wang et al., 2019b], to demonstrate the efficacy of our DML framework.

7.5.3.2 Classification with CoReL

The ablation studies signify the efficacy of NAMSS strategy and SoMP objective for improving DML performance. We incorporate the proposed DML methodology to CoReL to evaluate its significance in the joint learning framework. From Table 7.1 and Table 7.2, we set $D = 256$ and $b = 8$. Table 7.3 presents the best weighted F1 scores for the state-of-the-art approaches, the standalone CCE classifier, and the CoReL framework with various DML methodologies. We train each network three times with different weight initializations, and report the mean and standard deviations of weighted F1 scores in Figure 7.5. CoReL frameworks provide better results compared to standalone CCE classifier for all subsets of training data. CCE+Triplet:NAMSS performs

Table 7.5: Results on ICPR12 dataset. F1 scores for various learning strategies, including binary cross entropy alone, and CoReL with various deep metric losses and sampling strategies. The best F1 scores are shown in bold.

%Training Data	20%	40%	60%	80%	100%
DeepMitosis ([Li et al., 2018a])	-	-	-	-	**0.832**
SegMitos ([Li et al., 2019a])	-	-	-	-	0.771
IDSIA ([Roux et al., 2013])	-	-	-	-	0.782
BCE	0.712	0.724	0.766	0.789	0.806
BCE + Triplet: MSS	0.722	0.754	0.778	0.800	0.802
BCE + Triplet: NAMSS	0.736	0.759	0.782	0.795	0.806
BCE + SoMP(p3,n3): NAMSS	**0.747**	**0.775**	**0.798**	**0.814**	0.818

Table 7.6: Results on CoNSeP dataset. Weighted F1 scores for various learning strategies, including categorical cross entropy alone, and CoReL with various deep metric losses and sampling strategies. The best weighted F1 scores are shown in bold.

%Training Data	1%	5%	10%	25%	50%	75%	100%
CCE	0.605	0.654	0.740	0.799	0.809	0.824	0.831
CCE + Triplet: MSS	0.602	0.701	0.748	0.804	0.819	0.825	0.835
CCE + Triplet: NAMSS	0.627	0.691	0.752	0.811	0.824	0.836	0.839
CCE + SoMP(p3,n3): NAMSS	**0.672**	**0.742**	**0.772**	**0.816**	**0.832**	**0.841**	**0.844**

better than CCE+Triplet:MSS signifying the impact of neighborhood awareness based sampling. Further, CCE+SoMP(p1,n1):NAMSS outperforms CCE+Triplet:NAMSS indicating the contribution of SoMP loss. The performance of CoReL with SoMP objective increases with increasing the number of multiple positive and negative pair interactions, indicated by CCE+SoMP(p3,n3):NAMSS $>$ CCE+SoMP(p1,n1):NAMSS. CCE+SoMP(p3,n3):NAMSS provides significant gain compared to CCE classifier in the limited data regime. The gain gradually decreases with increasing the size of the training data. CCE+SoMP(p3,n3):NAMSS achieves the state-of-the-art results ([Sirinukunwattana et al., 2016]) by using 50% of the training data. On using complete training data, CCE+SoMP(p3,n3):NAMSS surpasses the state-of-the-art by 2.1%. We can not compare our results with [Shao et al., 2018; Zhou et al., 2018; Hamad et al., 2018; Li et al., 2019c] as they use different data splitting protocols, i. e., different number of cross-validation folds, and nuclei-level splitting instead of image-level splitting compared to the original work by [Sirinukunwattana et al., 2016].

7.5.4 AMIDA13

AMIDA13 ([Veta et al., 2015]) dataset consists of 12 subjects for training and 11 subjects for testing. The training set contains 311 high-power-fields (HPFs) and 550 annotated mitotic figures, and the test set contains 295 HPFs and 533 mitoses. The size of each HPF is 2000×2000 pixels, representing an area of 0.25mm^2. The HPFs are captured at

40× magnification with a spatial resolution of 0.25 μm/pixel by ScanScope XT whole-slide scanner. Centroid annotations for mitoses are provided by the consensus of two pathologists. We split the training set by subjects, and consider 9 subjects for training and 3 subjects [6, 9, 10] for validation. The training subjects include 230 HPFs and 452 mitoses, and the validation subjects include 81 HPFs and 98 mitoses. To add the limited data constraints, we split the 230 HPFs into incremental subsets of [12, 24, 61, 120, 176, 230] images corresponding to [5%, 10%, 25%, 50%, 75%, 100%] of the training data. These splits contain [28, 48, 107, 228, 340, 452] mitoses respectively.

We begin with normalizing the HPFs using color deconvolution [Stanisavljevic et al., 2018] to reduce the appearance variability. The nuclei in H&E stained HPFs possess high blue channel intensity. We convert the RGB images into blue-ratio images ([Chang et al., 2011]) and identify potential nuclei candidates by detecting high brightness objects. Morphological opening, Otsu thresholding and connected components analysis detect the bright objects. We consider connected components above an area of 100 pixels to define nuclei candidates, and extract patches of 72×72 pixels centered around their centroids. Nuclei detected within 20 pixels from the annotated mitoses are considered as mitoses. Extracted patches are randomly augmented using rotation (angles in multiples of 45^0), translation (±8 pixels) and flipping along horizontal and vertical axes.

We follow an incremental training strategy to address the high non-mitoses to mitoses ratio in the dataset. A preliminary network is trained using all mitoses and equal number of non-mitoses NM, a subset of complete non-mitotic set. The network identifies hard non-mitoses, i. e., non-mitoses predicted as mitoses with high probability, from the complementary set NM'. The hard non-mitoses are included to NM for subsequent training. NM is periodically extended by repeating the above strategy. The strategy prevents fitting suboptimal hypotheses by using evidence from the entire training set, and significantly limits the non-mitoses in the training phase. Additional to the implementation strategy in Section 7.5.1, Adam optimizer is used with learning rates set to 10^{-3}, 10^{-3} and 10^{-4}, respectively, at 0th, 20th and 40th epoch. Hard non-mitoses are included at 20th and 40th epochs. For a test HPF, we detect nuclei centroids and classify corresponding nuclei patches. Prediction probabilities are obtained for 5 augmentations of every nuclei. Class-wise mean of the probabilities is used to assign final class label. A predicted mitosis is considered a true-positive, if it lies within 30 pixels from the ground-truth mitotic location. Misdetected and undetected mitoses are considered as false-positives and false-negatives respectively. F1 score for mitotic class is used for evaluation. The highest F1 score on the validation data is used for model selection.

Ablation study:

- **Impact of embedding size:** DML classifiers are trained by optimizing SoMP$(p3, n3)$ loss with $m = 0.5$, NAMSS sampling strategy and $b = 8$ for varying embedding sizes on 5% and 100% of training data. F1 score increases with increasing embedding dimensions till 512 as shown in Table 7.1. The trend is consistent for both the training subsets.

- **Impact of base batch size:** Similar setting as 7.5.3.1A with $D = 256$ is evaluated with different base batch sizes. Table 7.2 indicates increase in F1 score for increasing base batch size from 4 to 8. Further increase in base batch size reduces F1 score.

- **Impact of triplet sampling strategies:** Sampling strategies are compared by training DML classifiers by optimizing triplet loss with $m = 0.5$, $D = 256$ and $b = 8$ on 5% and

Table 7.7: Results on Kather Multiclass datasets. Accuracy values for various learning strategies, including categorical cross entropy alone, and CoReL with various deep metric losses and sampling strategies. The best accuracies are shown in bold.

Methods	Kather Multiclass Internal							Kather Multiclass External							
%Training Data	1%	5%	10%	25%	35%	50%	70%	1%	5%	10%	25%	35%	50%	70%	100%
Kather19 ([Kather et al., 2019])	-	-	-	-	-	-	0.987	-	-	-	-	-	-	-	0.943
CCE	0.806	0.919	0.956	0.978	0.982	0.986	0.989	0.753	0.800	0.854	0.913	0.914	0.922	0.922	0.924
CCE + Triplet: MSS	0.871	0.959	0.978	0.984	0.990	0.991	0.993	0.763	0.808	0.894	0.918	0.920	0.924	0.927	0.930
CCE + Triplet: NAMSS	0.875	0.962	0.978	0.986	0.991	0.992	0.993	0.802	0.866	0.898	0.919	0.925	0.928	0.932	0.935
CCE + SoMP(p3,n3): NAMSS	**0.891**	**0.970**	**0.982**	**0.990**	**0.991**	**0.993**	**0.995**	**0.842**	**0.888**	**0.913**	**0.933**	**0.935**	**0.939**	**0.942**	**0.951**

100% of training data. Results in Figure 7.4(e) and Figure 7.4(f) present similar trend in classification performances like CRCHistoPhenotypes. The results indicate superior performance for NAMSS strategy on both training subsets.

- **Impact of DML objectives:** DML losses are analyzed for NAMSS strategy with $D = 256$ and $b = 8$. Figure 7.4(g) and Figure 7.4(h) demonstrate similar trend in classification performances like CRCHistoPhenotypes. SoMP($p3, n3$) performed the best among the DML objectives which includes the strategies proposed by [Wu et al., 2017] and [Wang et al., 2019b].

7.5.5 Classification with CoReL

We analyze the CoReL frameworks for jointly optimizing binary cross-entropy (BCE) and various DML strategies. Considering the binary classification task BCE is optimized instead of CCE. All experiments are conducted three times with different network weight initializations, $D = 256$ and $b = 8$ using different subsets of training data. The best F1 scores for all experiments and the state-of-the-art are presented in Table 7.4. The mean and standard deviation of F1 scores are presented in Figure 7.5(c). BCE+SoMP($p3, n3$):NAMSS performs the best across all training subsets. It outperforms the state-of-the-art classification approach ([Veta et al., 2015]) significantly and achieves comparable performance to the state-of-the-art segmentation approach ([Li et al., 2019a]). Results in Table 7.4 indicate the improvements provided by the NAMSS strategy and SoMP loss. For limited data regime, BCE+SoMP($p3, n3$):NAMSS significantly improves the detection performance compared to standalone BCE classifier. Also, the detection performance of the BCE classifier for using 100% training data is achieved by BCE+SoMP($p3, n3$):NAMSS for using only 50% of the training data. BCE+SoMP($p3, n3$):NAMSS sets the new state-of-the-art results for classification based mitotis detection.

7.5.6 ICPR12

The ICPR 2012 mitotis dataset ([Roux et al., 2013]) is extracted from a set of five breast cancer biopsy slides. 10 HPFs at 40× magnification are selected per slide by the pathologists. Among the 50 HPFs, 35 HPFs and 15 HPFs belong to train set and test set respectively. We use 25 images for training and 10 images for validation. The validation set contains 2 HPFs per slide. To experiment with limited data, we divide the training set of 25 images into five incremental subsets of [5, 10, 15, 20, 25] HPFs corresponding to [20%, 40%, 60%, 80%, 100%] of the training data. These splits contain [38, 63, 93, 137, 171] mitoses respectively. ICPR12 dataset contains pixel-wise annotation masks for mitoses. We consider the centroid of the mitoses masks as the ground-truth mitotic locations.

We follow the same pre-processing steps as AMIDA13 to detect nuclei candidates. We extract patches of size 96×96 around the centroids of the detected nuclei. Random augmentations are applied by rotating (angles in multiples of 45^0), translating (±8 pixels) and flipping around horizontal and vertical axes. We follow the same training and testing protocol as AMIDA13. Mitoses detected within 20 pixels from the ground-truth annotations are considered as true-positives. F1 score for the mitotic class is used to evaluate the algorithms.

Table 7.5 presents the F1 scores for several methodologies. Figure 7.5(d) presents the mean and standard deviation of F1 scores for all experiments trained three times with different weight initializations. All CoReL frameworks outperform the BCE classifier for all training subsets. BCE+SoMP($p3, n3$):NAMSS achieves the best F1 score. The BCE detector performance for using 100% training data is achieved by BCE+SoMP($p3, n3$):NAMSS for utilizing 50% of the training data. On using complete training data, BCE+SoMP($p3, n3$):NAMSS outperforms the state-of-the-art classification ([Roux et al., 2013]) and segmentation approach ([Li et al., 2019a]), and performs comparable to the mixed segmentation and classification approach ([Li et al., 2018a]).

7.5.7 CoNSeP

CoNSeP [Graham et al., 2019a]) dataset consists of 41 H&E stained images, each of size 1000×1000 pixels. Images are extracted at 40× magnification with an Omnyx VL120 scanner from 16 colorectal adenocarcinoma patients. The nuclei are categorized into four categories: epithelial, inflammatory, spindle-shaped and miscellaneous. Epithelial type consists of normal and tumor epithelial. Spindle-shaped type includes endothelial, muscle and fibroblast. Miscellaneous type contains necrotic, mitotic and cells that could not be categorized. Out of a total of 24332 nuclei annotations, 8751 are epithelial, 5579 are inflammatory, 9070 are spindle-shaped and 932 are miscellaneous. The train and test sets consists of 27 and 14 images respectively. We consider indices, [1, 6, 8, 10, 16, 25, 26], out of 27 images for validation. To experiment with limited data constraints, we divide the nuclei from 20 images into seven incremental training subsets, i.e, [1%, 5%, 10%, 25%, 50%, 75%, 100%]. We do not split at the image-level due to the uneven distribution of nuclei types across images. Patches of size 64×64 pixels are extracted around nuclei centroids. We follow the data augmentations and testing procedure specified for CRCHistoPhenotypes. Class-wise distributions of nuclei are used as weights to compute weighted F1 score.

Table 7.6 presents the weighted F1 scores for different methodologies. Figure 7.5(b) presents the mean and standard deviation of weighted F1 scores for all experiments trained three times with different weight initializations. All CoReL frameworks perform better than the CCE classifier for all subsets of training data, and CCE+SoMP($p3, n3$):NAMSS achieves the best score. It significantly outperforms the CCE classifier in the low data regime. The best weighted F1 score by the CCE classifier for using 100% training data is achieved by CCE+SoMP($p3, n3$):NAMSS for using only 40% of the training data. We can not compare our results to [Graham et al., 2019a], as they provide classification results for only 73.68% nuclei in the test data that are detected by the nuclei segmentation branch.

7.5.8 Kather Multiclass

Kather Multiclass ([Kather et al., 2019]) provides two datasets: Kather Multiclass-Internal (KMI) and Kather Multiclass-External (KME). KMI contains 100,000 tissue images of 224×224 pixels at 0.5 μm/pixel spatial resolution for 86 H&E colorectal cancer slides from

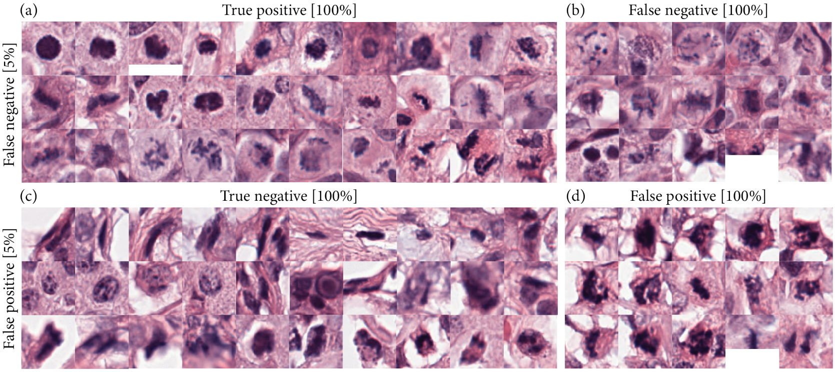


Figure 7.6: Qualitative assessment of the impact of the size of training data on CoReL classification performance. The evaluation is performed on AMIDA13 dataset by training the CoReL framework on 5% and 100% training data.

NCT biobank and UMM pathology archive. KME contains 7180 images from 25 H&E colorectal cancer slides from DACHS study in the NCT biobank. The images are stain normalized with [Macenko et al., 2009] algorithm. KMI and KME contain nine tissue types, i. e., adipose tissue, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal cancer epithelium. Similar to [Kather et al., 2019], we divide KMI into 70%, 15% and 15% for training, validation and testing. Entire KMI is split into eight incremental subsets, i. e., [1%, 5%, 10%, 25%, 35%, 50%, 70%, 100%]. The first seven sets are used for training, and the incremental 30% in the last subset is split to define the validation and test set. Similar to [Kather et al., 2019], KME is used as an external validation set for networks trained using KMI. Thus, all eight subsets are used for training while validating on KME. We employ DenseNet architecture as the embedding module instead of VGG19 architecture employed by [Kather et al., 2019] considering the similar classification performance of DenseNet with a significant reduction in network complexity in terms of trainable network parameters.

Table 7.7 presents test accuracies on KMI and validation accuracies on KME data for the state-of-the art approach, CCE baseline and CoReL frameworks with various DML strategies. Figure 7.5(e) and Figure 7.5(f) present the mean and standard deviation of accuracies for all experiments trained three times with different weight initializations. In both the datasets, CCE+SoMP($p3, n3$):NAMSS performs the best for all training subsets. In the limited data regime, it significantly outperforms the standalone CCE classifier. It achieves the state-of-the-art results on KMI and KME data for using 25% and 70% of training data respectively. CCE+SoMP($p3, n3$):NAMSS outperforms the state-of-the-art by using the complete training data. However, results on KME can further be improved by using a better embedding module.

7.5.9 Qualitative analysis

The size of the training data impacts the learning, hence the discriminability power of CoReL framework. To assess this impact, we present qualitative results for CCE+SoMP(p3, n3):NAMSS trained with 5% and 100% training data from AMIDA13 dataset, denoted as CoReL-5 and CoReL-100 respectively. Unlike CRCHistoPhenotypes, AMIDA13 is a binary classification task, thus it is selected for simplified qualitative assessment. Figure 7.6 presents qualitative results by illustrating false negative (FN) and false positive (FP) mitotic samples for CoReL-5 and CoReL-100 frameworks.

Figure 7.6(a,b) and Figure 7.6(c,d) present FN and FP instances for CoReL-5 respectively. CoReL-100 disagrees with CoReL-5 for the instances in Figure 7.6(a,c), and agrees with CoReL-5 for the instances in Figure 7.6(b,d). Specifically, Figure 7.6(a) includes samples from different mitotic phases indicating the learning inefficiency of CoReL-5 due to the low intra-class variability in the small mitotic training data. However, Figure 7.6(b) presents difficult mitotic instances for both CoReL-5 and CoReL-100, which primarily contains complex and ambiguous patterns, and noise such as blurring, and boundary mitotic figures. Figure 7.6(c) displays FPs for CoReL-5 which can be characterized as samples with non-mitotic figures such as, stromal nuclei, lymphocytes, cancerous epithelial nuclei, and samples with dense nuclei arrangements, ambiguous patterns, and noise. The small mitotic training data of CoReL-5 accounts for the misclassifications. Further, Figure 7.6(d) presents FPs common to both CoReL-5 and CoReL-100 which primarily include complex discernible patterns. The precision of CoReL-5 and CoReL-100 are 0.395 and 0.636 respectively, and the recall of CoReL-5 and CoReL-100 are 0.642 and 0.724 respectively. A comparatively higher gain in precision than recall and the qualitative results indicate that increasing training data significantly lowers the FP mitoses, and learns to identify the mitotic phases.

7.5.10 Runtime analysis

Training time statistics for CRCHistoPhenotypes and AMIDA13 datasets are provided in Table 7.8. We present the absolute training time per epoch for CCE and CoReL (CCE+SoMP($p3, n3$):NAMSS) framework, and the relative training time of CoReL with respect to CCE for different percentages of training data. Table 7.8 depicts that the CoReL framework is computationally 1.568 ± 0.03 and 1.113 ± 0.03 times more expensive than standalone CCE classifier for CRCHistoPhenotypes and AMIDA13, respectively. This is likely due to the DML component in CoReL that involves expensive triplet sampling. Further, the effective batch sizes for training CoReL frameworks on CRCHistoPhenotypes and AMIDA13 are 52 and 20, respectively. This leads to a more expensive DML component for CRCHistoPhenotypes, which results in higher relative training time. However, CoReL possesses the same inference time as the standalone CCE classifier as both the frameworks consist of an equal number of network parameters, and the DML component in CoReL is not computed during the inference phase.

7.6 Discussion

The classification performances for optimizing CCE and DML objectives are impacted differently by the training data size. To understand the contribution of the individual objectives to the CoReL framework, we experiment with loss-tradeoff parameter α for different training data sizes. We evaluate for the best performing CCE+SoMP($p3, n3$):NAMSS framework, and the SoMP loss is optimized with $m = [0.5, 1.0, 1.5]$ to eliminate the impact of margin. For every dataset, we rank the mean of the classification performances for varying α values on each subset of the training data. A higher rank indicates a higher classification performance. In Figure 7.7, for lower training data regions, higher α val-

Table 7.8: Training time analysis of CoReL framework and standalone CCE framework on CRCHistoPhenoTypes and AMIDA13 dataset for different percentages of training data.

	CRCHistoPhenotypes				
% Data/Time(in s)	12.5%	25%	50%	75%	100%
CCE	40	60	100	140	170
CoReL	60	95	160	220	270
Relative time	1.50	1.58	1.60	1.57	1.59

	AMIDA13					
% Data/Time(in s)	5%	10%	25%	50%	75%	100%
CCE	130	140	150	180	200	220
CoReL	140	150	170	200	230	250
Relative time	1.08	1.07	1.13	1.11	1.15	1.14

ues result in higher ranks, thus emphasizing on DML. In these regions, CCE objective does not acquire sufficient class label information. Whereas, DML exploits the available data by jointly optimizing multiple pair-wise interactions per anchor to capture spatial distribution information of the samples in the embedding space. Further, DML strongly regularizes the framework, and forces it focus less on the class labels. to prevent overfitting. Thus, DML drives the performance of the CoReL framework under limited training data. In Figure 7.7, for higher training data regions, lower α values result in higher ranks, thus emphasizing on CCE. In these regions, CCE objective acquires vital class label information and overpowers DML objective to drive the CoReL performance. DML objective still boosts the class discriminability by providing additional structural information and regularization. Figure 7.7 demonstrates that the α value with the highest rank gradually reduces with the increase in the training dataset size. This indicates a smooth transition between the amount of contributions by the two objectives.

Among the evaluated DML strategies, SoMP($p3, n3$): NAMSS framework achieves the highest F1 score and the lowest DB score as shown in Figure 7.4. Collective assessment of Figure 7.4 and Figure 7.5 depict that SoMP($p3, n3$):NAMSS outperforms the standalone CCE classifier, and performs poorer than the proposed CoReL framework. The observations are consistent for both low and high training data regimes on CRCHistoPhenotypes and AMIDA13 datasets. The observations conclude the proposed NAMSS with SoMP to be the best DML strategy among the competing DML and standalone CCE classifiers.

The proposed CoReL framework provides consistent performance improvement for class-imbalanced datasets by using simple class balancing techniques. For instance, the class-wise distribution of epithelial, inflammatory, spindle-shaped, and miscellaneous categories is 11:4:3:1 in CRCHistoPhenotypes, and 26:9:6:1 in CoNSep. CoReL employs simple data augmentations as described in Section 7.5.3 and Section 7.5.7 to consistently outperform other compared methodologies. Similarly, for AMIDA13 and ICPR12 datasets with mitosis to non-mitosis ratio of, respectively, 1:1000 and 1:265, CoReL handles the

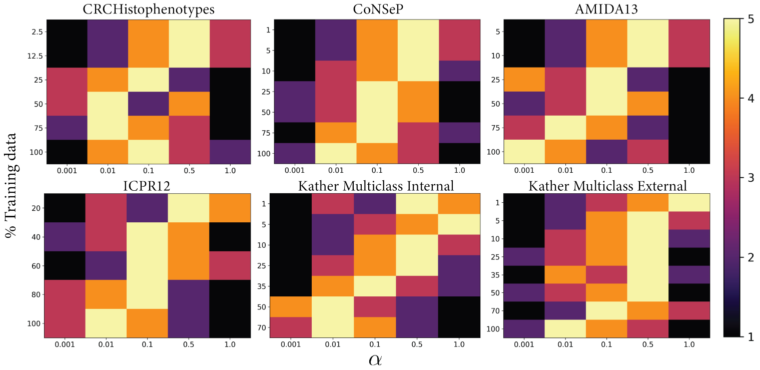


Figure 7.7: Effect of loss trade-off parameter α, signifying the relative contribution of DML towards the classification performance of the CoReL framework for using varying sizes of training data. A higher rank in the colorbar indicates a higher performance.

high class-imbalance by incrementally sampling difficult non-mitoses as explained in Section 7.5.4 and Section 7.5.6.

Furthermore, for all five datasets, the best performing CoReL framework, CCE+SoMP(p3, n3):NAMSS, outperforms the standalone CCE classifier and significantly improves the performance in the limited training data regime. The framework achieves the current state-of-the-art performances by using significantly less training data, i. e., approximately 50% for all evaluation datasets. On using complete training data, CCE+SoMP($p3, n3$):NAMSS outperforms the current state-of-the-art classification results on all datasets. The performance graphs of CCE+SoMP($p3, n3$):NAMSS in Figure 7.5 present a positive slope in the high training data regime. It indicates that the framework possesses the ability to learn further on increasing the training data. The framework contains the same number of trainable parameters as the baseline CCE classifier. Thus, CCE+SoMP($p3, n3$):NAMSS delivers higher efficacy with no additional inference cost. The classification performance values for all the frameworks in Figure 7.4, Figure 7.5 and Figure 7.7 are presented in Appendix A of the supplementary material.

Despite several advantages, CoReL includes certain anticipated limitations. Our proposed DML methodology may be sensitive to the training batch size, where the classification performance increases with increasing batch size for datasets with high variability. This may lead to a high computational cost and long training time per epoch. Furthermore, the joint optimization of CCE and DML objectives may be sensitive to the loss-tradeoff hyperparameter α, which may be computationally expensive to tune.

7.7 Conclusion

Digital pathology can greatly benefit from the advancements in artificial intelligence and computer vision, but the annotation effort required for training these tools pose a major challenge. In this paper, we adapt the training strategy to maximize the learning capabilities of deep neural networks in digital pathology, where the annotation cost is extremely high due to the size, variability, and complexity of the data.

In this paper, we have proposed a co-representation learning for classification (CoReL) framework to enhance the classification performance under data scarcity constraints. CoReL utilizes the annotated class-label information via optimizing a CCE objective, and captures the local spatial distribution information of the data in the embedding space via optimizing a DML objective. The CCE objective maximizes the per-sample mutual information between the ground-truth class-label distribution and the predicted class-label probability distribution. On a side note, the DML objective exploits the pair-wise relationship between the samples from the same class and the samples from different classes. DML optimizes the per-sample mutual information between the anchor embedding and the embeddings of the corresponding positive and negative identities in the local spatial distribution to bring the anchor and the positive samples together while pushing the anchor and the negative samples apart. We showcase the complementarity of the objectives and jointly optimize them to maximize the information utilization from the available data. Our approach has three major components: 1) a loss optimization component that jointly optimizes the CCE objective and the DML objective, 2) a sampling component that selects informative pairs to learn the spatial distribution of data, and 3) a deep metric loss component that simultaneously optimizes the interactions among multiple positive pairs and negative pairs per anchor to accelerate DML. In this regard, we propose a novel informative sampling strategy that uses per-sample neighborhood distribution, and per-pair self-similarity and relative similarities to assess the informativeness of each pair. Further, we propose a soft-multi-pair loss to jointly optimize the interactions of multiple similar pairs and dissimilar pairs to accelerate DML.

We target three classification tasks in digital pathology, i. e., nuclei classification, mitosis detection, and tissue type categorization in this work. We consider five benchmark datasets in this regard to compare the performance of CoReL with the baseline classifier trained by CCE alone. CoReL significantly outperforms the baseline frameworks across all datasets for training with limited annotated data. CoReL achieves state-of-the-art results for utilizing much-reduced training data. On using the complete training data, CoReL outperforms the state-of-the-art classification approaches on all five datasets. Ablation studies establish the impact of the proposed sampling strategy and the proposed deep metric loss towards improving the DML performance. Additional analysis demonstrates the higher contribution by the DML objective to the CoReL performance under limited training data. On increasing the training data, the CCE objective gradually takes over the role of the major contributor.

The CoReL framework poses a great value for majority of the digital pathology classification tasks as they suffer from data scarcity issues. It can cater to a variety of tasks as it does not involve any dataset-specific component. Considering the enhanced performance under limited data, it can be employed in active learning or semi-supervised learning setting to further strengthen the classification performance. Also, the framework is generic enough to be extended to classification tasks in other domains.

7.8 Appendices

Table 7.9: Ablation study: Impact of triplet sampling strategies on classification performance for CRCHistoPhenotypes and AMIDA13 datasets.

	CRCHistoPhenotypes						AMIDA13			
%Training data	12.5%		50%		100%		5%		100%	
Loss+Sampling	F1	DB	F1	DB	F1	DB	F1	DB	F1	DB
Triplet: RS	0.615	4.02	0.725	2.38	0.769	2.16	0.366	1.21	0.607	0.82
Triplet: BHS	0.601	4.18	0.720	3.25	0.753	2.75	0.382	1.43	0.621	0.79
Triplet: DWS	0.620	3.75	0.734	2.41	0.781	2.11	0.404	1.35	0.623	0.78
Triplet: MSS	0.639	3.57	0.735	2.36	0.781	2.19	0.409	1.28	0.633	0.72
Triplet: NADWS	0.634	3.35	0.737	2.34	0.783	2.01	0.415	1.22	0.637	0.67
Triplet: NAMSS	0.642	3.29	0.743	2.14	0.785	1.96	0.420	0.95	0.643	0.60

Table 7.10: Ablation study: Impact of DML losses on classification performance for CRCHistoPhenotypes and AMIDA13 datasets.

	CRCHistoPhenotypes						AMIDA13			
%Training data	12.5%		50%		100%		5%		100%	
Loss+Sampling	F1	DB	F1	DB	F1	DB	F1	DB	F1	DB
Margin: DWS ([Wu et al., 2017])	0.605	4.52	0.740	2.35	0.761	2.12	0.362	1.68	0.619	0.76
MS(p3,n3): MSS ([Wang et al., 2019b])	0.599	3.73	0.734	2.35	0.794	2.25	0.421	1.22	0.627	0.67
Triplet: NAMSS	0.642	3.29	0.743	2.14	0.785	1.96	0.420	1.11	0.643	0.60
Npair(p1,n1): NAMSS	0.627	3.93	0.748	2.34	0.782	2.10	0.400	1.18	0.624	0.70
Npair(p1,n3): NAMSS	0.648	3.43	0.749	2.03	0.789	2.01	0.424	1.05	0.647	0.58
SoMN(p1,n1): NAMSS	0.645	3.32	0.750	2.28	0.784	1.83	0.403	1.34	0.637	0.62
SoMN(p1,n3): NAMSS	0.654	3.19	0.753	2.05	0.792	1.76	0.427	0.99	0.643	0.57
SoMN(p3,n3): NAMSS	0.657	2.98	0.756	1.95	0.799	1.62	0.458	0.89	0.654	0.54

Table 7.11: CRCHistoPhenotypes dataset: Mean and standard deviation of weighted F1 scores for various learning strategies.

%Training Data	2.5%	12.5%	25%	50%	75%	100%
CCE	0.550±0.007	0.610±0.002	0.696±0.005	0.756±0.003	0.779±0.001	0.788±0.004
CCE + Triplet: MSS	0.560±0.003	0.638±0.004	0.698±0.002	0.762±0.006	0.780±0.006	0.793±0.003
CCE + Triplet: NAMSS	0.568±0.003	0.641±0.002	0.708±0.004	0.765±0.004	0.784±0.004	0.796±0.002
CCE + SoMP(p3,n3): NAMSS	0.631±0.004	0.676±0.002	0.724±0.001	0.777±0.003	0.796±0.002	0.808±0.004

Table 7.12: AMIDA13 dataset: Mean and standard deviation of weighted F1 scores for various learning strategies.

%Training Data	5%	10%	25%	50%	75%	100%
CCE	0.363±0.006	0.428±0.012	0.515±0.008	0.583±0.008	0.616±0.005	0.630±0.006
CCE + Triplet: MSS	0.427±0.006	0.513±0.012	0.570±0.006	0.624±0.006	0.638±0.004	0.658±0.004
CCE + Triplet: NAMSS	0.449±0.002	0.534±0.007	0.578±0.007	0.631±0.004	0.640±0.005	0.664±0.002
CCE + SoMP(p3,n3): NAMSS	0.482±0.005	0.546±0.017	0.592±0.006	0.635±0.003	0.648±0.002	0.669±0.001

Table 7.13: ICPR12 dataset: Mean and standard deviation of weighted F1 scores for various learning strategies.

%Training Data	20%	40%	60%	80%	100%
CCE	0.691±0.017	0.733±0.015	0.753±0.008	0.780±0.011	0.798±0.005
CCE + Triplet: MSS	0.712±0.009	0.744±0.007	0.770±0.005	0.798±0.002	0.796±0.005
CCE + Triplet: NAMSS	0.718±0.012	0.749±0.008	0.775±0.005	0.792±0.002	0.803±0.002
CCE + SoMP(p3,n3): NAMSS	0.733±0.010	0.771±0.003	0.795±0.002	0.810±0.004	0.810±0.005

Table 7.14: Kather Multiclass Internal dataset: Mean and standard deviation of weighted F1 scores for various learning strategies.

%Training Data	1%	5%	10%	25%	35%	50%	70%
CCE	0.801±0.004	0.904±0.004	0.954±0.001	0.977±0.001	0.981±0.001	0.982±0.001	0.988±0.001
CCE + Triplet: MSS	0.869±0.002	0.959±0.001	0.975±0.002	0.983±0.001	0.989±0.001	0.990±0.001	0.993±0.001
CCE + Triplet: NAMSS	0.873±0.002	0.960±0.002	0.977±0.002	0.986±0.001	0.990±0.001	0.991±0.001	0.992±0.001
CCE + SoMP(p3,n3): NAMSS	0.884±0.002	0.968±0.001	0.981±0.001	0.989±0.001	0.990±0.001	0.993±0.001	0.994±0.001

Table 7.15: Kather Multiclass External dataset: Mean and standard deviation of weighted F1 scores for various learning strategies.

%Training Data	1%	5%	10%	25%	35%	50%	70%	100%
CCE	0.748±0.005	0.796±0.004	0.851±0.003	0.910±0.002	0.912±0.001	0.918±0.003	0.919±0.003	0.922±0.001
CCE + Triplet: MSS	0.759±0.003	0.803±0.003	0.890±0.003	0.916±0.002	0.918±0.001	0.920±0.002	0.923±0.002	0.928±0.001
CCE + Triplet: NAMSS	0.788±0.001	0.862±0.002	0.895±0.002	0.918±0.002	0.923±0.001	0.923±0.003	0.928±0.003	0.933±0.001
CCE + SoMP(p3,n3): NAMSS	0.835±0.007	0.865±0.001	0.908±0.003	0.929±0.007	0.932±0.001	0.936±0.004	0.940±0.001	0.949±0.002

Table 7.16: Impact of α on the classification results of CoReL trained using varying training data sizes for CRCHistoPhenotypes and CoNSeP datasets.

	CRCHistoPhenotypes						CoNSeP						
α/%Data	2.5%	12.5%	25%	50%	75%	100%	1%	5%	10%	25%	50%	75%	100%
0.001	0.587	0.656	0.705	0.769	0.789	0.795	0.629	0.712	0.745	0.795	0.818	0.830	0.830
0.01	0.593	0.656	0.708	0.771	0.792	0.802	0.632	0.712	0.753	0.799	0.821	0.835	0.833
0.1	0.601	0.658	0.715	0.769	0.790	0.805	0.645	0.727	0.754	0.8160	0.824	0.839	0.832
0.5	0.621	0.6659	0.699	0.769	0.789	0.798	0.652	0.728	0.759	0.804	0.821	0.832	0.832
1.0	0.599	0.658	0.695	0.767	0.784	0.796	0.644	0.720	0.752	0.794	0.817	0.831	0.827

Table 7.17: Impact of α on the classification results of CoReL trained using varying training data sizes for AMIDA13 and ICPR12 datasets.

	AMIDA13						ICPR12				
α/%Data	5%	10%	25%	50%	75%	100%	20%	40%	60%	80%	100%
0.001	0.419	0.413	0.571	0.619	0.628	0.649	0.686	0.733	0.756	0.787	0.801
0.01	0.425	0.471	0.570	0.619	0.645	0.648	0.703	0.734	0.759	0.797	0.806
0.1	0.455	0.485	0.575	0.628	0.638	0.647	0.699	0.745	0.771	0.799	0.802
0.5	0.468	0.509	0.555	0.620	0.627	0.648	0.727	0.744	0.770	0.786	0.792
1.0	0.425	0.477	0.552	0.604	0.607	0.646	0.714	0.727	0.767	0.784	0.786

Table 7.18: Impact of α on the classification results of CoReL trained using varying training data sizes for Kather Multiclass Internal and Kather Multiclass External datasets.

	Kather Multiclass Internal							Kather Multiclass External							
α/%Data	1%	5%	10%	25%	35%	50%	70%	1%	5%	10%	25%	35%	50%	70%	100%
0.001	0.821	0.963	0.976	0.987	0.988	0.992	0.993	0.812	0.836	0.882	0.908	0.911	0.921	0.923	0.928
0.01	0.851	0.963	0.977	0.987	0.989	0.992	0.994	0.824	0.845	0.892	0.915	0.930	0.924	0.923	0.935
0.1	0.836	0.966	0.979	0.987	0.990	0.992	0.994	0.825	0.869	0.900	0.918	0.927	0.928	0.934	0.931
0.5	0.883	0.966	0.979	0.987	0.989	0.992	0.993	0.826	0.874	0.905	0.924	0.931	0.932	0.929	0.928
1.0	0.853	0.967	0.979	0.987	0.989	0.992	0.993	0.827	0.858	0.886	0.906	0.926	0.920	0.929	0.926

8

Weakly Supervised Segmentation and Classification of Prostate Cancer Slides

Segmenting histopathology images into diagnostically relevant regions is imperative to support timely and reliable decisions by pathologists. To this end, computer-aided techniques have been proposed to delineate relevant regions in histopathology slides. However, the techniques necessitate task-specific large datasets of annotated pixels, which is tedious, time-consuming, expensive, and infeasible to acquire for many histopathology tasks. Therefore, weakly-supervised semantic segmentation techniques are proposed to leverage image-level *inexact* annotations, which are cheaper and quicker to acquire. In this chapter, we propose WholeSIGHT, a weakly-supervised semantic segmentation method using tissue-graphs, to simultaneously segment and classify whole-slide images (WSIs) of arbitrary shape and size. Formally, WholeSIGHT first constructs a tissue-graph representation for an input image, where the nodes depict tissue regions, and the edges describe interactions among tissue regions. Subsequently, the method employs a *graph-classification head* to classify WSIs, followed by a post-hoc feature attribution technique to derive node-level pseudo labels. Finally, a *node classification head* is trained using the pseudo labels to segment WSIs. We evaluated WholeSIGHT on three public prostate cancer WSI datasets from three pathology labs. Our method achieved state-of-the-art weakly-supervised segmentation performance on all the datasets, and resulted in better or comparable classification performance with respect to state-of-the-art weakly-supervised WSI classification methods. Further, two Bayesian variants, WholeSIGHT-MCD and WholeSIGHT-DE, based on MC-dropout and deep ensembles, respectively, are proposed, which improves the generalization of WholeSIGHT over external test datasets. The generalization capabilities of the methods are quantified in terms of segmentation and classification performance, uncertainty estimations, and model calibration analyses.

Preliminary version of this chapter has been published as: Valenin Anklin[†], Pushpak Pati[†], Guillaume Jaume[†], Behzad Bozorgtabar, Antonio Foncubierta-Rodriguez, Jean-Philippe Thiran, Maria Gabrani, Orcun Goksel. “Learning Whole-Slide Segmentation from Inexact and Incomplete Labels using Tissue Graphs”, In: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 636-646, Strasbourg, France, Sep 2021, doi: 10.1007/978-3-030-87196-3_59, [arXiv]

This chapter with comprehensive extensions has been prepared to be submitted as: Pushpak Pati[†], Guillaume Jaume[†], Behzad Bozorgtabar, Jean-Philippe Thiran, Maria Gabrani, Orcun Goksel “Weakly Supervised Learning for Joint Whole-Slide Segmentation and Classification in Prostate Cancer”, 2021, for review as a journal publication.

8.1 Introduction

Prostate cancer is the second most frequently diagnosed cancer in men in the U.S., with 250,000 new registered cases resulting in 35,000 deaths in 2021. In contrary, the number of pathologists, who play a pivotal role in the diagnosis and management of cancer patients, is gradually decreasing. In the U.S., a decrease of 18% is recorded between 2007 and 2017, with a consequence of 42% increase in average workload [Wilson et al., 2018]. Moreover, the practice of uro-pathology has its own challenges [Amin et al., 2015]. Indeed, even though diagnostic criteria for prostate cancer grading are established [Tan et al., 2019], the continuum of histologic features phenotyped across the diagnostic spectrum leaves room for inconsistencies, with significant intra- and inter-observer variability [Gomes et al., 2014; Elmore et al., 2015]. Manual slide inspection is also a tedious and time-consuming which would benefit from automation and standardization. All these elements demand the development of AI-based automated CAD tools for diagnosing prostate cancer.

With the advancements in AI, in particular DL, several supervised CAD tools are proposed to assist the diagnosis of prostate cancer [Linkon et al., 2021; Tataru et al., 2021]. Recent studies have also demonstrated that AI-assisted prostate cancer diagnosis significantly outperforms standalone pathologist-based diagnosis [Bulten et al., 2021; Campanella et al., 2019]. Although these DL-based tools achieve remarkable performance, they often require task- and tissue-specific pixel or patch annotations on large datasets. Acquiring such annotations is laborious, time-consuming, and often infeasible.

To alleviate the burden of annotation requirements, several DL-based weakly-supervised methods are proposed across different types of tissues, which can leverage readily available WSI-level annotations. Most of these methods, that are scalable to WSIs, focus on classification tasks, e. g., MIL [Campanella et al., 2019; Lu et al., 2021] or compression-based representation learning [Tellez et al., 2021; Shaban et al., 2020]. Though methods classifying WSIs are important, their applicability is limited due to their poor ability to assist pathologists' focus during diagnosis [Wang et al., 2019c]. To address this limitation with classification methods, semantic segmentation methods are desired that can delineate diagnostically relevant regions in WSIs and speed up the diagnosis by directly guiding their focus to informative regions. Further, a pixel-level segmentation can also enable the quantification of tumor areas for better patient stratification, tailored treatment selection, and strengthening trust between the DL methods and pathologists. However, semantic segmentation of WSIs is more annotation-demanding, i. e., requiring pixel-level labeling, compared to WSI classification. Therefore, weakly-supervised semantic segmentation (WSS) methods are imperative for pathological diagnosis.

While DL-based WSS methods have shown great successes on natural images, they encounter several challenges when applied to histopathology images [Chan et al., 2021], as histopathology images (1) contain finer-grained objects with large intra-class variations [Xie et al., 2019]; (2) often include ambiguous boundaries among different histology components [Xu et al., 2017b]; (3) can be as large as several giga-pixels with arbitrary tissue shapes. Nevertheless, some WSS methods have been proposed for a number of histopathology applications. For instance, the methods by [Xu et al., 2014; Hou et al., 2016; Jia et al., 2017; Xu et al., 2019a; Ho et al., 2021] perform WSS at patch-level. These methods are limited as they require patch-level labels and cannot incorporate global tissue microenvironment to perform contextualized WSI segmentation. While [Chan et al., 2019; Silva-Rodrìguez et al., 2021] propose to analyse larger image-tiles compared to patches, they are constrained in terms of computational complexity and memory requirements to operate on WSIs in an end-to-end manner. Further, the method by [Chan et al., 2019] requires *exact* fine-grained tile-level annotations, i. e., a precise denomination of the presence of each lesion type in the image-tiles, which requires pathologists to annotate images beyond standard clinical needs and norms. Differently, recent WSI classification methods propose to use learned attention weights or feature attribution techniques to highlight salient regions in a WSI that drive the model's prediction [Lu et al., 2021; Tellez et al., 2021]. The identified salient regions are informative for visual assessment, but are insufficient, incomplete, and blurry for accurately delineating diagnostically relevant regions. Further, the saliency of a region signifies its relevance towards the model prediction, but do not convey the class label of the region. In addition, these methods typically require densely overlapping patch-level predictions to obtain a granular saliency map, which is computationally expensive while working with WSIs.

In addition to the above shortcomings, the aforementioned approaches do not include uncertainty estimate analyses, which are crucial to understanding *when* to trust the model predictions. Indeed, DL methods typically tend to produce overconfident predictions and do not indicate when they are likely to be incorrect [Fort et al., 2019], especially when generalizing predictions to unseen cohorts. This can be partially explained by the lack of confidence and uncertainty estimates in neural network parameters, also known as *epistemic uncertainty*. Intuitively, epistemic uncertainty can be correlated to the inter-observer variability in pathology diagnosis, which is known to be high for challenging tasks. Each pathologist, with his/her experience, develops an own understanding of the task. Thus pathologists can be considered as different "models", with different decision boundaries that induce uncertainty in challenging cases. Further uncertainty can be induced due to data, also known as *aleatoric uncertainty*. In pathology, aleatoric uncertainty is caused by, (1) the difficulty of matching the continuum of histologic features to the diagnostic spectrum, (2) intra- and inter-patient tumor heterogeneity, and (3) visualization artifacts that create ambiguous cases. Consequently, *aleatoric* and *epistemic* uncertainty are inherently part of pathology practice and should be considered when developing CAD tools.

Given the above, it is imperative to develop a WSS method that can (1) operate on arbitrarily large histopathology images, e. g., on WSIs; (2) utilize both local and global contexts to conduct precise segmentation; (3) perform simultaneous classification and segmentation; (4) leverage readily available annotations in a clinical setting, without any task-specific assumptions or post-processing; and (5) provide reliable uncertainty estimates as confidence to diagnostic predictions as well as to detect any domain shifts when applied to new datasets.

To address the aforementioned requirements, we propose WHOLESIGHT, "Whole-slide SegmentatIon using Graphs for HisTopathology". Formally, WHOLESIGHT represents a histopathology image using a superpixel-based tissue-graph (TG), and transforms the segmentation task into a *node-classification* task. WHOLESIGHT incorporates both local and global inter-tissue-region relationships to perform contextualized segmentation, principally in agreement with inter-pixel relation-based WSS method [Ahn et al., 2019]. To account for *epistemic* uncertainty, we further propose two Bayesian variants of WSS based on MC-dropout [Gal et al., 2016; Kendall et al., 2017] (MCD) and deep ensembles [Lakshminarayanan et al., 2017; Fort et al., 2019] (DE), respectively. Our major contributions are:

- WholeSIGHT, a novel weakly-supervised semantic segmentation and classification method that can scale to WSIs.
- A thorough evaluation of WholeSIGHT on three prostate cancer datasets for Gleason pattern segmentation and Gleason grading, and comparison against state-of-the-art WSI classification algorithms. WholeSIGHT directly predicts the Gleason pattern associated to each pixel, i. e., Benign (B), grade3 (G3), grade4 (G4), and grade5 (G5), along with the WSI-level grade defined as the combination of the most common (*primary*, P) and the second most common (*secondary*, S) cancer growth patterns in the image.
- A study of the generalization ability of WholeSIGHT, WholeSIGHT-MCD, and WholeSIGHT-DE by testing on *in-domain* and *out-of-domain* cohorts. The generalizability is quantified in terms of segmentation and classification performance, uncertainty estimation, and calibration of neural network predictions.

8.2 Related work

8.2.1 Weakly-supervised histopathology image classification

Most of the weakly-supervised methods in CP are proposed to *classify* histopathology images, i. e., tissue microarrays and whole-slides. EM-CNN is introduced in [Hou et al., 2016], a patch-based method that is trained using image-level labels. It employs an Expectation Maximization (EM)-based method to identify discriminative patches by utilizing the inter-patch spatial relationships, and subsequently uses a decision fusion model to aggregate the patch-level predictions. A two-step approach is proposed in [Campanella et al., 2019], which first identifies informative patches using a patch-level MIL framework, and then adopts a RNN-based strategy to aggregate patch-level predictions for WSI classification. Another MIL approach, CLAM, is proposed in [Lu et al., 2021] that learns class-level attention weights to discriminate diagnostically relevant regions. CLAM is further optimized by learning an instance-level clustering over the patches to constrain and refine the learned feature space. Differently, two-step compression-based procedures are proposed in [Tellez et al., 2021] and [Shaban et al., 2020] to analyse WSIs. First, they extract patch-level embeddings using a network pre-trained on an auxiliary task [Tellez et al., 2021; Shaban et al., 2020], e. g., contrastive learning, or using unsupervised learning [Tellez et al., 2021], e. g., a Variational Auto-Encoder (VAE). Then, they build a compressed feature cube representation of the input WSI, which is further processed by a CNN classifier. Despite the success of these weakly-supervised classification approaches, they cannot directly be extended for semantic segmentation.

8.2.2 Weakly-Supervised histopathology image segmentation

A few methods in literature have been proposed to perform WSS of histopathology images. DWS-MIL is proposed in [Jia et al., 2017], which trains a binary-classifier to generate pixel-level predictions, and then produces image-level prediction using a softmax function. The network is trained to optimize the image-level predictions, and thereby improving the pixel-level predictions. A MIL-based label enrichment method, CAMEL, is proposed in [Xu et al., 2019a] for WSS. It splits an image into latticed instances and automatically generates instance-level labels. After label enrichment, the instance-level labels are further assigned to the corresponding pixels, producing the approximate pixel-level labels and making fully supervised training of segmentation models possible. A deep multi-magnification network is introduced in [Ho et al., 2021] which performs patch-wise multi-class tissue segmentation by using concentric patches across multiple magnifications. This method leverages scribble annotations of regions in WSIs during the training phase. HistoSegNet, proposed in [Chan et al., 2019], performs WSS of histological tissue types in two steps. First, a CNN is trained at tile-level using tile-level annotations to predict the presence of different tissue types in a tile. Then, Grad-CAM, a feature attribution technique is employed to derive pixel-level class predictions. To further improve the segmentation, HistoSegNet employs a complex hand-crafted class-specific post-processing steps. As a main limitation, the aforementioned methods cannot perform WSS on giga-pixel WSIs using only image-level labels, and cannot adapt to WSIs of different sizes. Comparatively, WeGleNet proposed in [Silva-Rodrìguez et al., 2020] is scalable to WSIs. WeGleNet includes a multi-class segmentation layer and a global-aggregation layer to perform image-level classification during training and pixel-level prediction during inference. It aggregates class-wise pixel-level softmax activations to perform image-level task, and significantly upsample the pixel-level activations to segment an image. However, the method is insufficient to precisely delineate different lesions in an image, and is incomplete to highlight multiple occurrings of lesions. Further, it also requires to extract densely-overlapping patches to render fined-grained segmentation. In contrast, our proposed WSS approach can perform WSS by leveraging image-level labels, while efficiently scaling to WSIs with arbitrary shape and size.

8.2.3 Domain shift, generalization, and uncertainty in computational pathology

Domain shift and generalization: Building models that are in the same time robust to domain shifts and able to provide reliable uncertainty estimates is fundamental to deploy CAD tools in real-world [Tellez et al., 2018; Tellez et al., 2019]. Domain shifts are challenging to model and detect in DL. This is prevalent in CP, where domain-level biases are introduced due to various reasons, such as different staining protocols, manufacturing devices, materials, and scanning devices with respective color response [Aubreville et al., 2021]. Nevertheless, several approaches have been proposed to alleviate such domain shifts by developing data- and model-level adaptation mechanisms.

Stain normalization [Reinhard et al., 2001; Macenko et al., 2009; Vahadane et al., 2016; Ren et al., 2019] is a widely employed technique in this direction. It objectively operates at data-level by standardizing the input. Stain normalization is model-agnostic and has been shown to improve generalization performance of DL models [Tellez et al., 2018; Tellez et al., 2019]. Differently, color augmentations are proposed to model staining variations, e. g., by adding additive and multiplicative noise to the input [Tellez et al., 2018; Faryna et al., 2021]. These techniques offer good compromise between ease of integration in DL pipelines and performance gain.

In another scenario, when (unlabeled) samples from the target domain are available during training, domain adversarial learning [Ganin et al., 2016; Aubreville et al., 2020] methods are proven effective for domain adaptation. However, the availability of target domain samples during training is often impractical due to, (1) lack of knowledge about the setup where the model will be deployed, and (2) limitations related to data privacy and regulations. Further, a pre-trained model on a source domain can be fine-tuned by leveraging a few labeled target domain samples, but at the cost of compromising the generalization capabilities of the model.

Uncertainty estimation: While the aforementioned approaches propose various mechanisms to alleviate the impact of the distribution shifts, they do not address the scenario where the distribution on unseen cohorts is drastically different. In this case, accurate *uncertainty* estimates are crucial to know *when* to trust the model. This task is challenging for neural networks, which often provide over-confident predictions, as studied in [Guo et al., 2017; Lakshminarayanan et al., 2017; Fort et al., 2019]. This may hinder real-life deployment in clinics, where CAD must be transparent.

However, CP research along uncertainty estimation is scarce. [Thagaard et al., 2020] benchmarked the detection of adenocarcinoma in H&E lymph node sections from breast cancer under various real-life distribution shifts. Their work concluded that Bayesian neural networks based on deep ensembles [Fort et al., 2019] and MC-dropout [Gal et al., 2016; Kendall et al., 2017; Fort et al., 2019] provide better uncertainty estimates than classical approaches. Our proposed generalization and uncertainty analysis further ascertains the findings of [Thagaard et al., 2020] for WSI-level Gleason grading.

8.3 Methodology

This section presents the proposed WholeSIGHT methodology for scalable WSS of histopathology images. First, an input image is transformed into a tissue-graph (TG) representation, where the nodes and edges of the graph denote tissue regions and their interactions, respectively. Then, a GNN is employed to learn contextualized node embeddings. The resulting node embeddings are processed by a *graph-classification head* for primary and secondary Gleason classification. Upon training the *graph-head*, a feature attribution technique and a node selection strategy are employed to determine pseudo labels for a subset of the nodes, which are further used to train a *node-classification head*. The outcomes of the *node-head* are used to segment the Gleason patterns in the image. An overview of WholeSIGHT is provided in Figure 8.1.

8.3.1 Notation and preliminaries

We define a graph $G \in \mathcal{G}$ as a pair (V_G, E_G), where V_G and E_G represent the set of nodes and edges, respectively, of G, and $\mathcal{G}$ represents the set of graphs. The neighborhood of a node $v \in V_G$ is denoted as $\mathcal{N}(v) := \{u \in V_G \mid (v,u) \in E_G \ \vee \ (u,v) \in E_G\}$. We denote the cardinality of a set as $|.|$, e. g., $|\mathcal{N}(v)|$ indicates the number of neighbors of $v \in V_G$. We are concerned with *attributed* graphs, where $G \in \mathcal{G}$ is associated with d-dimensional node-level attributes H. Attributed graph G is denoted as $G := (V_G, E_G, H)$, where $H \in \mathbb{R}^{|V| \times d}$, or denoted at node-level as $H_{v,.} := h(v) \in \mathbb{R}^d$.

GNNs [Kipf et al., 2017; Xu et al., 2019b; Hamilton et al., 2017; Veličković et al., 2018] are a class of neural architectures that can learn from graph-structured data. In a typical message-passing GNN, the node features are iteratively updated via a two-step procedure to contextualize their feature representation in accordance with their neighborhood node

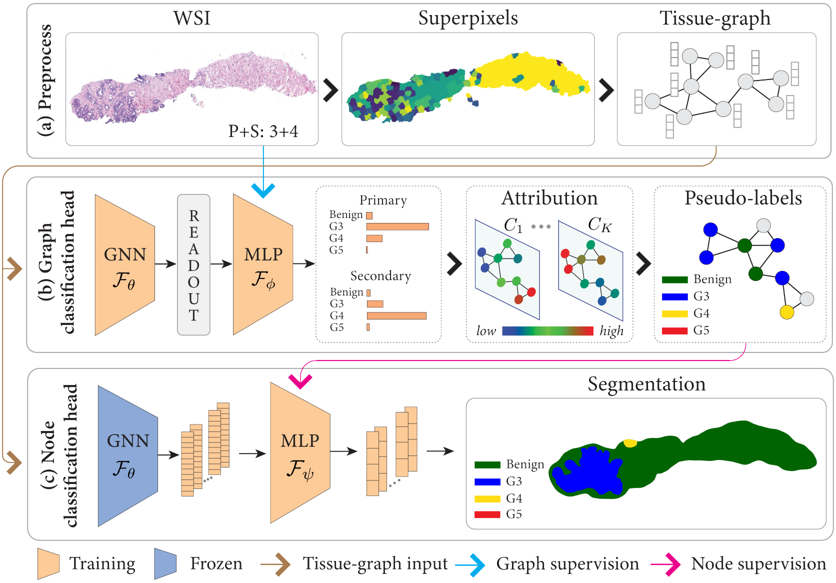


Figure 8.1: Overview of the WholeSIGHT method. (a) Preprocessing transform a WSI into a TG, where the nodes denoting tissue regions are identified via superpixels. (b) *Graph-classification head* contextualizes node-embeddings via a GNN and classifies the TG. In a post-hoc step, a feature attribution method and an importance-based node selection strategy derive node-level pseudo-labels. (c) *Node-classification head* contextualized node-embeddings by the trained GNN, and leverages the pseudo-labels to train a node-classifier, which generates the segmentation output.

information. First in an Aggregate step, for each node $v \in V_G$, the features of the neighboring nodes $\mathcal{N}(v)$ are aggregated by a differentiable and permutation-invariant function. Next in an Update step, the current features of v and the aggregated feature vector of $\mathcal{N}(v)$ are processed by a differentiable operator to produce the new features of v. The above procedure is repeated T times, where T denotes the number of GNN layers.

In this work, we use a version of the GIN architecture [Xu et al., 2019b], where the Aggregate step is based on a *mean*-operator, and the Update step combines the *aggregated* features with the current node features $h(v)$ via a multi-layer perceptron (MLP). Formally, the Aggregate and the Update steps are given as,

$$h^{(t+1)}(v) = \mathrm{MLP}\Big(h^{(t)}(v) + \frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} h^{(t)}(u)\Big) \tag{8.1}$$

These steps are repeated T times, which acquires context information up to T-hops for each $v \in V_G$. The GNN can be denoted as a function $\mathcal{F}_\theta$ that maps the graph nodes to embeddings, where θ are learnable parameters. For a graph classification, a fix-sized graph-level embedding h_G is derived by pooling the node-level feature representations $h^T(v)$, $\forall v \in V_G$ by a Readout step, e. g., a mean-Readout operation. Subsequently, the graph-level embeddings can be mapped to target classes by a neural network classifier $\mathcal{F}_\phi$, where ϕ are learnable parameters. Similarly, for a node classification task, the node-level feature representations $h^T(v)$, $\forall v \in V_G$ can be classified by a neural network classifier $\mathcal{F}_\psi$, where ψ are learnable parameters.

Formally, classification aims to predict a target label $y \in \mathcal{K}$ for an input $x \in \mathcal{X}$, where $\mathcal{K}$ and $\mathcal{X}$ denote the set of classes and the set of inputs, respectively. Given a set of sample pairs $\{(x_i, y_i)\}_{i=1}^N$, where N is the number of samples and $(x_i, y_i) \sim p(x, y)$, the data likelihood can be expressed as $p(Y|X, \theta, \phi) = \Pi_{i=1}^N p(y_i|x_i, \theta, \phi)$. The optimal parameters $(\hat{\theta}, \hat{\phi})$ are obtained by Maximum Likelihood Estimation (MLE), or equivalently by minimizing the Negative Log-Likelihood (NLL) $-\sum_{i=1}^N \log p(y_i|x_i, \theta, \phi)$. In practice, NLL is expressed as a cross-entropy loss, where the model weights are updated by Stochastic Gradient Descent (SGD), or a similar gradient-based optimizer. In a graph classification setting, a sample pair is denoted as (y_G, G), $y_G \in \mathcal{K}_\mathcal{G}$, $G \in \mathcal{G}$. In node classification, a sample pair is denoted as (y_V, v), $y_V \in \mathcal{K}_\mathcal{V}$, $v \in \mathcal{V}$. For the considered task at hand, the set of graph- and node-level classes are the same, simplifying notation to $\mathcal{K} := \mathcal{K}_\mathcal{G} = \mathcal{K}_\mathcal{V}$.

We further introduce the notion of model *calibration*. Intuitively, the probability of outcomes, i. e., confidence scores, of a calibrated model should match its performance. For example, the samples predicted with an average confidence of 60% by a model should have an average accuracy of 60%. Formally, for a given network, $f : \mathcal{X} \rightarrow \mathcal{K}$, and $p(X, Y)$ a joint distribution over the data and the labels, $f(x)$ is said to be calibrated with respect to p if, $\mathbb{E}_p[Y|f(X) = \beta] = \beta, \forall \beta \in [0, 1]$. The *calibration* can be visualized with a *reliability diagram* [DeGroot et al., 1983]. Namely, all the samples in the dataset are assigned to bins according to their predicted confidence scores by the network. Then, the network performance, e. g., accuracy, is computed for all the samples in each bin. The network performance is plotted against the binned confidence scores, where deviations from the diagonal represent uncalibrated bins.

8.3.2 Preprocessing and tissue-graph construction

The input H&E stained images in the dataset are first stain-normalized using the algorithm by [Vahadane et al., 2016] to reduce any appearance variability across the images due to tissue preparation. Stain normalization is crucial since such variabilities can adversely impact the computational methods for downstream diagnosis [Veta et al., 2014; Tellez et al., 2019]. In the next step, a stain normalized image is transformed into a TG, Figure 8.1(a), where the nodes and the edges of the TG denote tissue regions and inter-tissue interactions, respectively. Motivated by [Bejnordi et al., 2015], we consider superpixels as the visual primitives to encode the tissue regions for this work. In comparison to rectangular patches, superpixels are flexible units to accommodate arbitrary shapes in accordance with the local homogeneity of the tissue in an image. The homogeneity constraint also restricts the superpixels to span across multiple distinct structures and include different morphological regions.

The high-level steps in a TG construction are, (1) the construction of superpixels to define the nodes V_G, (2) characterization of the superpixels to define the node features H, and (3) the construction of the graph topology to define the edges E_G. For identifying the superpixels in an input image, a two-step procedure is adopted. First, unsupervised Simple Linear Iterative Clustering (SLIC) algorithm [Achanta et al., 2012] emphasizing on space proximity is employed on the image to produce over-segmented superpixels. The SLIC algorithm is applied on a low magnification of the image to capture homogeneity, while offering a good compromise between granularity and smoothing-out noise. Second, the over-segmented superpixels are hierarchically merged according to their channel-wise color similarity at high magnification. The color similarity is quantified in terms of channel-wise 8-bin color histograms, mean, standard-deviation, median, energy, and skewness. The resulting merged tissue regions form the nodes of the TG. The merging allows to semantically group the superpixels and render meaningful tissue regions. In addition, the merging reduces the node complexity of the TG, thereby enabling the scaling of TG to large dimensional histopathology images and contextualization to distant nodes.

To characterize the nodes of the TG, we extract morphological and spatial features from the tissue regions constituting the nodes. Considering the arbitrary dimensions of the superpixels, a two-step process is adopted to extract deep learning-based morphological features. First, patches of size 144×144 pixels are extracted from a superpixel, resized to 224×224 size, and encoded into 1280-dimensional features by processing through a MobileNetV2 network [Sandler et al., 2018] pre-trained on ImageNet [Deng et al., 2009]. Then, the corresponding node-level morphological features are computed as the mean of the individual patch-level features. Further, spatial features of the nodes are computed by normalizing the superpixel centroids by the image dimensions. The normalization ensures the invariability of the spatial features with respect to the varying dimensions of the input histopathology images. Finally, the TG topology is defined by constructing a Region Adjacency Graph (RAG) [Potjer, 1996] using the spatial connectivity of superpixels. To this end, we assume that adjacent tissue regions biologically interact the most, and thus should be connect in the TG topology.

8.3.3 Contextualized node embeddings

Given a TG, we aim to learn discriminative node embeddings (see Figure 8.1(b)) by utilizing the context information of the nodes, i. e., the tissue micro-environment and the inter-tissue interactions. The contextualized node embeddings are subsequently used for WSI classification and WSS. To contextualize the node embeddings, we use a GIN [Xu et al., 2019b] graph neural network, denoted as $\mathcal{F}_\theta$ and parametrized by the learnable parameters θ. Since GNNs can operate on graphs of arbitrary and varying sizes, they allow to encode histopathology images represented in the form of TGs without the need for tile-based processing. As the discriminative information, dependent on the sub-graph structures, can lie at different abstraction levels in the GNN, we employ a Jumping Knowledge (JK) strategy to incorporate multi-level node representations. Namely, the final node-level embedding after T GIN-layers is defined as,

$$h^{(T)}(v) = \text{CONCAT}(h^{(t)}(v),\ \forall t \in \{1, ..., T\}) \tag{8.2}$$

where, CONCAT denotes a concatenation operation.

8.3.4 WSI classification

Following the contextualized node embeddings, a *graph-classification head* is employed to classify the TG by leveraging image-level *inexact* labels. To this end, first, a READOUT averages out the information from all the nodes $h^{(T)}(v),\ \forall v \in V_G$ to build a fix-sized graph-level embedding h_G. Subsequently, the graph-level embedding is fed to a multi-task classifier for primary and secondary Gleason grading. Specifically, the classifier is composed of two parallel MLPs, denoted as $\mathcal{F}_\phi = \{\mathcal{F}_{\phi_1}, \mathcal{F}_{\phi_2}\}$, which are parametrized by trainable parameters $\phi = \{\phi_1, \phi_2\}$. The two MLPs individually predict the primary, i. e., the worst Gleason pattern, and secondary, i. e., the second worst Gleason pattern, in the WSI. Each MLP solves a multi-class problem with $|\mathcal{K}|$ Gleason pattern classes, i. e., B, G3, G4, G5. The final Gleason grade is derived as the sum of the predicted primary and secondary Gleason patterns. $\mathcal{F}_\theta$ and $\mathcal{F}_\phi$ are optimized jointly by minimizing the weighted multi-label cross-entropy loss,

$$\mathcal{L}_G = \lambda \mathcal{L}_{CE}(y_{G_P}, \hat{y}_{G_P}) + (1-\lambda)\mathcal{L}_{CE}(y_{G_S}, \hat{y}_{G_S}) \tag{8.3}$$

where, P and S denote the primary and the secondary labels of ground truth y_G and prediction $\hat{y}_G$, and $\lambda \in [0,1]$ is a hyper-parameter used to balance the two terms. Gleason grading is typically imbalanced, where WSIs with higher grade patterns are less frequent. To address this, we define class-weights as $w := \{\log(\frac{\sum_i N_i}{N_i}),\ i = \{1, ..., |\mathcal{K}|\}\}$, where N_i is the count of class-wise Gleason patterns. The weights are designed such that a higher value is assigned to classes with lower frequency.

8.3.5 Weakly supervised semantic segmentation

The nodes in a TG are identified by superpixels that denote morphologically homogeneous tissue regions. Since each Gleason pattern is characterized by *distinct* morphological patterns, we assume that each tissue region, depicted by a node of the TG, includes a *unique* Gleason pattern. Thereby, the WSI segmentation task is transformed into a classification task of the nodes in the TG. In the presence of only image-level labels, the node classification task is achieved in two steps. First, pseudo-node labels are generated by leveraging the image-level annotations, and subsequently the pseudo-node labels are used to train a node classifier.

8.3.5.1 Pseudo node label generation

Following the image-level classification in Section 8.3.4, a post-hoc *feature attribution* technique is employed to measure the importance of each node for TG classification. Specifically, we use GraphGrad-CAM [Pope et al., 2019; Jaume et al., 2021b], an extension of Grad-CAM [Selvaraju et al., 2017] technique to operate with GNNs. For a graph G, GraphGrad-CAM produces class-wise node attribution maps, A_k, $\forall k \in \mathcal{K}$. The attribution maps highlight the importance $\forall v \in V_G$ towards the classification of G into $|\mathcal{K}|$ categories, as demonstrated in Figure 8.1. Given the importance scores of a node $v \in V_G$ towards $|\mathcal{K}|$ classes, a simple and straight-forward approach would be to assume that the class label of v is $k \in \mathcal{K}$ if the highest importance score corresponds to class k. At this stage, an *argmax* operation across the class-wise importance scores $\forall v \in V_G$ can be considered to classify the nodes. However, such node classification strategy carries several disadvantages.

- An *argmax* operation for a node greedily selects the class label with the highest importance score. However, some nodes only marginally contribute to the graph classification, e. g., background nodes, and bear low importance scores for all $k \in \mathcal{K}$.

An *argmax* operation would confidently label such nodes into one of the $\mathcal{K}$ classes, which reduces confidence in the node classification.

- The class labels of the nodes that highly contribute towards a certain class cannot be guaranteed to be the same as the corresponding class label. For example, a node can bear high importance if it provides useful complementary information for tie-breaking or ruling out another class possibility. Formally, if the set of nodes $V_k \subset V$ have high importance scores for class k, then the class labels of V_k are not ensured to be k. Further, the labels of $v \in V_k$ are also not ensured to be the same, e. g., a node $v \in V_k$ can be an evidence of the *absence* of all classes $\mathcal{K} \setminus \{k\}$, thus bearing high importance for classifying the graph as k, while not being of this class.
- GRAPHGRAD-CAM does not necessarily highlight all the nodes that belong to a class in the corresponding class attribution map. Depending on the complexity of a classification task, a classifier may utilize only a subset of the informative nodes corresponding to a class for graph classification. Formally, if the set of nodes $V_k \subset V$ have high importance scores for class k, then V_k may not include all the nodes in $\mathcal{V}_k \subset V$ that have the actual label k, i. e., $V_k \subset \mathcal{V}_k$.
- There are several feature attribution techniques in literature that can be employed to assign node-wise importance scores and perform node classification. However, as demonstrated in [Jaume et al., 2021b], differences in the underlying mechanisms of these techniques lead to different node-wise importance scores. Therefore, a single feature attribution technique, e. g., GRAPHGRAD-CAM, may not be trusted for a score-based node classification.

Therefore, we devise a strategy to use the highlighted nodes by GRAPHGRAD-CAM as pseudo-labels to train a *node-classification head*. The strategy aims to create pseudo-labels while minimizing the class-wise false positives and false negatives. Specifically, for a graph G with Gleason score $P + S$, such that $P, S \in \mathcal{K}$, we compute the node importance scores I_P and I_S $\forall v \in V_G$. I_P and I_S are computed by using un-normalized GRAPHGRAD-CAM on the P-th class and the S-th class in the primary and secondary graph-classification heads. Since the importance scores by GRAPHGRAD-CAM are unbounded, employing a fixed threshold on the importance scores across all samples is sub-optimal. Therefore, we select the top $n\%$ nodes, denoted as V_P and V_S, based on the respective importance scores I_P and I_S. n is a hyperparameter, which is tuned during the training phase. For a node $v \in V_P$ and $v \in V_S$, we compute $\arg\max(I_P(v), I_S(v))$ to assign v to either of the sets. This ensures that $V_P \cap V_S = \emptyset$. Subsequently, we label the nodes $v \in V_P$ as P and the nodes $u \in V_S$ as S. This process ensures to select the most important set of nodes corresponding to the ground truth image-level label of G, and create the pseudo-labels, denoted as $y_{\tilde{V}}$. Continuing this process for all the TGs in the dataset produces pseudo-node labels across all classes, denoted as $Y_{\tilde{V}}$.

8.3.5.2 Node classification

The pseudo-node labels $Y_{\tilde{V}}$ are used to train a *node-classification head*, as shown in Figure 8.1. Specifically for a graph G, we extract the node embeddings $h^{(T)}(v)$, $\forall v \in V_G$ using $\mathcal{F}_{\hat{\theta}}$, where $\hat{\theta}$ are the parameters from the graph classification in Section 8.3.4. $\mathcal{F}_{\hat{\theta}}$ is kept frozen during the node classification to ensure that the *same* GNN backbone can be used for both segmentation and classification, thereby reducing the number of trainable parameters. The node embeddings are processed by an MLP classifier $\mathcal{F}_\psi$, parameterized by learnable parameters ψ, to predict the pseudo-node labels. The *node-classification head* $\mathcal{F}_\psi$ is trained by minimizing a weighted multi-class cross-entropy objective and a deep metric learning objective (as described in Chapter 7), denoted as,

$$\mathcal{L}_V = \lambda w \mathcal{L}_{CE}(y_{\tilde{V}}, \hat{y}_V) + (1-\lambda)\mathcal{L}_{DML}(y_{\tilde{V}_a}, y_{\tilde{V}_p}, y_{\tilde{V}_n}) \tag{8.4}$$

The (a, p, n) denotes *(anchor, positive, negative)* triplets that are selected in an online manner by using batch-hard sampling technique. In this work, triplet loss is used as $\mathcal{L}_{DML}$. The metric learning objective assists the node classification cross-entropy objective by further constraining the node-embedding space. Similar to the graph classification setting, the class-weights are defined as $w := \{\log(\frac{\sum_i N_i}{N_i}),\ i = \{1, ..., |\mathcal{K}|\}\}$, where N_i is the number of annotated nodes of class i. The node-wise predicted class labels are finally used to obtain the segmentation prediction.

We refer to our proposed method, the simultaneous WSI classification and pseudo-node labeling-based WSS, as WholeSIGHT. Noticeably, unlike [Chan et al., 2019], WholeSIGHT does not involve any customized post-processing, thus being a generic method that can be applied to various organs, tissue types, segmentation tasks, etc.

8.3.6 Extension to Bayesian models

We propose two Bayesian variants of WholeSIGHT to incorporate uncertainty estimates into model predictions. We assume that *aleatoric* uncertainty, i. e., data uncertainty, is already modeled during network training and reflected in the predicted probabilities of WholeSIGHT. Since *epistemic* uncertainty is not explicitly captured by WholeSIGHT, we propose to model this using WholeSIGHT-MCD based on MC-dropout [Gal et al., 2016; Kendall et al., 2017] as well as WholeSIGHT-DE based on deep ensembles [Lakshminarayanan et al., 2017; Fort et al., 2019]. These methods are built on the fact that there exist several sets of parameters that can explain a given dataset equally well, i. e., a set of WSIs and WSI labels. The underlying principle of these methods aims to utilize multiple optimal models to capture the variations in the decision boundaries of the individual models, thereby accounting for the epistemic uncertainty. These methods are also crucial when generalizing to unseen cohorts, including distribution shifts in the data.

Deep Ensembles: Deep ensembles are realized by training several models with *different network initializations*, herein exploring diverse modes in function space. In our graph classifier, recall that the conditional distribution $p(y_G|G, \theta, \phi)$ is approximated by $\mathcal{F}_\phi(\mathcal{F}_\theta(G))$, which learns an optimal parameter set $(\hat{\theta}, \hat{\phi})$ by MLE. Using different network weight initializations, we can learn potentially *different* optimal parameters $\{\hat{\theta}^{(m)}, \hat{\phi}^{(m)}\}_{m=1}^{M}$, where $m \in \{1, ..., M\}$ denotes different models. Then, for a test sample $G^* \in \mathcal{G}$, WholeSIGHT-DE output is obtained by computing the average prediction from all the models, i. e.,

$$\hat{p}(y_G^*|G^*) := \frac{1}{M}\sum_{m=1}^{M} p(y_G^*|G^*, \hat{\theta}^{(m)}, \hat{\phi}^{(m)}) \tag{8.5}$$

For node classification and WSI segmentation, a similar approach is employed where $p(y_V|v, \theta, \psi)$ is approximated by $\mathcal{F}_\psi(\mathcal{F}_\theta(v))$.

MC-dropout: MC-dropout [Gal et al., 2016; Gustafsson et al., 2020] follows the same principle to propose a modification of the use of *dropout* layer in the network. Unlike the standard DL networks which utilize dropout only during training, MC-dropout proposes to retain the dropout layers during inference as well. Due to the dropout layer, that randomly switches-off some neurons in the network, each forward pass during inference operates on a different network defined as a *random* subset of the original network. The randomly sampled networks can be viewed as an *ensemble* of networks that provide different decision boundaries, and thus different predictions. As in deep ensembles, the output WholeSIGHT-MCD predictions are obtained by averaging the network predictions over N passes with different dropout patterns.

8.4 Datasets

We evaluate our proposed method on three prostate cancer datasets that are acquired from three independent data sources, and consist of whole-slide prostate cancer needle biopsies. We use these datasets for simultaneously segmenting Gleason patterns in the WSIs and classify the WSIs into different Gleason grades. The Gleason patterns range from G3, characterized by moderately differentiated nuclei and the presence of poorly-formed and cribiform glands, to G4, that include poorly differentiated nuclei and irregular masses, to grade G5, characterized by even less differentiated nuclei and lack or only occasional glands. Normal glands and non-epithelial tissue regions are categorized as B. The Gleason grade is estimated from a Gleason score which is presented as *primary + secondary*, where the *primary* and the *secondary* denote the worst and the second worst Gleason patterns, respectively. Details of the datasets are presented as follows:

Radboud dataset: The Radboud dataset [Bulten et al., 2020a] is composed of 5,759 core needle biopsies extracted from 1,243 patients. The dataset were acquired between January 1, 2012 and December 31, 2017, from patients who underwent prostate biopsy for suspected cancer at the Radboud University Medical Center. The slides were scanned with a 3D Histech Panoramic Flash II 250 scanner at 20× magnification, pixel resolution 0.24μm, and were further downsampled to 10×. The annotations include WSI-level Gleason scores and noisy pixel-level segmentation masks of Gleason patterns, which were made available as part of the Prostate cANcer graDe Assessment (PANDA) challenge [Bulten et al., 2020b]. These segmentation masks were cleaned for the purpose of Gleason pattern segmentation by using standard image manipulation techniques, such as contextualized noise removal, hole filling, and edge smoothing. In the absence of large public datasets with pixel-level annotated prostate cancer WSIs, we utilized the Radboud dataset for the development and evaluation of our methods.

Karolinska dataset: The Karolinska dataset [Ström et al., 2019] comprises of 5,662 core needle biopsies from 1,222 patients. The data were acquired on men aged between 50 and 69 years, between 2012 and 2015 from various hospitals in Stockholm, Sweden. The slides were scanned with a Hamamatsu C9600-12 and an Aperio Scan Scope AT2 scanner at 20× magnification, with pixel resolution of 0.45202μm and 0.5032μm, respectively. Gleason scores of the biopsies were annotated by an expert uro-pathologist.

Sicap dataset: The Sicap dataset [Silva-Rodrìguez et al., 2020] contains 18,783 patches of size 512×512 with *complete* pixel-level annotations and WSI-level Gleason scores from 155 WSIs extracted on 95 patients. The original WSIs and annotation masks

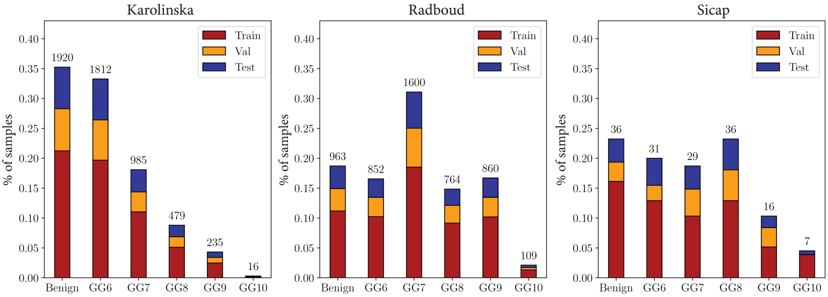


Figure 8.2: Class distribution of the Karolinska, Radboud, and Sicap datasets.

were reconstructed by stitching the patches in the dataset. The WSIs were scanned at 40× resolution with a Ventana iS-can Coreo scanner, and further downsampled to 10× magnification for processing. Pixel- and WSI-level annotations were acquired by a group of expert urogenital pathologists at the Hospital Clínico of Valencia.

Each dataset is split into train, validation, and test in a ratio of 60%, 20%, and 20% at Gleason grade-level, using a random stratified partition that preserves the percentage of samples in each class. No further sample-level analysis was performed to partition the data. The Gleason grade-wise dataset distribution is displayed in Figure 8.2, which highlights the different class-level imbalances across the three datasets. Karolinska dataset is more skewed towards benign and low-grade Gleason categories. The Gleason grade-wise distribution is the most balanced in Radboud dataset. Notably, all three datasets contain a lower fraction of high-grade Gleason categories.

8.5 Results

8.5.1 Implementation and metrics

WholeSIGHT is implemented by using PyTorch [Paszke et al., 2019], DGL [Wang et al., 2019b], and Histocartography [Jaume et al., 2021a]. The experiments were conducted on NVIDIA Tesla P100 GPUs and POWER9 CPUs.

To develop the WholeSIGHT network architecture, the GNN backbone $\mathcal{F}_\theta$, the *graph-classification head* $\mathcal{F}_\phi$, and the *node-classification head* $\mathcal{F}_\psi$ were developed by setting and optimizing their respective hyperparameters. First, $\mathcal{F}_\theta$ and $\mathcal{F}_\phi$ were trained by using image/graph-level labels, and afterwards pseudo-node labels were created to train $\mathcal{F}_\psi$. The segmentation output was obtained via node classification from $\mathcal{F}_\psi$. The number of GIN layers in $\mathcal{F}_\theta$ are optimized for the values $\{3, 4, 5\}$, where the Update function was defined as a 2-layer MLP with 64 hidden units, and Rectified Linear Unit (ReLU) activations. The *graph-classification head* $\mathcal{F}_\phi$ contains two heads for classifying *primary* and *secondary* Gleason categories, where each head consists of a 2-layer MLP with 128 hidden units and ReLU activations. The *node-classification head* $\mathcal{F}_\psi$ contains a 2-layer MLP with 128 hidden units and ReLU activations.

For the Sicap dataset, that consists of a few WSIs, node-level augmentation techniques are employed to augment the graph dataset. Specifically, random node rotations $\{90, 180, 270\}$ degrees, and horizontal and vertical mirroring are used for augmenting the nodes. The batch size and the learning rate were optimized from $\{4, 8, 16\}$ and $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$ set of values, respectively. Dropout layers with dropout rates 0.2, 0.5, and 0.5 were included in the MLPs belonging to $\mathcal{F}_\theta$, $\mathcal{F}_\phi$, and $\mathcal{F}_\psi$, respectively.

Following the hyperparameter tuning, eight WholeSIGHT models were trained with *different* network initializations. The reported WholeSIGHT results correspond to the mean and standard deviation obtained over these eight models. A similar approach was employed for WholeSIGHT-MCD, where each model was run 25 times on different sampled networks created randomly by using the dropout layers. WholeSIGHT-DE was defined by randomly sampling five out of the eight trained models. This process was repeated eight times to obtain different ensemble-based predictions. All the algorithms were trained with Adam optimizer [Kingma et al., 2015].

The model selection criteria during training relied on the version of the WholeSIGHT method. For the first version, model with the best Gleason grade weighted-F1 on the validation set was selected, whereas for the other two versions, model with the best node-classification weighted-F1 score on the validation set was selected. For creating the pseudo-node labels, several percentages of the most important nodes were selected, where the experimented percentage values were $\{5, 10, 15, 20\}$.

Classification metrics: WSI classification performance is measured by the weighted-F1 score of the predicted Gleason grade. In accordance with the prior work [Bulten et al., 2020a; Bulten et al., 2021], we also report the quadratic kappa score (κ^2) of the predicted ISUP grade [Epstein et al., 2005; Epstein et al., 2016]. ISUP grading is an alternative grading system which corresponds with Gleason grade as, Benign → ISUP-0, GG(3+3) → ISUP-1, GG(3+4) → ISUP-2, GG(4+3) → ISUP-3, GG8 → ISUP-4, and GG$\geq$9 → ISUP-5. κ^2 measures the level of disagreement between the prediction and ground truth.

Segmentation metrics: The segmentation performance is measured by the Dice score between the ground truth and the predicted Gleason pattern segmentation masks. The Dice score is equivalent to F1-score at pixel-level predictions. In view of the imbalance of the Gleason patterns in the datasets, we also report the per-pattern Dice score.

Uncertainty metrics: Following the previous work of [Gomariz et al., 2021], we evaluate the classification and segmentation uncertainties by computing the Brier score s_B (lower is better) and the NLL s_NLL (lower is better) over a set of N test samples, expressed as,

$$s_\text{B} = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{|\mathcal{K}|} (y_i - \hat{y}_i)^2, \quad s_\text{NLL} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{|\mathcal{K}|} p(y_i) \log \hat{p}(y_i) \tag{8.6}$$

Intuitively, the uncertainty estimates will be good, (1) when the model performance is high, and (2) when the misclassified samples are not highly confident in their predictions.

Calibration metrics: Reliability diagrams provide an intuitive understanding of model calibration. To quantify the observations in a reliability diagram, we use the Expected Calibration Error (ECE) metric [Kumar et al., 2018]. It computes the weighted average deviation of the confidence scores over all the bins. Formally, it is expressed as,

$$c_\text{ECE} = \sum_{b=1}^{B} \frac{N_b}{N} |\text{acc}(b) - \text{conf}(b)|, \tag{8.7}$$

where n_b represents the number of samples in bin b, $\mathrm{acc}(b)$ and $\mathrm{conf}(b)$ denote the accuracy and average confidence of samples in the bin b, respectively.

8.5.2 Baselines

WholeSIGHT(Graph, GraphGrad-CAM): This variant of WholeSIGHT uses only image/ graph-level supervision during training. Compared to the proposed WholeSIGHT method, this baseline contains only the GNN backbone $\mathcal{F}_\theta$ and the graph-classification head $\mathcal{F}_\phi$. It does not create or utilize pseudo labels, and the segmentation output is obtained by taking the *argmax* over the class-wise GraphGrad-CAM attribution maps.

WholeSIGHT(Multiplex, NC): This variant of WholeSIGHT leverages both *inexact* image- and *complete* pixel-level supervision during training. It acts as the upper-bound for WholeSIGHT method. As pixel-level annotations are available, the node-classification head is trained by using ground-truth node-level labels, instead of generated pseudo-node labels. It constitutes of the same GNN backbone $\mathcal{F}_\theta$, graph-classification head $\mathcal{F}_\phi$, and node-classification head $\mathcal{F}_\psi$ as the WholeSIGHT architecture. In this setting, $\mathcal{F}_\theta$, $\mathcal{F}_\phi$, and $\mathcal{F}_\psi$ are trained jointly by optimizing a multi-task objective, i. e., WSI-level primary and secondary Gleason score prediction along with node-level Gleason pattern prediction. This variant of WholeSIGHT was proposed in our preliminary work, as described in [Anklin et al., 2021].

CLustering-constrained Attention Multi Instance Learning (CLAM): CLAM [Lu et al., 2021] is a clustering-constrained attention MIL approach designed for WSI classification. Our experiments are based on the publicly available implementation of CLAM [1]. Minor modifications were performed to adapt the algorithm for a multi-task objective, i. e., primary- and secondary Gleason score classification. Specifically, patches of size 256 × 256 were extracted from a WSIs. Each patch was further processed by a ResNet50 model pretrained on ImageNet, where features after the third residual block are extracted with an adaptive mean-spatial pooling operation, which resulted in a 1024-dimensional feature representations. The attention module was using a self-attention network with sigmoid activations and 0.25 dropout. The clustering module, that learns class-level representations, was trained by using outputs of the attention network as pseudo-labels and a smooth top1 SVM loss. The attention-weighted patch features were finally passed to a linear classifier for classifying the primary and secondary Gleason scores.

Neural Image Compression (NIC): NIC [Tellez et al., 2021] creates feature cube representations of WSIs to learn a mapping between deep patch features and WSI-level class labels. Our implementation and experiments are partially based on the publicly available implementation [2], which required to be completed with training utilities, dataloaders, and model translation in PyTorch. Specifically, input WSIs were resized to the dimensions of the largest WSI in our datasets with padding. It allowed to associate each WSI to WSI-level label without further processing. Different patch feature extraction strategies were experimented to extract the compressed WSI representations. In our experiments, we found that NIC with BiGAN features (see [Tellez et al., 2021] for implementation details) led to the best performance. A custom CNN with eight convolutional layers was trained from scratch, where each layer has 128 channels, a batch normalization module,

---

[1] CLAM publicly available code: https://github.com/mahmoodlab/CLAM

[2] NIC publicly available code: https://github.com/davidtellez/neural-image-compression

0.2 dropout, and stride 1. As a significant portion of the input is background, the average pooling was replaced by max pooling to extract the most relevant regions per channel. Then, the primary and secondary Gleason pattern classifiers were implemented as 2-layer MLPs with 128 channels and LeakyReLU activations. The network was trained with a multi-class cross entropy loss.

For all the baselines, hyper-parameter search was conducted to find the best learning-rate and batch size, if applicable. Subsequently, eight models were re-trained from scratch with the optimal set of parameters. For each experiment, we report the average and standard deviation over these runs without further model selection.

8.5.3 WSS performance analysis

Training setting: We study the classification and segmentation performance of the proposed WholeSIGHT method, and compare against the aforementioned baselines on three datasets, i. e., Karolinska, Radboud, and Sicap datasets. These evaluations measures the standalone applicability of the methods across independent train and test datasets.

Results analysis: Table 8.1 presents the classification and segmentation results on the Sicap dataset. The analyses are performed under two supervision settings, namely *complete* ($\mathcal{C}$) and *inexact* ($\mathcal{IE}$). The $\mathcal{C}$ setting utilizes both *inexact* image-level labels and the pixel-level annotations. Whereas, the $\mathcal{IE}$ setting only uses the *inexact* image-level labels. WholeSIGHT reaches 39.3% average Dice score, which significantly outperforms WholeSIGHT(Graph, GraphGrad-CAM) by +8.6% in absolute. Further, WholeSIGHT significantly outperforms HistoSegNet in terms of both classification and segmentation metrics. WholeSIGHT(Multiplex, NC), which acts as the upperbound, results in slight improvement in classification and a significant gain in segmentation comapared to WholeSIGHT. The per-class Dice scores indicate that the benign patterns, that constitute most of the tissue area, have a high detection rate compared to less occurring Gleason patterns. For the classification task, WholeSIGHT outperforms NIC and CLAM methods both in terms of Gleason grade weighted-F1 and ISUP κ^2. However, considering the small size of the Sicap test set, the classification performance assessment on the Radoud and Karolinka datasets reveal a more confident picture.

Table 8.2 presents the classification and segmentation results on Radboud. WholeSIGHT renders an absolute gain of +10.33% in average Dice score over WholeSIGHT(Graph, GraphGrad-CAM). This confirms the utility of pseudo-node labels for a superior segmentation. WholeSIGHT(Multiplex, NC) remains a good upper-bound with an average Dice score of 64.99 ± 0.4. The observations of class-wise Dice scores are consistent with Sicap, where the benign patterns have a high detection rate, followed by G3, G4, and G5 patterns. As the Radboud dataset includes more G5 patterns than Sicap, we observe a significant gain in detecting high-grade patterns. For the classification task, the observations are consistent with the observations on the Sicap dataset. Noticeably, the complementarity of the image- and pixel-level annotations results in a better classification performance for WholeSIGHT(Multiplex, NC) than WholeSIGHT.

Table 8.3 presents the classification results on Karolinska. In the absence of ground truth pixel-level annotations, the segmentation performances could not be computed. WholeSIGHT outperforms NIC and produces comparable classification performance with respect to CLAM. The Gleason grade weighted-F1 score is higher for Karolinska

Table 8.1: Classification and segmentation results on Sicap dataset. The best performances for using image-level supervision are highlighted in **bold**.

Annot.		per-class Dice				avg. Dice	GG wF1	ISUP κ^2
	Method	Benign	Grade3	Grade4	Grade5			
$\mathcal{C}$	WholeSIGHT (Multiplex, NC)	$91.1_{\pm 1.0}$	$39.4_{\pm 1.6}$	$52.9_{\pm 1.4}$	$10.6_{\pm 5.4}$	$48.7_{\pm 1.3}$	$55.0_{\pm 1.7}$	$86.2_{\pm 3.1}$
$\mathcal{IE}$	NIC [Tellez et al., 2021]	-	-	-	-	-	$35.3_{\pm 5.0}$	$44.5_{\pm 14.2}$
	CLAM [Lu et al., 2021]	-	-	-	-	-	$53.8_{\pm 3.5}$	$61.8_{\pm 5.5}$
	HistoSegNet [Chan et al., 2019]	$71.5_{\pm 1.4}$	$1.5_{\pm 0.7}$	$8.4_{\pm 0.9}$	$1.6_{\pm 0.3}$	$22.4_{\pm 0.3}$	$16.7_{\pm 4.3}$	$36.7_{\pm 2.8}$
	WholeSIGHT (Graph, GraphGrad-CAM)	$65.5_{\pm 2.3}$	$23.3_{\pm 4.2}$	$30.0_{\pm 5.5}$	$4.1_{\pm 1.4}$	$30.7_{\pm 2.1}$	$54.1_{\pm 4.1}$	$79.2_{\pm 2.9}$
	WholeSIGHT (Graph + Pseudo, NC)	**73.0**$_{\pm 3.1}$	**34.7**$_{\pm 1.2}$	**43.8**$_{\pm 5.3}$	**5.7**$_{\pm 0.4}$	**39.3**$_{\pm 1.4}$	**54.7**$_{\pm 4.6}$	**81.4**$_{\pm 5.2}$

Table 8.2: Classification and segmentation results on Radboud dataset. The best performances for using image-level supervision are highlighted in **bold**.

Annot.		per-class Dice				avg. Dice	GG wF1	ISUP κ^2
	Method	Benign	Grade3	Grade4	Grade5			
$\mathcal{C}$	WholeSIGHT (Multiplex, NC)	$91.6_{\pm 0.1}$	$64.3_{\pm 0.3}$	$65.9_{\pm 0.8}$	$38.2_{\pm 1.1}$	$65.0_{\pm 0.2}$	$61.7_{\pm 0.4}$	$76.3_{\pm 1.3}$
$\mathcal{IE}$	NIC [Tellez et al., 2021]	-	-	-	-	-	$35.1_{\pm 1.2}$	$45.0_{\pm 2.2}$
	CLAM [Lu et al., 2021]	-	-	-	-	-	$55.8_{\pm 1.1}$	$73.7_{\pm 1.7}$
	WholeSIGHT (Graph, GraphGrad-CAM)	$63.8_{\pm 2.3}$	$23.8_{\pm 3.8}$	$22.6_{\pm 1.9}$	$12.1_{\pm 0.7}$	$30.6_{\pm 1.0}$	**58.0**$_{\pm 0.8}$	**73.8**$_{\pm 1.6}$
	WholeSIGHT (Graph + Pseudo, NC)	**83.8**$_{\pm 0.6}$	**36.3**$_{\pm 1.1}$	**23.1**$_{\pm 2.3}$	**20.6**$_{\pm 0.3}$	**40.9**$_{\pm 0.5}$	**58.0**$_{\pm 0.8}$	**73.8**$_{\pm 1.6}$

compared to Radboud. This is due to the presence of more high-grade Gleason grade WSIs in Karolinska. This observation is substantiated by the confusion matrix of Gleason grade classification for the WholeSIGHT-DE method, as shown in Figure 8.3.

8.5.4 Generalization: performance, uncertainty, and calibration

Training setting: To study the generalization capability of WholeSIGHT, we propose a modified training setting. Specifically, we build a new training dataset that comprises of Karolinska and Radboud training WSIs. Thus, we create one large multi-source dataset by encompassing better sample variability and including more diagnostically challenging cases than their standalone counterparts. The trained models on this curated dataset are tested individually on the Karolinska and Radboud test WSIs, herein studying the *in-domain* performance. Further, we test on the entire Sicap dataset to analyze performance on *out-of-domain* WSIs.

Table 8.3: Classification results on Karolinska dataset. The best performances for using image-level supervision are highlighted in **bold**.

		GG wF1	ISUP κ^2
$\mathcal{IL}$	NIC [Tellez et al., 2021]	44.0$_{\pm 1.0}$	45.7$_{\pm 2.4}$
	CLAM [Lu et al., 2021]	66.3$_{\pm 1.0}$	**78.1**$_{\pm 1.5}$
	WholeSIGHT (Graph)	**67.1**$_{\pm 0.9}$	77.4$_{\pm 1.2}$

Table 8.4: Classification and segmentation results on Radboud, Karolinska, and Sicap datasets for models trained using both Radboud and Karolinska datasets.

Annot.		Radboud			Karolinska		Sicap		
	Method	avg. Dice	GG wF1	ISUP κ^2	GG wF1	ISUP κ^2	avg. Dice	GG wF1	ISUP κ^2
$\mathcal{C}$	WholeSIGHT (Multiplex, NC)	64.8$_{\pm 0.6}$	58.5$_{\pm 1.4}$	74.0$_{\pm 1.5}$	67.6$_{\pm 1.4}$	78.8$_{\pm 1.2}$	55.8$_{\pm 0.6}$	75.0$_{\pm 3.9}$	92.8$_{\pm 3.0}$
$\mathcal{IL}$	NIC[Tellez et al., 2021]	-	27.6$_{\pm 5.0}$	40.6$_{\pm 7.2}$	43.1$_{\pm 2.4}$	45.0$_{\pm 4.7}$	-	27.3$_{\pm 6.3}$	36.1$_{\pm 9.1}$
	CLAM[Lu et al., 2021]	-	**57.6**$_{\pm 2.3}$	**73.8**$_{\pm 2.3}$	65.5$_{\pm 1.3}$	77.3$_{\pm 2.8}$	-	56.4$_{\pm 2.7}$	75.0$_{\pm 7.5}$
	WholeSIGHT (Graph, Grad-CAM)	29.0$_{\pm 1.2}$	56.5$_{\pm 0.5}$	72.0$_{\pm 1.5}$	**68.1**$_{\pm 0.6}$	**77.4**$_{\pm 0.9}$	24.2$_{\pm 2.1}$	**64.2**$_{\pm 4.7}$	**86.9**$_{\pm 4.4}$
	WholeSIGHT (Graph + Pseudo, NC)	**46.0**$_{\pm 0.4}$	56.5$_{\pm 0.5}$	72.0$_{\pm 1.5}$	**68.1**$_{\pm 0.6}$	**77.4**$_{\pm 0.9}$	**41.6**$_{\pm 0.5}$	**64.2**$_{\pm 4.7}$	**86.9**$_{\pm 4.4}$
Bayes	WholeSIGHT-MCD	43.9$_{\pm 1.8}$	58.2$_{\pm 0.8}$	73.7$_{\pm 3.1}$	67.9$_{\pm 1.1}$	77.7$_{\pm 1.0}$	44.5$_{\pm 3.0}$	61.4$_{\pm 3.6}$	75.2$_{\pm 6.7}$
	WholeSIGHT-DE	46.3$_{\pm 0.2}$	60.6$_{\pm 0.6}$	76.5$_{\pm 0.7}$	68.6$_{\pm 0.4}$	78.1$_{\pm 0.6}$	46.6$_{\pm 1.7}$	66.0$_{\pm 1.5}$	84.5$_{\pm 1.2}$

Performance analysis: Table 8.4 compares the classification performance of WholeSIGHT, its Bayesian variants, CLAM, and NIC. For the Gleason grade weighted-F1 on the in-domain Karolinska and Radboud datasets, WholeSIGHT reaches a comparable performance to CLAM, and significantly outperforms NIC. Similar observations are prevailed for the ISUP κ^2 metric for both the in-domain datasets. However, the variances of Gleason grade weighted-F1 and ISUP κ^2 of the CLAM models are much higher than WholeSIGHT. For testing on the *out-of-domain* Sicap dataset, WholeSIGHT achieves significantly better Gleason grade weighted-F1 and ISUP κ^2 compared to competing CLAM and NIC. Even though the WholeSIGHT variance on Sicap is larger compared to Karolinska and Radboud, it remains significantly lower than CLAM and NIC.

WholeSIGHT-MCD performs comparable to WholeSIGHT, without highlighting a clear performance gain for any of the datasets. Further, the variances of WholeSIGHT-MCD are significantly higher than standalone WholeSIGHT. However, WholeSIGHT-DE shows a significant gain in classification and segmentation performances for all datasets. The deep ensemble-based methods result in clear advantages over MC-dropout-based methods, which are consistent with the observations by [Thagaard et al., 2020]. Noticeably, the gain in performances are higher on the *out-of-domain* dataset, compared to *in-domain* datasets. This finding corroborates the conclusion of [Gustafsson et al., 2020] which showed that deep ensemble improves generalization to unseen cohorts. Overall, WholeSIGHT-DE is the best performer across all datasets for all the evaluation metrics. Figure 8.3 presents

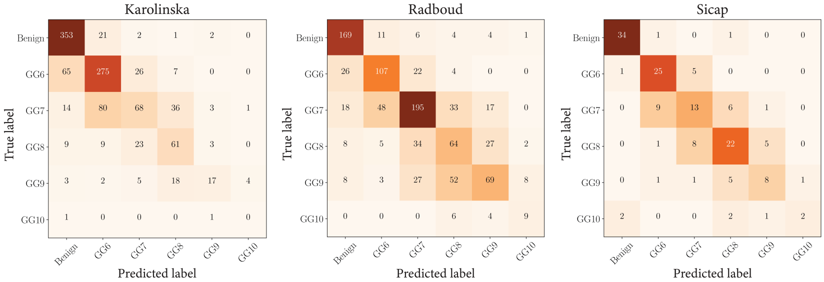


Figure 8.3: Confusion matrix of Gleason grade classification for the WholeSIGHT-DE method on Karolinska, Radboud, and Sicap datasets.

the Gleason grading confusion matrices of WholeSIGHT-DE on the three considered datasets. It can be observed that the most misclassifications lie close to the diagonal. Majority of the confusion occurs between GG6 and GG7, i. e., $\mathrm{GG}(3+3)$ versus $\mathrm{GG}(3+4)$ and $\mathrm{GG}(4+3)$. Such ambiguity is prevalent among pathologists, as presented in [Ozkan et al., 2016; Salmo, 2015]. Further confusion matrices for Gleason grading, ISUP grading, primary classification, and secondary classification are presented in Figure 8.9

Table 8.4 and Figure 8.4 present the segmentation and its generalizability assessment for WholeSIGHT, and its Bayesian variants. Both WholeSIGHT-MCD and WholeSIGHT-DE significantly outperform WholeSIGHT by improving the mean Dice by +2.9% and +5.0%, respectively, on Sicap. In consistence with the classification results, WholeSIGHT-DE is the best performer in terms of class-wise and aggregated Dice, and systematically reduces the variance in performance. Benign regions, being the most common class, results the highest Dice. Whereas, the less encountered Gleason patterns, i. e., G3, G4, G5, have comparatively lower Dice. This drop primarily occurs due to the ambiguities among the cancerous patterns and false positive benign regions.

Uncertainty estimate analysis: Figure 8.5 presents the classification uncertainties of WholeSIGHT, and its Bayesian variants, in terms of NLL (Figure 8.5(a)) and Brier score (Figure 8.5(b)), on Karolinska, Radboud, and Sicap. The Bayesian methods render a significantly lower NLL than WholeSIGHT across all datasets, for primary, secondary, and Gleason grading (P+S). The relative gain of WholeSIGHT-DE is +34.1% for P+S on Karolinska, +44.71% on Radboud, and +51.59% on Sicap. Interestingly, the gain is higher for the *out-of-domain* dataset, showing that Bayesian models, in particular deep ensembles, provide better uncertainty estimates. These observations are also consistent for the Brier score. WholeSIGHT-DE consistently outperforms WholeSIGHT, with a relative gain of +13.37% on Karolinska, +15.45% on Radboud, and +21.87% on Sicap. Noticeably, the NLL and Brier scores are consistently higher for predicting the secondary Gleason patterns compared to the primary patterns. This resonates with the fact that identifying secondary patterns is a harder task with higher ambiguity.

A similar analysis for quantifying the uncertainty in segmentation is shown in Figure 8.4(b). A relative gain of +21.49% and +1.44% in NLL and Brier score, respectively, is achieved by WholeSIGHT-MCD on average Dice score. Though WholeSIGHT-DE outperforms WholeSIGHT-MCD in terms of NLL, it performs inferior in terms of Brier.

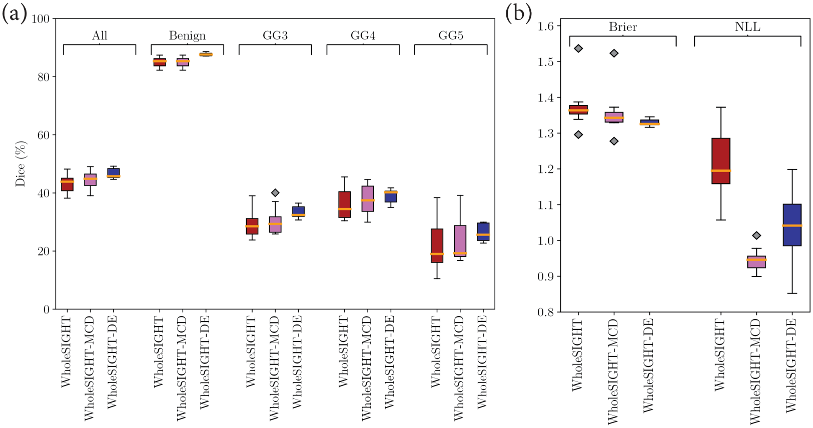


Figure 8.4: (a) Average and per-class Dice scores of WHOLESIGHT, WHOLESIGHT-MCD, and WHOLESIGHT-DE on Sicap dataset. (b) Uncertainty analysis of these methods in terms of Brier and NLL metrics on Sicap dataset.

Model calibration analysis: A model with a good uncertainty estimate should be well-calibrated, i. e., the model confidence should be close to the underlying model performance. Figure 8.5(c) presents the reliability diagrams of the primary classification head on Karolinska and Radboud datasets. WHOLESIGHT-DE shows significantly better calibration than WHOLESIGHT-MCD and WHOLESIGHT-DE in accordance with the uncertainty estimate analysis. However, we observe that WHOLESIGHT-DE remains over-confident as the model accuracy (in orange) is lower than the expected optimal calibration (in blue). Figure 8.6 shows a detailed analysis of model calibration. We observe that even if not perfectly aligned, the gap between model accuracy and model confidence, denoted as dashed vertical lines in black, is reduced for the Bayesian methods. This gain is quantified by computing the ECE. For instance, the Radboud secondary classification head calibration is improved by +27.7% for WHOLESIGHT-MCD and +46.4% for WHOLESIGHT-DE.

8.5.5 Qualitative analysis

We qualitatively analyze the results of our proposed WHOLESIGHT method by (1) visualizing overlaid segmentation masks on WSIs, (2) analysing the t-distributed stochastic neighbor (t-SNE) [Van der Maaten et al., 2008] node embeddings, and (3) correlating the segmentation outputs with pathological reasonings.

Visualizing WholeSIGHT segmentation masks: Figure 8.7 demonstrates segmentation predictions obtained with WHOLESIGHT and its variant, WHOLESIGHT(Multiplex, NC), on Sicap dataset. We can observe that WHOLESIGHT correctly delineates the cancerous regions in the WSIs. Zooming into different regions conclude that the tissue regions of TG, i. e., the nodes of TG, (outlined in black in Figure 8.7) encode meaningful units of *homogeneous* tissue. It substantiates the relevance of using TG representations for segmenting the tissue regions into Gleason patterns. We further notice that WHOLESIGHT, in a few cases, predicts benign regions adjacent to cancerous patterns as cancerous. For example, the benign region, primarily consisting of stroma, in Figure 8.7(c) is predicted as G5. We argue that these false positive detections do not inhibit the applicability of the method, as neighboring cancerous regions are correctly detected. In a few other cases, WholeSIGHT correctly detects cancerous regions that are missed in the ground truth annotations. For instance, in Figure 8.7(b), the missing G4 region in the upper part of the WSI is correctly identified by WholeSIGHT.

On comparing WholeSIGHT with WholeSIGHT(Multiplex, NC) in Figure 8.7(a), we see that several false positives are removed, thereby offering more accurate segmentation. However, the improvements by WholeSIGHT(Multiplex, NC) are achieved at the cost of training with pixel-level annotations, that are hardly available in real-world practice. Thus, WholeSIGHT appears to be an appealing compromise between segmentation performance and annotation requirement for Gleason pattern segmentation.

Visualizing tissue-level t-SNE feature space: A t-SNE visualization of the learned tissue-level embeddings is demonstrated in Figure 8.8 for Sicap dataset. The t-SNE algorithm projects the GNN node embeddings onto a two-dimensional feature space, allowing to analyse the connection between node embeddings and the Gleason pattern distribution.

Figure 8.8(a) displays the t-SNE feature space for the *correctly* classified nodes, which highlights demarcated clusters for each Gleason pattern. The large cluster of benign nodes indicate the diversity of the benign category. Several patches from each cluster are shown in Figure 8.8(d). We can observe the reduced nuclei differentiation across the patches from benign to grade 5. Further, Figure 8.8(b) and (c) display the t-SNE feature space for the misclassified nodes. Specifically, Figure 8.8(b) represents the ground truth node labels, and Figure 8.8(c) the predicted node labels. Different embedding locations are further selected and highlighted by different colored rectangles and put in relation with corresponding patches to indicate the inter-class ambiguities, as demonstrated in Figure 8.8(e). For example, the first row in Figure 8.8(e) showcases patches which are actually benign but are predicted as Gleason pattern-3. We can visually compare these patches with the Gleason pattern-3 patches in the third row of Figure 8.8(d). Similar ambiguities between other pairs of Gleason patterns are also included in Figure 8.8(e).

Interpreting model outcomes via predicted segmentations: Predicted segmentations provide human-understandable *interpretability* maps. For researchers, the segmentations allow to, (1) identify morphological patterns responsible for the WSI classification, (2) analyse failure cases by inspecting the pixel-level predictions, and ultimately (3) better understand the model behavior towards biomarker discovery. For pathologists, they assist to, (1) put in relation the predicted WSI-level Gleason scores and the highlighted pixel-level Gleason patterns, (2) confirm that the morphology of the identified cancerous regions align with the pre-established diagnosis criteria.

Additionally, in the perspective of developing AI-assisted human-in-the-loop tools, a Gleason grading system that can simultaneously *classify* and *segment* WSIs is closer to the latest pathological standards. Indeed, recent revisions of the Gleason grading system [Epstein et al., 2016] emphasized on the importance of reporting the percentage of each grade for better patient stratification and treatment selection [Cheng et al., 2007; Huang et al., 2014; Choy et al., 2016; Sharma et al., 2018]. These percentages can be trivially derived from the predicted segmentation maps by counting the number of pixels belonging to each pattern. Naturally, such information is not available in mere WSI classification systems. Reporting per-grade percentage is particularly important in ambiguous and borderline cases. For instance, consider two patients with Gleason score 3+4. When a small percentage of pattern-4 is present, e. g., 10%, the case can be considered as an intermediate risk cancer where active patient surveillance is enough [Amin et al., 2014]. However, a larger secondary pattern may require specific treatments. Reporting percentages of each grade allows to easily discriminate these two scenarios. Similarly, consider a Gleason score 4+3 with a small secondary Gleason pattern, e. g., 90% and 10% area for primary and secondary patterns, respectively. This case will be scored as 4+3, even though it is close to a score 4+4, which would lead to a different treatment protocol. By explicitly reporting the Gleason pattern percentages, such corner-cases can be avoided.

8.6 Conclusion

Accurate delineation of patterns in a giga-pixel sized whole-slide histopathology image by using a deep learning method typically demands pixel-level annotations. However, such exhaustive annotations are often impossible to acquire in a real-world scenario due to the bottlenecks in time, effort, and expense. Nonetheless, the semantic segmentation of diagnostically relevant patterns is crucial for disease diagnosis and treatment selection. To this end, we have proposed a novel weakly-supervised semantic segmentation method, WholeSIGHT, that can segment the relevant patterns of interest in histopathology images by leveraging only image-level supervision. To the best of our knowledge, WholeSIGHT is the first weakly-supervised semantic segmentation method that can operate in an end-to-end manner on histopathology images of arbitrary shape and size. First, WholeSIGHT transforms a histopathology image into a tissue-graph representation, where the nodes and edges of the graph denote tissue regions and tissue-to-tissue interactions. Second, the method employs a graph neural network to construct inter-tissue relationship-aware representations for the tissue regions. These contextualized representations are further used to classify the tissue-graph. Subsequently, pseudo-labels are generated for the tissue regions via a graph-feature-attribution technique, which enables the classification of the tissue regions and segments the input histopathology image. We evaluated our proposed method on three publicly available prostate needle biopsy datasets for Gleason grade classification and the delineation of different Gleason patterns in the biopsies. On comparing with several state-of-the-art methods, we demonstrated the classification and segmentation superiority of our proposed WholeSIGHT method. Further, we conducted extensive experimentation to assess the generalizability of WholeSIGHT on *out-of-domain* histopathology datasets. Further, we proposed a Bayesian extension of WholeSIGHT, i. e., WholeSIGHT-DE, to enhance the generalizability of the method to images from different data sources. The generalizability is quantified in terms of classification and segmentation performance metrics, uncertainty estimation, and model calibration analysis. Notably, the proposed WholeSIGHT method can utilize both image-level and pixel-level supervision to simultaneously perform image classification and segmentation tasks. Hence, WholeSIGHT performance on both tasks can be enhanced in presence of pixel-level partial annotations from pathologists. Though we have evaluated our method for H&E stained prostate cancer needle biopsies, the technology is easily extendable to other tissue types, e. g., breast, colon, and lungs, or imaging techniques, e. g., tissue microarrays, and resection biopsies, or image modalities, e. g., other staining types in histopathology, multiplexed histopathology images, etc., or domains, e. g., natural images, hyperspectral images, satellite images, and other medical imaging data.

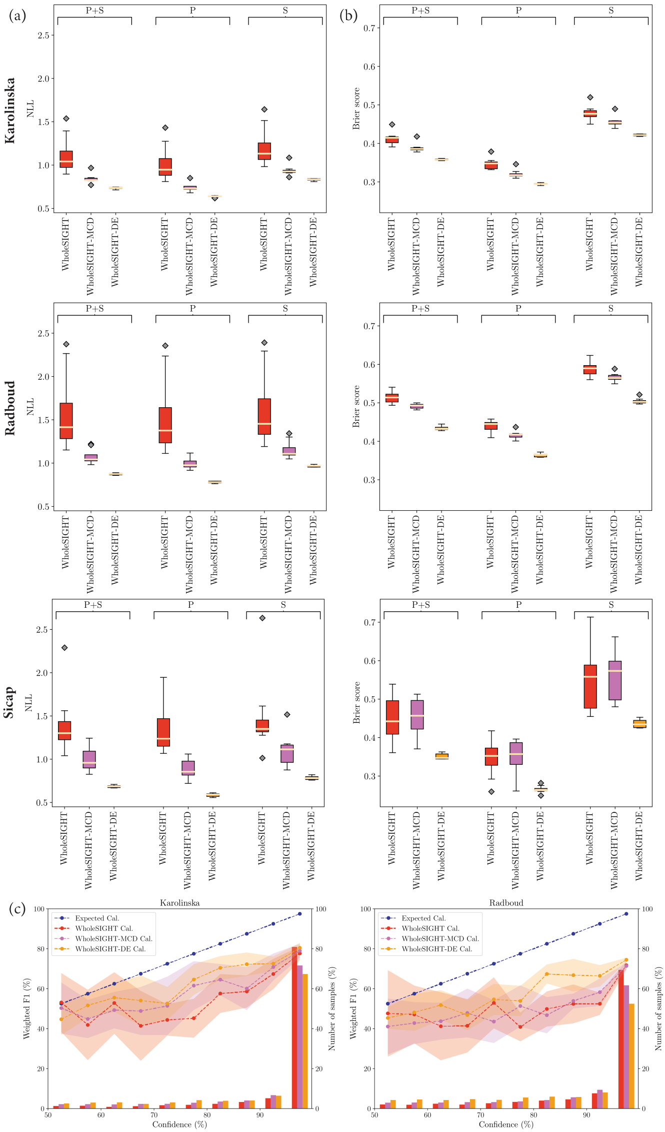


Figure 8.5: Uncertainty analysis of the WholeSIGHT, WholeSIGHT-MCD, and WholeSIGHT-DE models. (a) and (b) present Brier scores (lower is better) and NLL (lower is better), respectively, on Karolinska, Radboud and Sicap. (c) Reliability diagrams on Karolinska and Radboud test sets for the primary Gleason classification head. The expected calibration (blue) denotes a perfectly calibrated model. Calibrations of WholeSIGHT, WholeSIGHT-MCD, and WholeSIGHT-DE, and the fraction of samples in each confidence bin is shown in red, purple and orange, respectively.

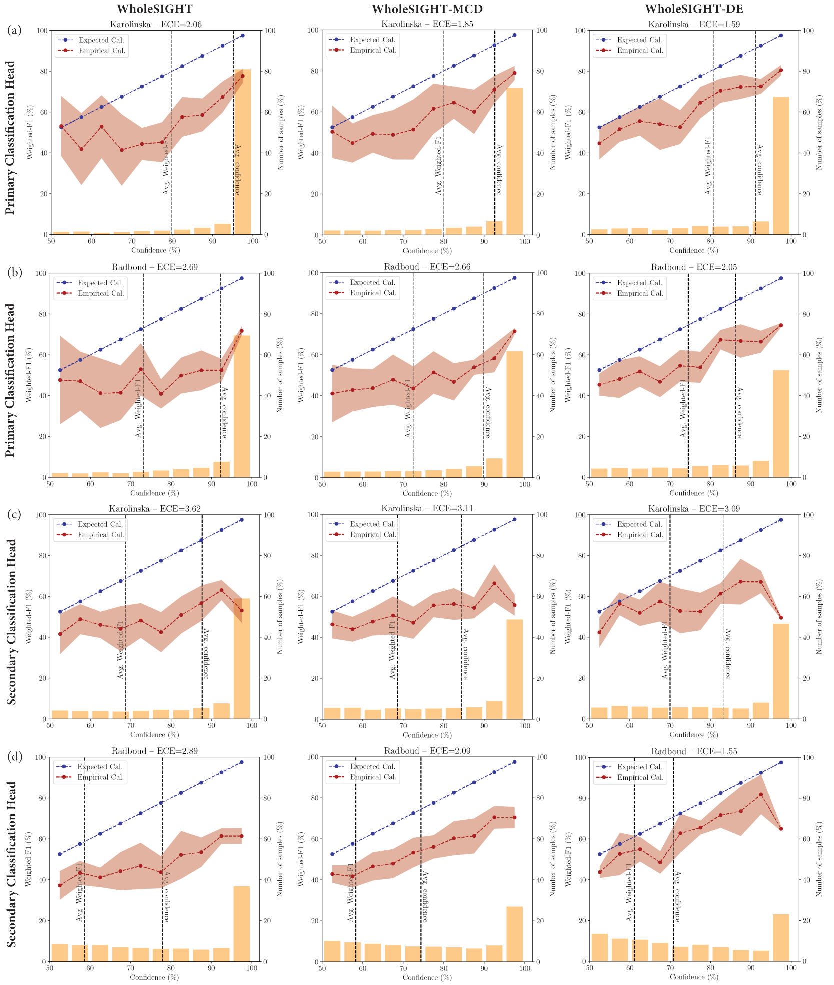


Figure 8.6: Reliability diagrams of WholeSIGHT, WholeSIGHT-MCD, and WholeSIGHT-DE tested on Karolinska and Radboud datasets for the primary and secondary Gleason classification heads. The expected calibration (blue) highlights a perfectly calibrated network. The observed network calibrations are shown in red. The fraction of samples in each confidence bin is shown in orange. Proximity of average confidence to average weighted-F1 indicates a better calibration. Note that ECE is the integral between the expected calibration (blue) and the network calibration (red).

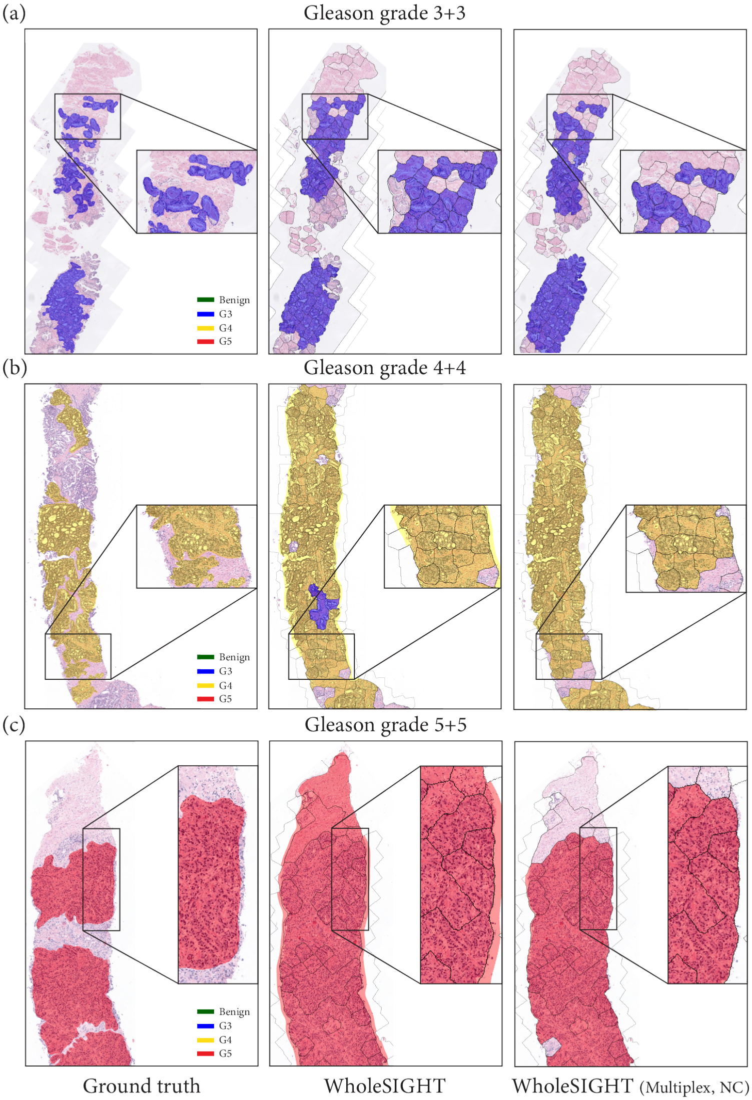


Figure 8.7: Samples of segmentation maps for the Sicap dataset. The columns indicate ground truth (left), the proposed WholeSIGHT (middle), and WholeSIGHT(Multiplex, NC) (right) segmentation maps, respectively. (a), (b), and (c) presents examples from GG (3+3), GG (4+4), and GG (5+5), respectively. The tissue regions, i. e., TG nodes, are highlighted in black overlay. For better visualization, the benign areas are not highlighted on the segmentation maps.

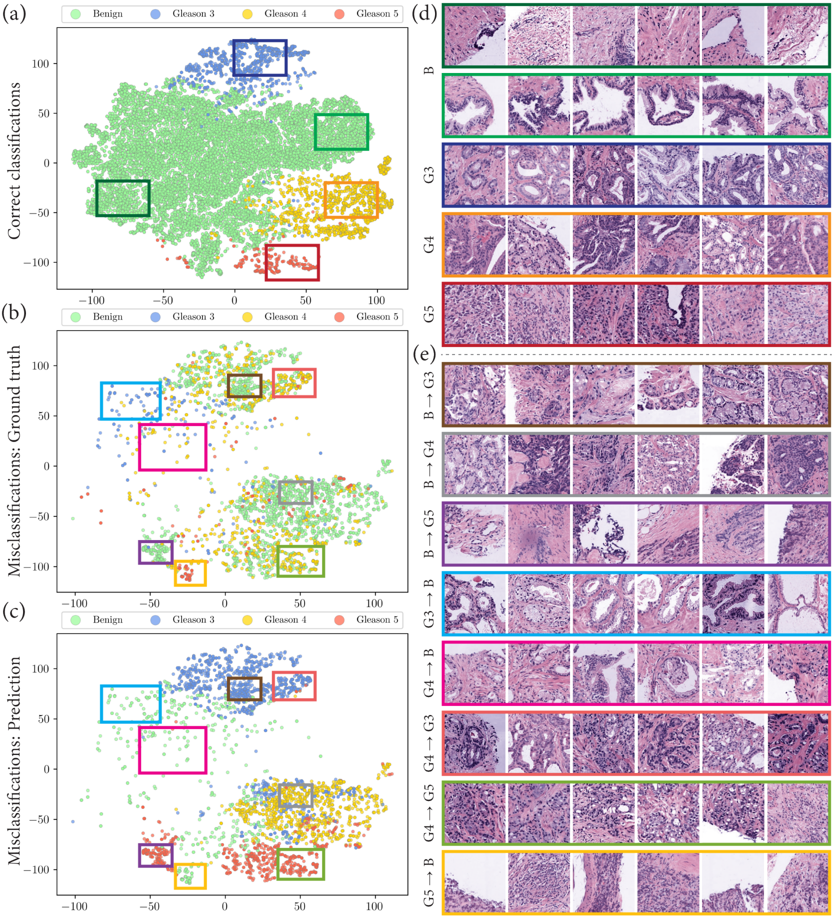


Figure 8.8: t-SNE visualization of node-level feature representations, extracted by WholeSIGHT, and example patches corresponding to several regions on the two-dimensional t-SNE feature space for Sicap dataset. (a) t-SNE visualization of correctly classified nodes. (b) and (c) display the t-SNE visualization of misclassified nodes, where (b) and (c) highlight the ground truth and predicted class labels of the nodes, respectively. (d) and (e) display square patches of size 224 × 224 at 10× magnification cropped around the node centroids selected from different regions on the t-SNE embedding space. (d) and (e) present the correctly and incorrectly classified patches, respectively. The labels of the patches in (e) are formatted as $Y \rightarrow \hat{Y}$, where Y and $\hat{Y}$ denote the ground truth and the predicted class labels. The colored rectangles around the patches in (d) and (e) correspond to respective colored rectangles in (a), (b), and (c).

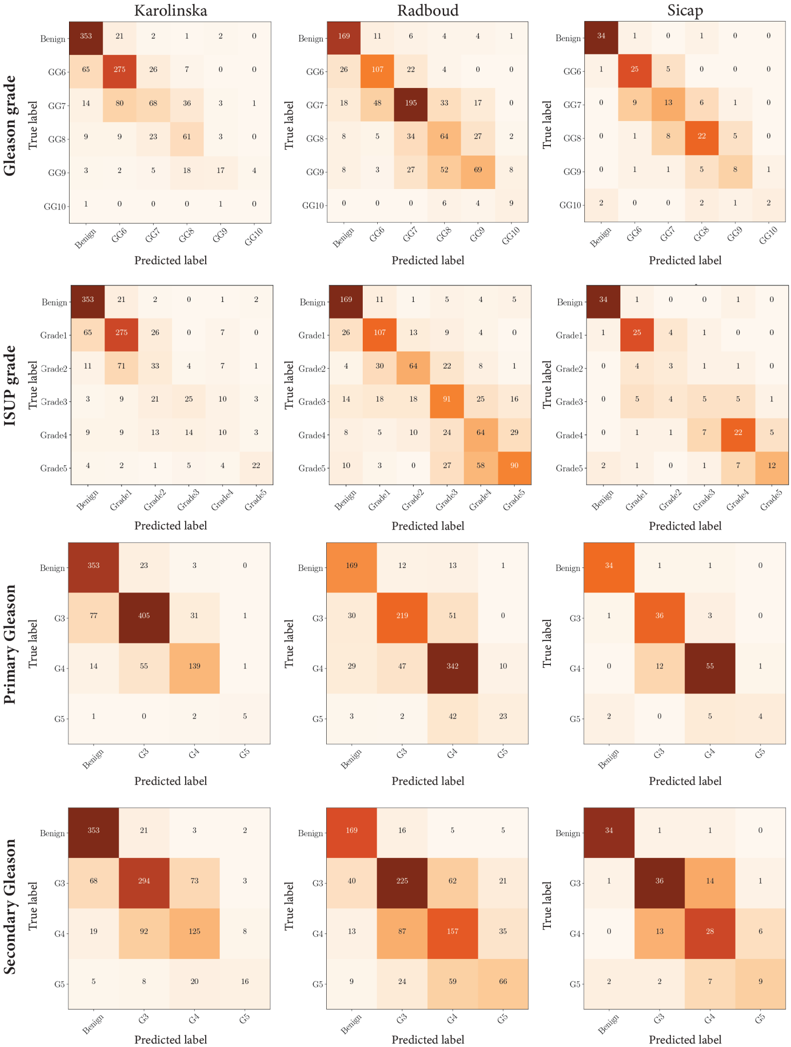


Figure 8.9: Gleason grade, ISUP grade, primary-, and secondary-classification confusion matrices obtained for WholeSIGHT-DE on Karolinska, Radboud, and Sicap datasets.

9

HistoCartography: A Toolkit for Graph Analytics in Digital Pathology

Advances in entity-graph analysis of histopathology images have brought in a new paradigm to describe tissue composition, and learn the tissue structure-to-function relationship. Entity-graphs offer flexible and scalable representations to characterize tissue organization, while allowing the incorporation of prior pathological knowledge to further support model explainability. However, their analysis requires prerequisites for image-to-graph translation and knowledge of state-of-the-art algorithms applied to graph-structured data, which can potentially hinder their adoption. In this work, we aim to alleviate these issues by developing HistoCartography, a standardized python API with necessary preprocessing, machine learning and explainability tools to facilitate graph-analytics in computational pathology. Further, we have benchmarked the computational time and performance on multiple datasets across different imaging types and histopathology tasks to highlight the applicability of the API for building computational pathology workflows. HistoCartography is available at `https://github.com/histocartography/histocartography`.

9.1 Introduction

Recent advancements in tissue-slide digitization have paved way for enhancing storage, sharing capabilities, and computer-aided inspection by leveraging DL. Most DL approaches analyze tissue images in three steps, namely patch generation, patch-level feature extraction, and feature aggregation to produce image-level embeddings for downstream pathology tasks. However, they suffer from several limitations, (1) the trade-off between operational resolution and adequate context per-patch, (2) the aggregation is often sub-optimal, (3) comprehensive modeling of tissue composition is missing, and (4) the lack of model transparency raises barriers to deployment in real life.

To circumvent these limitations, entity-graphs are proposed by [Demir et al., 2004] where the nodes and edges of the graphs denote tissue entities and their interactions, respectively. Entity-graphs, followed by GNNs-based processing, have recently gained popularity in addressing various pathology tasks [Zhou et al., 2019a; Chen et al., 2020a; Pati et al., 2022;

This chapter has been published as: Guillaume Jaume†, Pushpak Pati†, Valenin Anklin, Antonio Foncubierta-Rodriguez, Maria Gabrani "HistoCartography: A Toolkit for Graph Analytics in Digital Pathology", In: Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop on Computational Pathology (COMPAY), vol. 156, pp. 117-128, Strasbourg, France, Sep 2021, [arXiv]

Anklin et al., 2021; Jaume et al., 2021b]. The entities can be biologically-defined, e. g., nuclei, tissue regions, glands [Zhou et al., 2019a; Pati et al., 2022; Anklin et al., 2021], or can be patches [Adnan et al., 2020; Aygüneş et al., 2020]. The entity-graphs enable to simultaneously capture local entity environment and global tissue composition. They can seamlessly scale to arbitrary tissue dimensions by incorporating arbitrary number of entities and interactions, thus offering an alternate to MIL [Campanella et al., 2019; Lu et al., 2021]. The entity-graphs also enable to selectively operate on diagnostically relevant entities, instead of analyzing the entire tissue [Tellez et al., 2021; Shaban et al., 2020]. Furthermore, when the entities depict biological units, such as nuclei, and glands, the analysis allows pathologists to directly comprehend and reason with the outcomes [Jaume et al., 2020; Jaume et al., 2021b]. However, constructing an entity-graph based pathological workflow demands several prerequisites, such as entity detection, entity encoding, and constructing the graph topology, alongside standard preprocessing, such as stain normalization, and tissue detection. Additionally, the workflow requires to utilize the recent advancements in DL for processing graph-structured data. All these may inhibit the adoption of entity-graphs in computational pathology. In addition, the lack of a standardized framework with the aforementioned functionalities urge the researchers to reinvent the wheel, which is cumbersome, time-consuming, hampers reproducibility, and requires a wide range of technical acumen.

To overcome these constraints, we present HistoCartography, a novel open-source python library that facilitates graph-analytics in computational pathology. Specifically our contributions are: (1) a standardized, unit-tested python library that unifies a set of histology image manipulation tools, entity-graph builders, GNN models, and model explainability tools, (2) a benchmark assessment of performance and scalability on classification and segmentation tasks in pathology, (3) a comprehensive overview of graph representation and modeling in histology, and (4) a review of extant libraries for histological image analysis.

9.2 Related work

9.2.1 Graphs in computational pathology

Entity-graphs are proposed to realize the tissue composition-to-functionality relationship in terms of the phenotypical and structural characteristics of tissue. The entities can be nuclei [Demir et al., 2004; Zhou et al., 2019a; Wang et al., 2019a; Chen et al., 2020a; Pati et al., 2022], tissue regions [Pati et al., 2022], patches [Anand et al., 2019; Adnan et al., 2020; Aygüneş et al., 2020; Zhao et al., 2020b; Li et al., 2018c; Levy et al., 2021], etc. Typically nodes include handcrafted or DL features to characterize the entities, and the topology can depict the spatial or semantic relationship among the entities, e. g., k-NN, region adjacency, or probabilistic models. The graphs can be processed using classic ML [Sharma et al., 2016; Sharma et al., 2017a] or GNNs to outperform state-of-the-art CNN-based approaches for several pathology tasks across multiple organs [García-Arteaga et al., 2017; Zhou et al., 2019a; Zhao et al., 2020b; Adnan et al., 2020; Pati et al., 2022; Studer et al., 2021; Anklin et al., 2021]. Interestingly, when the graph-nodes depict biological entities, e. g., nuclei, tissue regions, the entity-graphs combined with feature attribution techniques can provide pathologist-friendly interpretations [Zhou et al., 2019a; Jaume et al., 2020; Sureka et al., 2020] and explanations [Jaume et al., 2021b], unlike pixelated blurry saliency maps. A detailed review of graphs in computational pathology is presented by [Ahmedt-Aristizabal et al., 2021].

9.2.2 Extant libraries in computational pathology

Several open-source libraries facilitate the development of computational pathology workflows. Most of them include helper functions to perform standard preprocessing and visualization. HistoLab [Marcolini et al., 2020] includes WSI-level tissue detection and tile extraction modules. Syntax [Byfield et al., 2020] provides the same features with abstraction where modules can be stacked and run in a pre-defined pipeline. StainTools [Byfield et al., 2019] provides tools for stain normalization and augmentation. HistomicsTK [Beezley et al., 2021] enables to perform tissue detection, object detection and segmentation, image filtering, stain normalization and deconvolution, and handcrafted feature extraction. Further, HistomicsTK allows nuclei segmentation and classification using classical ML approaches. It also provides a User Interface (UI) to run containerized modules and pipelines. Though HistomicsTK includes valuable functionalities, it caters limited DL tools. Similarly, OpenSlide [Gilbert et al., 2020] provides a UI to read and visualize histology images that supports most of the WSI formats. Finally, QuPath [Bankhead et al., 2021] offers a UI that allows to read, visualize and annotate WSIs. It also includes tools to perform stain normalization, nuclei and tissue detection, and implement basic ML models. However, QuPath is not a python API, which makes it difficult to integrate into existing workflow and DL frameworks, e. g., PyTorch, Tensorflow. Most importantly, none of the frameworks provide graph-related helpers. With the advent of graph-techniques as a new paradigm for analyzing histology images, a standardized library is desired for reinforcing the development.

9.3 HistoCartography: Graph analytics tool for pathology

In this section, we highlight the core modules of HistoCartography, (1) *Preprocessing* module: a set of histology image processing tools and entity-graph builders, (2) *ML* module: helpers to learn from entity-graphs, (3) *Explainability* module: a set of GNN model interpretability tools. List of module-wise functionalities are summarized in Table 9.1. To facilitate integration and reduce boilerplate code, HistoCartography includes a pipeline runner which allows to pre-define pipeline steps along with loading and saving utilities.

9.3.1 Preprocessing module

Stain normalization: Variation in H&E staining protocols for tissue specimens induces appearance variability that adversely impacts computational methods [Tellez et al., 2019]. To alleviate these variations, HistoCartography implements two popular normalization algorithms proposed by [Macenko et al., 2009] and [Vahadane et al., 2016], similar to StainTools and HistomicsTK, which supports both reference-based and reference-free normalization, i. e., with manual stain vectors. Figure 9.1 highlights a sample normalization output using our API.

Tissue Detection: A WSI usually includes significant non-tissue region. Identifying the tissue regions can confine the analysis and reduce computational effort. The tissue detector in HistoCartography iteratively applies Gaussian smoothing and Otsu thresholding

Table 9.1: Overview of HistoCartography functionalities, with the i/o, CPU and GPU compatibility, and availability in extant libraries for individual module. `I`, `M`, `X`, `G`, `P` and `S` denote an image (np.array [Harris et al., 2020]), a mask (np.array), features (torch.Tensor [Paszke et al., 2019]), a graph (DGLGraph [Wang et al., 2019b]), predictions (torch.Tensor) and importance scores (torch.Tensor), respectively.

Module	Function	Input	Output	Existing	CPU	GPU
Preprocessing	Vahadane Stain Norm	`I`	`I`	✓	✓	✗
	Macenko Stain Norm	`I`	`I`	✓	✓	✗
	Tissue Mask Detection	`I`	`M`	✓	✓	✗
	Nuclei Detection	`I`	`M`	✓	✓	✓
	Nuclei Concepts	`I`, `M`	`M`	✓	✓	✗
	Tissue Component Detection	`I`	`M`	✗	✓	✗
	Deep Feature Extraction	`I`, `M`	`X`	✗	✓	✓
	Feature Cube Extraction	`I`	`X`	✗	✓	✓
	k-NN Graph Building	`X`, `M`	`G`	✗	✓	✗
	RAG Graph Building	`X`, `M`	`G`	✗	✓	✗
ML	Cell-Graph Model	`G`	`P`	✗	✓	✓
	Tissue-Graph Model	`G`	`P`	✗	✓	✓
	HACT Model	`G`, `G`, `X`	`P`	✗	✓	✓
Explainers	GnnExplainer	`G`	`S`	✗	✓	✓
	GraphGrad-CAM	`G`	`S`	✗	✓	✓
	GraphGrad-CAM++	`G`	`S`	✗	✓	✓
	GraphLRP	`G`	`S`	✗	✓	✓

until the mean of non-tissue pixels is below a threshold. This module is common across Histolab, Syntax, HistomicsTK and QuPath.

Nuclei detection: This module enables to segment and locate nuclei in H&E images. Though it is well-studied in computational pathology, only a few public implementations are available. For instance, QuPath allows to detect nuclei but requires model training and fine-tuning. While providing flexibility, the module includes only elementary ML methods. HistoCartography integrates two checkpoints from the state-of-the-art HoVerNet model [Graham et al., 2019a] trained on PanNuke [Gamper et al., 2020] and MoNuSac [Verma et al., 2021] datasets for nuclei segmentation and classification.

Tissue Component Detection: HistoCartography includes an unsupervised superpixel-based approach to segment tissue regions. First, the tissue is oversegmented into homogeneous superpixels using SLIC [Achanta et al., 2012]. Then, neighboring superpixels are hierarchically merged using color similarity to denote meaningful tissue regions, e. g., epithelium and stroma regions. Superpixels depicting tissue regions are used by [Bejnordi et al., 2015; Pati et al., 2020; Pati et al., 2022].

Feature Extraction: HistoCartography includes two types of feature extractors, i. e., handcrafted- and CNN-based, to encode the entity characteristics. The handcrafted feature extractor computes entity-level morphological and topological properties. Morphological features capture the shape and size, e. g., entity area, eccentricity, and perimeter, and the texture captures chromaticity using the gray-level co-occurrence matrix. Topological features capture the local entity distribution using k-NN entity density estimation. A

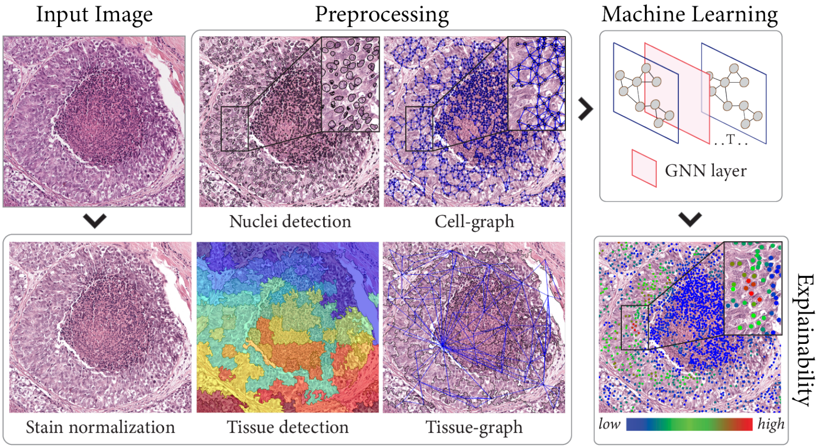


Figure 9.1: Overview of HistoCartography modules and functionalities.

comprehensive list is provided in the Appendix. Handcrafted features can be used for training DL algorithms [Demir et al., 2004; Zhou et al., 2019a; Pati et al., 2020; Studer et al., 2021], or concept-based post-hoc explainability [Jaume et al., 2021b].

The deep feature extractor allows to extract CNN features by using any pre-trained deep architecture, e. g., ResNet, MobileNet, embedded in torchvision [Marcel et al., 2010]. The module intakes patches centered around the entity and extracts features from the penultimate layer of the architectures. If the entity is larger than the specified patch size, then multiple patches within the entity, w/ or w/o overlapping, are processed, and the final feature is computed as the mean of the patch-level deep features, as in [Chen et al., 2020a; Pati et al., 2020; Pati et al., 2022]. Deep features can alternatively be extracted from WSI to build a feature-cube as suggested by [Shaban et al., 2020; Tellez et al., 2021].

Graph builders: HistoCartography presents two graph builders, i. e., the thresholded k-NN and the RAG. The k-NN graph builder defines the graph topology by connecting each entity to its k-closest neighbors. Connections between distant entities beyond a threshold can be pruned to have spatial sparsity in the graph. We recommend this builder to connect single entities, e. g., nuclei, glands. The RAG builder connects entities using spatial adjacency, i. e., entities sharing a common boundary. It builds a sound topology when dealing with dense segmentation maps, e. g., tissue regions. Figure 9.1 presents samples of cell- and tissue-graphs. Further, the module fuses the node features and the topological distribution to render a Deep Graph Library (DGL) graph for an image.

9.3.2 Graph machine learning module

HistoCartography includes a set of DL models, based on a GNN backbone to learn from graph-structured tissue representations. It includes two state-of-the-art GNN layers, i. e., GIN [Xu et al., 2019b] and PNA [Corso et al., 2020]. PNA proves to outperform GIN provided more computational resources [Dwivedi et al., 2020]. HistoCartography defines cell- and tissue-graph models, which are GNN-based abstractions to learn from biological entity-graphs. They offer efficient [Pati et al., 2022], scalable [Anklin et al., 2021] and explainable [Jaume et al., 2020; Jaume et al., 2021b] approaches to analyze histology images. Further, the library includes models to jointly represent and learn from cell- and tissue-graphs [Pati et al., 2022]. The models in HistoCartography are organized such that they can be adapted to various GNN backbones, tasks (e. g., regression, clustering, classification, segmentation), organs, and entity-types. These models provide the blueprints to accelerate the development of graph-based models in computational pathology. All the graph modules are implemented using DGL [Wang et al., 2019b], a state-of-the-art library for GNNs built around PyTorch.

9.3.3 Explainability module

HistoCartography includes four post-hoc feature attribution graph explainers, that can generate node-level saliency maps to highlight the node-wise contribution towards an output task. Namely, the library includes two gradient-based explainers (GraphGrad-CAM [Selvaraju et al., 2017; Pope et al., 2019] and GraphGrad-CAM++ [Chattopadhay et al., 2018; Jaume et al., 2021b]), a node pruning-based explainer (GnnExplainer [Ying et al., 2019]), and a layer-wise relevance propagation explainer (GraphLRP [Schwarzenberg et al., 2019]). The saliency map can be visualized by overlaying the node importances on the input image (see Figure 9.2). Alternatively, entities with high importances can be extracted and studied independently to assess their relevance [Jaume et al., 2021b].

9.4 Benchmarking HistoCartography

We benchmark HistoCartography in terms of run-time and performance for various histopathology tasks, i. e., stain normalization, tissue detection, tumor classification and segmentation, on images of various sizes. The CPU and GPU compatible modules are assessed on a single-core POWER8 processor and a NVIDIA P100 GPU, respectively.

9.4.1 Computational time

Analyzing the computational time for processing a histology image is imperative. We thoroughly analyze the run-time of HistoCartography functionalities on a set of RoIs and WSIs (Table 9.2). The preprocessing modules are observed to be the most time-consuming. For instance, Vahadane stain normalization can take up to 3 minutes to process a $11'000 \times 11'000$ image, whereas Macenko method is 2× faster for competitive result. The implementations are computationally similar to Histolab and StainTools, and scale linearly w.r.to image size. The cell- and tissue-graph construction take 2.5 and 4.1 seconds, respectively, for a 1000×1000 image with the following parameters. Nuclei detection is performed on patches of size 256×256 with an overlap of 164 pixels. Nuclei features are extracted from 72×72 patches centered around the nuclei, that are resized to 224×224 and processed by ResNet34 pretrained on ImageNet [Deng et al., 2009]. Finally, thresholded k-NN topology is built with $k = 5$ and a threshold distance of 50 pixels. For the tissue-graph, SLIC is used to extract 400 superpixels per image, that are subsequently merged to provide the tissue components. Tissue features are also extracted using ResNet34 with 144×144 size patches that are resized to 224×224. The graph buildings can be further optimized as per the task by downsampling the input image, reducing the patch overlap, or by using a lighter feature extractor. For extracting the

Table 9.2: Run time analysis of HistoCartography core functionalities (in seconds).

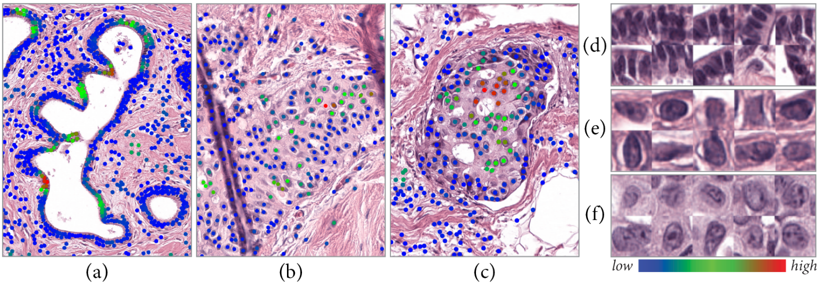
Mod.		Function / Image type	Tumor RoI			WSI		
			1000^2	2500^2	5000^2	5000^2	7500^2	11000^2
Preprocessing	Standard	Vahadane Normalization	1.77	6.46	29.03	30.67	68.27	186.10
		Macenko Normalization	0.80	2.86	11.19	15.98	32.37	81.72
		Tissue Mast Detection	-	-	-	1.04	2.11	8.09
		Feature Cube Extraction	0.24	1.61	5.92	6.27	11.97	29.79
	CG	Nuclei Detection	3.03	12.93	47.66	-	-	-
		Nuclei Concept Extraction	2.95	6.52	27.94	-	-	-
		Deep Nuclei Feature Extraction	0.010	0.30	1.28	-	-	-
		k-NN Graph Building	0.06	0.20	1.35	-	-	-
	TG	Super-pixel Detection	3.32	17.84	68.99	31.50	68.99	183.54
		Deep Tissue Feature Extraction	0.56	2.99	8.40	4.17	9.96	20.54
		RAG Graph Building	0.12	2.04	25.6	6.33	19.98	85.73
ML		Cell-Graph Model	0.028	0.033	0.040	-	-	-
		Tissue-Graph Model	0.011	0.015	0.026	0.039	0.056	0.069
		HACT Model	0.034	0.041	0.057	-	-	-
Explainers	CG	GnnExplainer	12.00	13.09	35.33	-	-	-
		GraphGrad-CAM	0.011	0.022	0.035	-	-	-
		GraphGrad-CAM++	0.011	0.023	0.035	-	-	-
		GraphLRP	0.020	0.024	0.90	-	-	-
	TG	GnnExplainer	11.23	11.28	11.38	-	-	-
		GraphGrad-CAM	0.011	0.012	0.018	0.025	0.030	0.033
		GraphGrad-CAM++	0.011	0.013	0.018	0.026	0.030	0.033
		GraphLRP	0.011	0.014	0.016	0.079	0.085	0.089

feature cube representation, we process patches of size 144 × 144 resized to 224 × 224 w/o overlap by pretrained ResNet34.

TRoIs are processed using a cell- and tissue-graph model, and hierarchical cell-to-tissue graph model [Pati et al., 2022]. They consist of three PNA layers with 64 hidden units followed by a 2-layer MLP with 128 hidden units for classification. WSIs are processed using SegGini [Anklin et al., 2021], a weakly supervised approach basdn on tissue-graphs, which contains six GIN layers with 64 hidden units followed by a 2-layer MLP with 128 hidden units. The models process in near real-time irrespective of the increment in the graph size. The graph explainers are based on GNNs with 3 GIN layers, each having a 2-layer MLP with 32 hidden units, and a 2-layer MLP head. GnnExplainer is the slowest among all as it requires to optimize a mask to explain each image.

Table 9.3: Benchmarking HistoCartography for classification and segmentation (in %).

Task	Dataset	Model	Image Type	Avg. #pixels	#classes	Avg. Dice	Weighted F1
Classification	BRACS	CG-GNN	Tumor RoI (TRoI)	3.9×10^6 (40×)	7	-	55.9 ± 1.0
	BRACS	TG-GNN	TRoI	3.9×10^6 (40×)	7	-	56.6 ± 1.3
	BRACS	HACT-Net	TRoI	3.9×10^6 (40×)	7	-	61.5 ± 0.9
	BACH	HACT-Net	TRoI	3.1×10^6 (20×)	4	-	90.7 ± 0.5
	SICAPv2	SegGini	WSI	121×10^6 (10×)	6	-	62.0 ± 3.6
	UZH	SegGini	TMA	9.6×10^6 (40×)	6	-	56.8 ± 1.7
Seg.	SICAPv2	SegGini	WSI	121×10^6 (10×)	4	44.3 ± 2.0	-
	UZH	SegGini	TMA	9.6×10^6 (40×)	4	66.0 ± 3.1	-



(a) (b) (c)

Figure 9.2: Qualitative explanations of sample breast RoIs: (a) Benign, (b) ADH, (c) DCIS (d, e, f) and highlight the ten most important nuclei for the respective samples.

9.4.2 Performance benchmark

Table 9.3 benchmarks the performance of HistoCartography for classification and segmentation tasks. Classification is performed on BRACS [Pati et al., 2022] and BACH [Aresta et al., 2019] datasets to characterize breast tumors using cell-graph model, tissue-graph model, and HACT-Net [Pati et al., 2022], and the performance is measures by weighted-F1 score. Segmentation is performed using SegGini [Anklin et al., 2021] to delineate Gleason patterns in prostate cancer images from UZH [Zhong et al., 2017] and SICAPv2 [Silva-Rodrìguez et al., 2020], and the performance is measured by average Dice score. We evaluate on various image types, i. e., tumor RoIs, tissue microarrays, and whole-slides, to highlight the scalability of entity-graphs in HistoCartography to arbitrary image dimensions.

Figure 9.2 presents the outcome of GraphGradCAM function in HistoCartography to interpret a cell-graph model. This function renders per-image explanations in terms of node-level saliency maps by applying post-hoc feature attribution methods on trained cell-graph model. Further, the cell-graph model can be interpreted by characterizing the highlighted important nuclei per-image, as shown in Figure 9.2.

9.5 Conclusion

We introduced HistoCartography, the first open source library, to the best of our knowledge, to facilitate graph analytics, i. e., graph representation, learning, and explainability, in computational pathology. It can potentially enable researchers to develop entity-graph based pathology workflows by leveraging the inbuilt helpers. As the library is built on python, the deep learning researchers can seamlessly customize and integrate the functionalities into their task-specific workflows. HistoCartography is constantly growing with new functionalities and improved implementations, aiming to promote the adoption of graph-based analysis in computational pathology.

9.6 Appendices

9.6.1 HistoCartography ecosystem

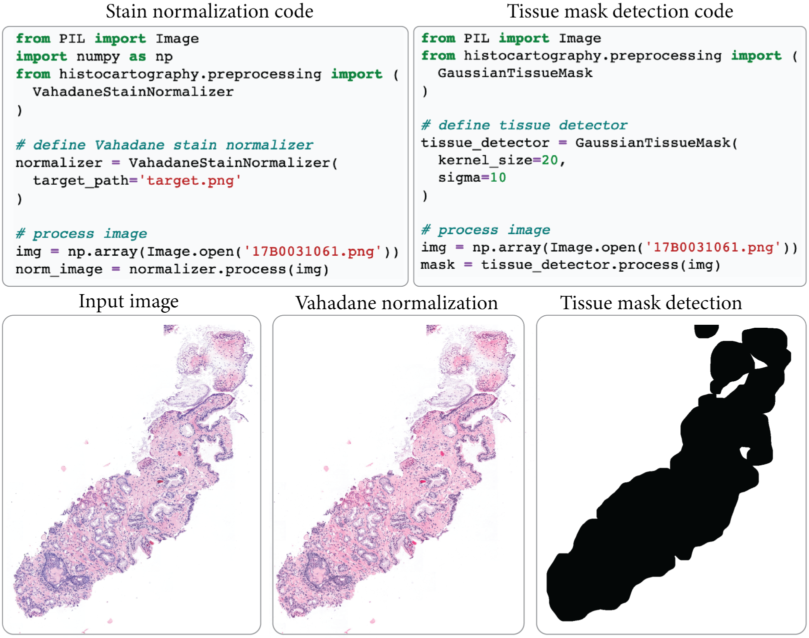
HistoCartography core functionalities can be tested using a set of examples available at url. Examples include stain normalization, cell- and tissue-graph generation, cell-graph explanation, and feature cube extraction. Additionally, a Jupyter Notebook presenting the library interpretability and explainability capabilities can be found at url. Individual functions are thoroughly unit tested (88% unit test coverage), and can be accessed at url. The code documentation, which provides a user-friendly approach to understanding HistoCartography architecture and modules can be accessed at `https://histocartography.github.io/histocartography/`. Finally, papers using HistoCartography can be found at `https://github.com/histocartography`.

9.6.2 HistoCartography syntax

In this section, we introduce the syntax to implement the functionalities of HistoCartography. Figure 9.3 presents code snippets to implement Vahadane stain normalization and tissue mask detection. Figure 9.4 shows the syntax for building cell- and tissue-graphs. Noticeably, these functionalities require only ten lines of code by using HistoCartography, which could have otherwise required a few hundred lines. In Figure 9.5, we present the syntax to declare and run a cell- and tissue-graph model. All the model parameters, e. g., GNN type, number of GNN layers, can be adapted and fine-tuned using a configuration file. Finally, Figure 9.6 shows code snippets to use the graph explainability modules. All explainers follow a similar syntax with the same input and output types, making implementation and integration straightforward.

9.6.3 Supported handcrafted features

In this section, we provide a comprehensive list of morphological and topological features which can be extracted per-entity by HistoCartography. Morphological features include shape, size and texture properties, namely, entity area, convex area, eccentricity, equivalent diameter, euler number, length of the major and minor axis, orientation, perimeter, solidity, convex hull perimeter, roughness, shape factor, ellipticity, roudness. Texture properties are based on gray-level co-occurrence matrices (GLCM). Specifically, we extract the GLCM contrast, dissimilarity, homogeneity, energy, angular speed moment and dispersion. The topological features are based on the entity density computed as the mean and variance



Stain normalization code

```
from PIL import Image
import numpy as np
from histocartography.preprocessing import (
  VahadaneStainNormalizer
)

# define Vahadane stain normalizer
normalizer = VahadaneStainNormalizer(
  target_path='target.png'
)

# process image
img = np.array(Image.open('17B0031061.png'))
norm_image = normalizer.process(img)
```

Tissue mask detection code

```
from PIL import Image
from histocartography.preprocessing import (
  GaussianTissueMask
)

# define tissue detector
tissue_detector = GaussianTissueMask(
  kernel_size=20,
  sigma=10
)

# process image
img = np.array(Image.open('17B0031061.png'))
mask = tissue_detector.process(img)
```

Input image

Vahadane normalization

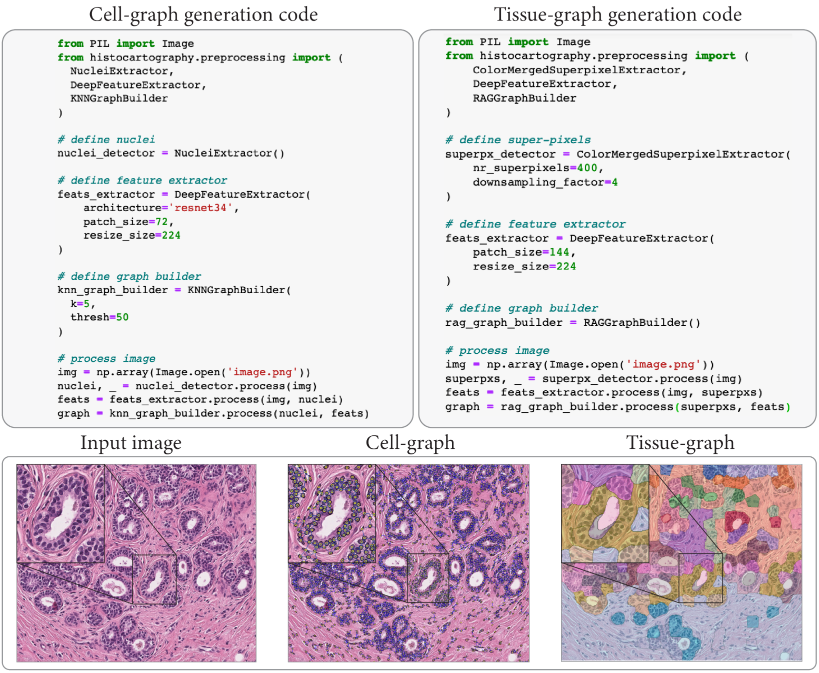
Tissue mask detection

Figure 9.3: Implementation of Vahadane stain normalization (left) and tissue mask detection (right) with the *Preprocessing* functionalities in the HistoCartography API.

of entity crowdedness. These features can be computed for the most important set of entities highlighted by the graph explainability techniques, and utilized along with prior pathological knowledge to interpret the trained entity-graph models.

9.6.4 Future of HistoCartography

HistoCartography development is only in its infancy, bugs will be fixed as people use it, new modules will be developed as the community develops novel graph-based methods and algorithms. Nevertheless, HistoCartography can already be used for developing new projects. Thanks to its modularity, pipelines can be developed by only partially using HistoCartography, e. g., only for building tissue-graphs, while novel components that require more flexibility and control can be developed on the side, e. g., for developing new models.



Cell-graph generation code

```python
from PIL import Image
from histocartography.preprocessing import (
  NucleiExtractor,
  DeepFeatureExtractor,
  KNNGraphBuilder
)

# define nuclei
nuclei_detector = NucleiExtractor()

# define feature extractor
feats_extractor = DeepFeatureExtractor(
    architecture='resnet34',
    patch_size=72,
    resize_size=224
)

# define graph builder
knn_graph_builder = KNNGraphBuilder(
  k=5,
  thresh=50
)

# process image
img = np.array(Image.open('image.png'))
nuclei, _ = nuclei_detector.process(img)
feats = feats_extractor.process(img, nuclei)
graph = knn_graph_builder.process(nuclei, feats)
```

Tissue-graph generation code

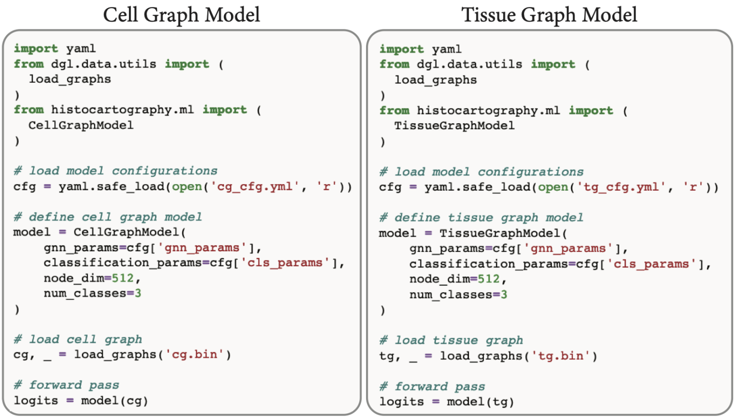
```python
from PIL import Image
from histocartography.preprocessing import (
    ColorMergedSuperpixelExtractor,
    DeepFeatureExtractor,
    RAGGraphBuilder
)

# define super-pixels
superpx_detector = ColorMergedSuperpixelExtractor(
    nr_superpixels=400,
    downsampling_factor=4
)

# define feature extractor
feats_extractor = DeepFeatureExtractor(
    patch_size=144,
    resize_size=224
)

# define graph builder
rag_graph_builder = RAGGraphBuilder()

# process image
img = np.array(Image.open('image.png'))
superpxs, _ = superpx_detector.process(img)
feats = feats_extractor.process(img, superpxs)
graph = rag_graph_builder.process(superpxs, feats)
```

Input image

Cell-graph

Tissue-graph

Figure 9.4: Implementation of cell-graph (left) and tissue-graph (right) generation using the graph builders in HistoCartography.



Cell Graph Model

```python
import yaml
from dgl.data.utils import (
  load_graphs
)
from histocartography.ml import (
  CellGraphModel
)

# load model configurations
cfg = yaml.safe_load(open('cg_cfg.yml', 'r'))

# define cell graph model
model = CellGraphModel(
    gnn_params=cfg['gnn_params'],
    classification_params=cfg['cls_params'],
    node_dim=512,
    num_classes=3
)

# load cell graph
cg, _ = load_graphs('cg.bin')

# forward pass
logits = model(cg)
```

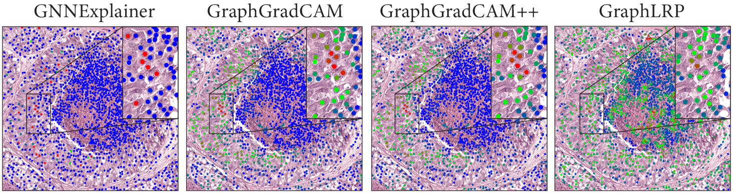
Tissue Graph Model

```python
import yaml
from dgl.data.utils import (
  load_graphs
)
from histocartography.ml import (
  TissueGraphModel
)

# load model configurations
cfg = yaml.safe_load(open('tg_cfg.yml', 'r'))

# define tissue graph model
model = TissueGraphModel(
    gnn_params=cfg['gnn_params'],
    classification_params=cfg['cls_params'],
    node_dim=512,
    num_classes=3
)

# load tissue graph
tg, _ = load_graphs('tg.bin')

# forward pass
logits = model(tg)
```

Figure 9.5: Implementation of the cell- (left) and tissue- graph (right) models by using the ML modules in HistoCartography.

Cell graph explainer code

```
from histocartography.interpretability import (
  GraphGradCAMExplainer,
  GraphGradCAMPPExplainer,
  GraphPruningExplainer,
  GraphLRPExplainer
)

# load pretrained model
model = CellGraphModel(config['gnn_params'], config['cls_params'], 512, pretrained=True)

# load cell graph
graph, _ = load_graphs('291_dcis_18.bin')

# define graph explainers
grad_cam_explainer = GraphGradCAMExplainer(model=model)
grad_campp_explainer = GraphGradCAMPPExplainer(model=model)
gnn_explainer = GraphPruningExplainer(model=model)
graph_lrp_explainer = GraphLRPExplainer(model=model)

# explain cell graph
grad_cam_scores, _ = grad_cam_explainer.process(graph)
grad_campp_scores, _ = grad_campp_explainer.process(graph)
gnn_explainer_scores, _ = gnn_explainer.process(graph)
graph_lrp_scores, _ = graph_lrp_explainer.process(graph)
```



GNNExplainer GraphGradCAM GraphGradCAM++ GraphLRP

Figure 9.6: Implementation of graph explainers in HistoCartography. The most important nodes are marked in red and the least important ones in blue.

10

Conclusions

In this chapter, we summarize the major contributions of this thesis, and highlight their strengths and limitations. Then, we postulate a set of future research directions, both on the methodological and clinical aspects.

10.1 Summary and limitations

The core concept of this thesis is to shift the analytical paradigm of histopathology images from pixels to histological entities. Operating in the entity-paradigm enables to address several limitations of the pixel-based analysis, such as, (1) disregard for histological entities, (2) inability to simultaneously capture both local and global context, (3) intensive computational requirements for processing large WSIs, and (4) difficulty in comprehending the model interpretations by pathologists. In the entity-based analytical workflow, first, a histopathology image is transformed into an entity-graph, where the nodes and edges denote histological entities and entity-to-entity interactions. Subsequently, DL methods operating on graph structured data, in particular GNNs, are devised to address various tasks. An entity-graph is built in three steps, i. e., (1) a task-relevant entity identification to form the graph nodes, (2) encoding the entities for node characterization, and (3) a task-relevant topology builder to define the graph edges. The nodes and topology of a graph encode the local phenotype and global tissue microenvironment to comprehensively represent the tissue composition. The GNNs process these graphs to construct context-aware entity- and graph-representations to enable enhanced tissue-structure to histopathology task mapping. The advantages of DL on entity-guided representations, in terms of superior performance, scalability, and interpretability, are demonstrated across different chapters in this thesis.

The motivation behind emphasizing on histological entities in an image, instead of the entire image, is described in **Chapter 3**. Specifically, diagnostically relevant entities from a HER2-stained immunohistochemisty image, i. e., cell membranes and secondary stained regions, are analyzed to quantify the quality of HER2 staining. The extracted entities across a dataset of images are characterized to define disease-specific *staining quality metrics* (SQMs). Further, sensitivity analyses of the SQMs over the staining-parametric space are performed to determine disease-specific optimal staining-process parameters. Such optimization enables to standardize the staining quality assessment, and significantly reduces the search space for determining an appropriate staining protocol. The resulting protocols are further substantiated by comparing against clinical protocols through staining several standard cell-blocks. The findings motivated to explore the potential of the entity-paradigm through advanced entity encoding and computational techniques.

Encoding of a tissue composition in form of an entity-graph, and processing the entity-graph via a GNN is proposed in **Chapter 4**. The chapter illustrates the potential and advantages of leveraging pathological prior into comprehensive tissue representation. Further, the scalability of entity-graph-based processing is highlighted by processing histopathology images of arbitrary shapes and sizes. Specifically, a Hierarchical Cell-to-Tissue (HACT) representation is proposed to encode both cell-level and tissue-level information in a hierarchical fashion. The cell-level information is encoded as a cell-graph, where nodes and edges denote nuclei and inter-nuclei interactions. The tissue-level information is captured as a tissue-graph, where nodes are tissue regions and edges are relationships among adjacent regions. The intra-level hierarchies are encoded in terms of edges denoting the relative spatial distribution among nuclei and tissue regions. Subsequently, a novel GNN, HACT-Net, is devised to hierarchically process the HACT graphs. HACT combined with HACT-Net is benchmarked on the curated BRACS dataset, to date the largest cohort of H&E breast histopathology TRoIs. The framework outperformed several CNNs and performed comparable to domain-expert pathologists. While being promising, the approach suffers from several challenges, listed as follows. (1) The construction of an entity-graph requires task-specific entity detectors. For example, to represent a WSI using glands as entities, a gland detector is required, which is a non-trivial task. (2) To construct domain-specific HACT representations, either HACT-Net is required to be made end-to-end trainable, or histopathology-specific pre-trained feature extractors [Tellez et al., 2021; Shaban et al., 2020] are to be developed. A domain-specific representation refers to encoding histopathology-specific features in the nodes of the HACT graph. The nodes of the proposed HACT graphs are encoded using pre-trained CNNs on ImageNet dataset. In the current form, HACT-Net can in theory be trained end-to-end but at a really high computational cost, as the graphs need to be built on-the-fly during batch construction. Notably, this challenge is common to other CNN-based methods in CP as well. Further, the downstream gain in classification performance would be marginal if not nonexistent. (3) The impact of other inter-level graph topologies are not studied. We hypothesize that any topology enforcing spatial connectivity, e. g., k-NN, and radial topology, would lead to similar performance. This is a reasonable assumption as the uni-level entity-graphs are homophilous, i. e., a node and its neighbors share the same functional properties. For example, the nuclei belonging to a gland bear similar morphological features, compared to the nuclei outside the gland, i. e., stromal nuclei or lymphocytes. In this sense, cell-graph GNNs act as low-pass filters to learn discriminative relation-aware nuclei phenotypes.

Chapter 5 and **6** propose methods to interpret and explain an entity-graph in pathologist-understandable terminologies. A post-hoc perturbation-based interpretability technique (or explainer), CGExplainer is proposed in **Chapter 5** to identify salient nuclei in a cell-graph, i. e., an explanation, processed by a cell-GNN. It illustrated the sparseness of informative content in cell-graphs which drive the model prediction. Considering the limitations of qualitative evaluation of the explanations, generated by several explainers, a set of quantitative metrics, i. e., *maximum*, *average*, and *correlated separability*, are proposed in **Chapter 6**. The metrics are built on the statistics of class separability due to pathologically measurable *concepts*, e. g., shape, size, and chromaticity of nuclei. The limitations of the approaches are listed as follows. (1) Gathering a universal prior for computing correlated separability is not always feasible, which requires the prior knowledge of task-specific expert pathologists. (2) The conceptualization of the set of concepts requires domain- and task-specific knowledge. Even the transferability of the metrics, built on a set of concepts, across related tasks or domains may require the supervision of domain experts.

A co-representation learning framework (CoReL) for classification tasks is proposed in **Chapter 7** to learn from limited annotated data. CoReL simultaneously captures the class-label information and the local spatial distribution information of the data points in the embedding space to enhance the learning capability of the model. A key contribution of the method includes a novel context-aware pair mining strategy and a soft-multi-pair objective to boost the efficacy of the DML component in the method. Notably, CoReL achieves state-of-the-art classification on five benchmark datasets across three histopathology tasks. However, CoReL includes certain anticipated limitations. (1) The DML component may be sensitive to the training batch size. Empirically, the performance increases with increasing batch size for datasets with high variability. This may lead to a high computational cost and long training time per epoch. (2) The joint optimization of the multiple objectives may be sensitive to the loss trade-off hyperparameter, which may be computationally expensive to tune.

Chapter 8 addresses the challenge of learning from weak-supervision in CP. Specifically, the chapter proposes Whole-slide SegmentatIon using Graphs for HisTology (WholeSIGHT) to simultaneously segment and classify WSIs by using weakly-supervised WSI-level annotations. To this end, the method leverages the potentials of entity-graph representation and learning, graph interpretability technique, and DML. At first, a WSI is encoded into a tissue-graph and a GNN is trained for graph classification. Then, a graph explainer generates pseudo node-level labels, which are used to train a node classifier. The class prediction and segmentation are obtained from the graph- and node-classifiers, respectively. The main limitation of WholeSIGHT is the module for detecting tissue regions, which is both time-consuming and sub-optimal. This module relies on basic image processing to identify tissue regions in form of superpixels. To optimize this module, either a DL-based superpixel detector or a dedicated tissue region detector is required. Despite this limitation, WholeSIGHT is applicable to real-world setting. It provides better fine-grained segmentation compared to state-of-the-art weakly-supervised MIL methods, which typically process densely overlapping patches for an acceptable segmentation. Further, the chapter proposes a Bayesian variant of WholeSIGHT for better generalization to out-of-domain datasets. The method achieves state-of-the-art weakly-supervised Gleason grading and Gleason pattern segmentation performances on prostate cancer needle biopsies.

Chapter 9 introduces HistoCartography, a generic open-source python library to facilitate effective graph analytics in digital histopathology.

10.2 Future work

A thesis is a continuum of projects and tasks, through which a plethora of new ideas emerge. Below we list some research directions that could be promising to further explored.

Pathological prior-aware tissue modeling: One of the key strengths of entity-graphs is their flexibility to encode complex relationships among histological entities. The choice of the entity and inter-entity interactions up to arbitrary spatial distances can be modelled by an entity-graph. More complex heterogeneous entity-graphs can also be designed by using multiset of entities and intra-entity relationships for a comprehensive encoding of tissue composition. Such degree of flexibility enables the incorporation of well-established prior knowledge into tissue modeling. For example, to evaluate the Gleason grade in prostate cancer, the prior dictates the importance of analyzing the phenotype and topological distribution of glands, the constituting nuclei, and associated stroma [Gordetsky et al., 2016]. Therefore, a hierarchical entity-graph representation, similar to HACT, can be constructed by encoding the characteristics and distribution of the mentioned relevant entities, for an improved diagnosis. Further, the entity-graphs allow to selectively operate on a subset of entities. For instance, to characterize tumor infiltrating lymphocytes, only the distribution of tumorous epithelial nuclei and lymphocytes, among other nuclei types, can be evaluated. Such selective assessment can improve the diagnosis by reducing the amount of uninformative content and emphasizing more on the informative substance.

Patient-level-graphs with heterogeneous concepts: The explanations of entity-guided GNNs, generated by interpretability techniques, illustrate that different subset of *concepts* are informative for differentiating different disease pairs. Further, the relevant subsets are also patient-dependent. These observations convey that there exist a level of homogeneity and heterogeneity among patients at concept-level for disease stratification. However, DL methods, operating under *i.i.d.* assumption, do not leverage the inter-sample relationships. An exact modeling of these relations is also infeasible for a direct incorporation into DL methods. Nevertheless, the relations can be learned through a *patient-graph*, where the nodes encode patients. The edges can exploit the homogeneity and heterogeneity among the patients at concept-levels, such that the graph can encode heterogeneous information in abstraction. A utilization of this additional information can potentially improve data modeling, inducing better generalization.

Multi-modal entity-graphs: Entity-graphs are generic and applicable to other imaging modalities and medical domains. These graphs can be employed to perform early- or late-fusion of multi-modal information from diverse modalities, such as pathology, radiology, clinical records, and multi-omics. Within the pathology domain, multi-modal graphs can also be constructed by leveraging different types of tissue stains. As each stain is targeted to highlight specific tissue structures, an inclusion of a variety of stains can suffice a comprehensive tissue encoding. The modalities can be simply incorporated across registered images by considering each stain as an image channel. However, considering the complexities of registering images across stains, stain-specific relevant entity-graphs can be built, and the context information of the entities can be utilized to include cross-stain relationships. As this would be registration-free, it can seamlessly include diverse stains highlighting different tissue structures, which are challenging to register.

Bibliography

[Pati et al., 2022] Pati P., Jaume G., Foncubierta-Rodriguez A., Feroce F., Anniciello A., Scognamiglio G., Brancati N., Fiche M., Dubruc E., Riccio D., et al. (2022). "Hierarchical Graph Representations in Digital Pathology". In: *Medical Image Analysis* 75, p. 102264.

[Ahmedt-Aristizabal et al., 2021] Ahmedt-Aristizabal D., Armin M., Denman S., Fookes C., and Petersson L. (2021). "A Survey on Graph-Based Deep Learning for Computational Histopathology". In: *arXiv:2107.00272*.

[Anklin et al., 2021] Anklin V., Pati P., Jaume J., Bozorgtabar B., Foncubierta-Rodriguez A., Thiran J., Sibony M., Gabrani M., and Goksel O. (2021). "Learning Whole-Slide Segmentation from Inexact and Incomplete Labels using Tissue Graphs". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 636–646.

[Aubreville et al., 2021] Aubreville M., Bertram C., Veta M., Klopfleisch R., Stathonikos N., Breininger K., ter Hoeve N., Ciompi F., and Maier A. (2021). "Quantifying the Scanner-Induced Domain Gap in Mitosis Detection". In: *Medical Imaging with Deep Learning (MIDL)*.

[Bankhead et al., 2021] Bankhead P. et al. (2021). *QuPath*. `https://qupath.github.io/`.

[Beezley et al., 2021] Beezley J. et al. (2021). *HistomicsTK*. `https://github.com/DigitalSlideArchive/HistomicsTK`.

[Budd et al., 2021] Budd S., Robinson E., and Kainz B. (2021). "A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis". In: *Medical Image Analysis*, p. 102062.

[Bulten et al., 2021] Bulten W., Balkenhol M., Belinga J., Brilhante A., Cakir A., Egevad L., Eklund M., Farré X., Geronatsiou K., Molinié V., et al. (2021). "Artificial Intelligence Assistance Significantly Improves Gleason Grading of Prostate Biopsies by Pathologists". In: 34 (3), pp. 660–671.

[Businesswire, 2021] Businesswire (2021). *Paige Receives First Ever FDA Approval for AI Product in Digital Pathology*. `https://www.businesswire.com/news/home/20210922005369/en/Paige-Receives-First-Ever-FDA-Approval-for-AI-Product-in-Digital-Pathology`.

[Chan et al., 2021] Chan L., Hosseini M., and Plataniotis K. (2021). "A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains". In: *International Journal of Computer Vision* 129 (2), pp. 361–384.

[Faryna et al., 2021] Faryna K., van der Laak J., and Litjens G. (2021). "Tailoring Automated Data Augmentation to H&E-Stained Histopathology". In: *Medical Imaging with Deep Learning (MIDL)*.

[Folmsbee et al., 2021] Folmsbee J., Brandwein-Weber M., and Doyle S. (2021). "Whole Slide Semantic Segmentation: Large Scale Active Learning for Digital Pathology". In: *SPIE Medical Imaging 2021: Digital Pathology*. Vol. 11603, 116030G.

[Gomariz et al., 2021] Gomariz A., Portenier T., Nombela-Arrieta C., and Goksel O. (2021). "Probabilistic Spatial Analysis in Quantitative Microscopy with Uncertainty-Aware Cell Detection using Deep Bayesian Regression of Density Maps". In: *arXiv:2102.11865*.

[Ho et al., 2021] Ho D., Yarlagadda D., D'Alfonso T., Hanna M., Grabenstetter A., Ntiamoah P., Brogi E., Tan L., and Fuchs T. (2021). "Deep Multi-Magnification Networks for Multi-Class Breast Cancer Image Segmentation". In: *Computerized Medical Imaging and Graphics* 88 (101866).

[Jaume et al., 2021a] Jaume G., Pati P., Anklin V., Foncubierta A., and Gabrani M. (2021a). "HistoCartography: A Toolkit for Graph Analytics in Digital Pathology". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshops*, pp. 117–128.

[Jaume et al., 2021b] Jaume G., Pati P., Bozorgtabar B., Foncubierta-Rodríguez A., Feroce F., Anniciello A., Rau T., Thiran J., Gabrani M., and Goksel O. (2021b). "Quantifying Explainers of Graph Neural Networks in Computational Pathology". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8106–8116.

[Koohbanani et al., 2021] Koohbanani N., Unnikrishnan B., Khurram S., Krishnaswamy P., and Rajpoot N. (2021). "Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations". In: *IEEE Transactions on Medical Imaging*.

[Lai et al., 2021] Lai Z., Wang C., Oliveira L., Dugger B., Cheung S., and Chuah C. (2021). "Joint Semi-Supervised and Active Learning for Segmentation of Gigapixel Pathology Images With Cost-Effective Labeling". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 591–600.

[Levy et al., 2021] Levy J., Haudenschild C., Barwick C., Christensen B., and Vaickus L. (2021). “Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks”. In: *Pacific Symposium on Biocomputing* 26, pp. 285–296.

[Li et al., 2021] Li Z., Liu F., Yang W., Peng S., and Zhou J. (2021). “A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects”. In: *IEEE Transactions on Neural Networks and Learning Systems*.

[Linkon et al., 2021] Linkon A., Labib M., Hasan T., Hossain M., Marium E., et al. (2021). “Deep Learning in Prostate Cancer Diagnosis and Gleason Grading in Histopathology Images: An Extensive Study”. In: *Informatics in Medicine Unlocked*, p. 100582.

[Lu et al., 2021] Lu M., Williamson D., Chen T., Chen R., Barbieri M., and Mahmood F. (2021). “Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images”. In: *Nature Biomedical Engineering* 5, pp. 555–570.

[Mahapatra et al., 2021a] Mahapatra D., Poellinger A., Shao L., and Reyes M. (2021a). “Interpretability-Driven Sample Selection Using Self Supervised Learning For Disease Classification And Segmentation”. In: *IEEE Transactions on Medical Imaging*.

[Mahapatra et al., 2021b] Mahapatra D., Kuanar S., Bozorgtabar B., and Ge Z. (2021b). “Self-supervised Learning of Inter-label Geometric Relationships for Gleason Grade Segmentation”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pp. 57–67.

[Mahapatra et al., 2021c] Mahapatra D., Bozorgtabar B., Kuanar S., and Ge Z. (2021c). “Self-supervised Multimodal Generalized Zero Shot Learning for Gleason Grading”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pp. 46–56.

[Marini et al., 2021] Marini N., Otálora S., M"uller H., and Atzori M. (2021). “Semi-supervised Learning with a Teacher-Student Paradigm for Histopathology Classification: A Resource to Face Data Heterogeneity and Lack of Local Annotations”. In: *International Conference on Pattern Recognition (ICPR)*, pp. 105–119.

[Mohseni et al., 2021] Mohseni S., Zarei N., and Ragan E. (2021). “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems”. In: *ACM Transactions on Interactive Intelligent Systems* 11 (3-4), pp. 1–45.

[Otálora et al., 2021] Otálora S., Marini N., Müller H., and Atzori M. (2021). “Combining Weakly and Strongly Supervised Learning Improves Strong Supervision in Gleason Pattern Classification”. In: *BMC Medical Imaging* 21 (1), pp. 1–14.

[Ozen et al., 2021] Ozen Y., Aksoy S., Kösemehmeto
(2021). “Self-Supervised Learning with Graph Neural Networks for Region of Interest Retrieval in Histopathology”. In: *International Conference on Pattern Recognition (ICPR)*, pp. 6329–6334.

[Pati et al., 2021] Pati P., Foncubierta-Rodriguez A., Goksel O., and Gabrani M. (2021). “Reducing Annotation Effort in Digital Pathology: A Co-Representation Learning Framework for Classification Tasks”. In: *Medical Image Analysis* 67, p. 101859.

[Pinckaers et al., 2021] Pinckaers H., Bulten W., van der Laak J., and Litjens G. (2021). “Detection of Prostate Cancer in Whole-Slide Images Through End-To-End Training with Image-Level Labels”. In: *IEEE Transactions on Medical Imaging* 40 (7), pp. 1817–1826.

[Pocevičiūtė et al., 2021] Pocevičiūtė M., Eilertsen G., and Lundström C. (2021). “Unsupervised Anomaly Detection in Digital Pathology Using GANs”. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1878–1882.

[Roy et al., 2021] Roy M., Kong J., Kashyap S., Pastore V., Wang F., Wong K., and Mukherjee V. (2021). “Convolutional Autoencoder based Model Histocae for Segmentation of Viable Tumor Regions in Liver Whole-Slide Images”. In: *Scientific Reports* 11 (1), pp. 1–10.

[Shi et al., 2021] Shi J., Wang R., Zheng Y., Jiang Z., Zhang H., and Yu L. (2021). “Cervical Cell Classification with Graph Convolutional Network”. In: *Computer Methods and Programs in Biomedicine* 198, p. 105807.

[Silva-Rodrìguez et al., 2021] Silva-Rodrìguez J., Colomer A., and Naranjo V. (2021). “WeGleNet: A Weakly-Supervised Convolutional Neural Network for the Semantic Segmentation of Gleason Grades in Prostate Histology Images”. In: *Computerized Medical Imaging and Graphics* 88 (101846), pp. 6329–6334.

[Srinidhi et al., 2021] Srinidhi C., Ciga O., and Martel A. (2021). “Deep Neural Network Models for Computational Histopathology: A Survey”. In: *Medical Image Analysis* 67, p. 101813.

[Studer et al., 2021] Studer L., Wallau J., Dawson H., Zlobec I., and Fischer A. (2021). “Classification of Intestinal Gland Cell-Graphs Using Graph Neural Networks”. In: *International Conference on Pattern Recognition (ICPR)*, pp. 6329–6334.

[Su et al., 2021] Su L., Liu Y., Wang M., and Li A. (2021). “Semi-HIC: A Novel Semi-supervised Deep Learning Method for Histopathological Image Classification”. In: *Computers in Biology and Medicine* 137, p. 104788.

[Sung et al., 2021] Sung H., Ferlay J., Siegel R., 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”. In: *CA Cancer Journal for Clinicians* 71 (3), pp. 209–249.

[Tataru et al., 2021] Tataru O., Vartolomei M., Rassweiler J., Virgil O., Lucarelli G., Porpiglia F., Amparore D., Manfredi M., Carrieri G., Falagario U., et al. (2021). “Artificial Intelligence and Machine Learning in Prostate Cancer Patient Management—Current Trends and Future Perspectives”. In: *Diagnostics* 11 (2), p. 354.

[Tellez et al., 2021] Tellez D., Litjens G., Laak J., and Ciompi F. (2021). “Neural Image Compression for Gigapixel Histopathology Image Analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2), pp. 567–578.

[Verma et al., 2021] Verma R., Kumar N., Patil A., Kurian N., Rane S., and Sethi A. (2021). “Multi-organ Nuclei Segmentation and Classification Challenge 2020”. In: *IEEE Transactions on Medical Imaging* 39, pp. 1380–1391.

[Wang et al., 2021] Wang X., Yang S., Zhang J., Wang M., Zhang J., Huang J., Yang W., and Han X. (2021). “TransPath: Transformer-based Self-supervised Learning for Histopathological Image Classification”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 186–195.

[Yao et al., 2021a] Yao J., Zhu X., Jonnagaddala J., Hawkins N., and Huang J. (2021a). “Whole Slide Images based Cancer Survival Prediction using Attention Guided Deep Multiple Instance Learning Networks”. In: *Medical Image Analysis* 65 (473), pp. 6329–6334.

[Yao et al., 2021b] Yao K., Huang K., Sun J., and Jude C. (2021b). “AD-GAN: End-to-end Unsupervised Nuclei Segmentation with Aligned Disentangling Training”. In: *arXiv:2107.11022*, pp. 6329–6334.

[Yu et al., 2021] Yu G., Xie T., Xu C., Shi X., Wu C., Sun K., Meng R., Meng X., Wang K., Xiao H., et al. (2021). “Accurate Recognition of Colorectal Cancer with Semi-Supervised Deep Learning on Pathological Images”. In: *Nature Communications* 12 (1), pp. 1–13.

[Abbet et al., 2020] Abbet C., Zlobec I., Bozorgtabar B., and Thiran J. (2020). “Divide-and-rule: Self-supervised Learning for Survival Analysis in Colorectal Cancer”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 480–489.

[Adnan et al., 2020] Adnan M., Kalra S., and Tizhoosh H. (2020). “Representation Learning of Histopathology Images using Graph Neural Networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4254–4261.

[Arrieta et al., 2020] Arrieta A., Diaz-Rodriguez N., Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., and Herrera F. (2020). “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”. In: *Information Fusion* 58, pp. 82–115.

[Aubreville et al., 2020] Aubreville M., Bertram C., Jabari S., Marzahl C., Klopfleisch R., and Maier A. (2020). “Inter-Species, Inter-Tissue Domain Adaptation for Mitotic Figure Assessment: Learning New Tricks from Old Dogs”. In: *Bildverarbeitung für die Medizin 2020*, pp. 1–7.

[Aygüneş et al., 2020] Aygüneş B., Aksoy S., Cinbiş R., Kösemehmeto (2020). “Graph Convolutional Networks for Region of Interest Classification in Breast Histopathology”. In: *SPIE Medical Imaging: Digital Pathology*. Vol. 11320, 113200K.

[Bulten et al., 2020a] Bulten W., Pinckaers H., van Boven H., Vink R., de Bel T., van Ginneken B., van der Laak J., Hulsbergen-van de Kaa C., and Litjens G. (2020a). “Automated Deep-learning System for Gleason Grading of Prostate Cancer using Biopsies: A Diagnostic Study”. In: *The Lancet Oncology* 21 (2), pp. 233–241.

[Bulten et al., 2020b] Bulten W., Litjens G., Pinckaers H., Ström P., Eklund M., Egevad L., Grönberg H., Kartasalo K., Ruusuvuori P., Häkkinen T., Dane S., and Demkin M. (2020b). *Prostate cANcer graDe Assessment (PANDA) Challenge*. `https://panda.grand-challenge.org/`.

[Byfield et al., 2020] Byfield P. et al. (2020). *Syntax*. `https://github.com/jgamper/compay-syntax/tags`.

[Caron et al., 2020] Caron M., Misra I., Mairal J., Goyal P., Bojanowski P., and Joulin A. (2020). “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 11320, 113200K.

[Chen et al., 2020a] Chen R., Lu M., Wang J., Williamson D., Rodig S., Lindeman N., and Mahmood F. (2020a). “Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis”. In: *IEEE Transactions on Medical Imaging* 11320, 113200K.

[Chen et al., 2020b] Chen T., Kornblith S., Norouzi M., and Hinton G. (2020b). “A Simple Framework for Contrastive Learning of Visual Representations”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, pp. 1597–1607.

[Ciga et al., 2020] Ciga O., Martel A., and Xu T. (2020). “Self Supervised Contrastive Learning for Digital Histopathology”. In: *Machine Learning with Applications* 11320, p. 100198.

[Corso et al., 2020] Corso G., Cavalleri L., Beaini D., Liò P., and Velicković P. (2020). “Principal Neighbourhood Aggregation for Graph Nets”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 11320, 113200K.

[De Cao et al., 2020] De Cao N., Schlichtkrull M., Aziz W., and Titov I. (2020). “How Do Decisions Emerge Across Layers in Neural Models? Interpretation with Differentiable Masking”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. Vol. 11320, 113200K.

[Deng et al., 2020] Deng S., Zhang X., Yan W., Chang E., Fan Y., Lai M., and Xu Y. (2020). “Deep Learning in Digital Pathology Image Analysis: A Survey”. In: *Frontiers in Medicine* 11320, pp. 1–18.

[Dwivedi et al., 2020] Dwivedi V., Joshi C., Laurent T., Bengio Y., and Bresson X. (2020). “Benchmarking Graph Neural Networks”. In: *arXiv:2003.00982*. Vol. 11320, 113200K.

[Gamper et al., 2020] Gamper J., Koohbanani N., Benes K., Graham S., Jahanifar M., Khurram S., Azam A., Hewitt K., and Rajpoot N. (2020). “PanNuke Dataset Extension, Insights and Baselines”. In: *arXiv:2003.10778*. Vol. 11320, 113200K.

[Gilbert et al., 2020] Gilbert B. et al. (2020). *OpenSlide*. `https://github.com/openslide/openslide-python/`.

[Graziani et al., 2020] Graziani M., Andrearczyk V., Marchand-Maillet S., and Müller H. (2020). “Concept Attribution: Explaining CNN Decisions to Physicians”. In: *Computers in Biology and Medicine*. Vol. 123, p. 103865.

[Gustafsson et al., 2020] Gustafsson F., Danelljan M., and Schon T. (2020). “Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Vol. 11320, pp. 318–319.

[Hägele et al., 2020] Hägele M., Seegerer P., Lapuschkin S., Bockmayr M., Samek W., Klauschen F., Müller K., and Binder A. (2020). “Resolving Challenges in Deep Learning-based Analyses of Histopathological Images using Explanation Methods”. In: *Scientific Reports* 10 (1), pp. 1–12.

[Hamilton, 2020] Hamilton W. (2020). “Graph Representation Learning”. In: *Synthesis Lectures on Artifical Intelligence and Machine Learning* 14 (3), pp. 1–159.

[Harris et al., 2020] Harris C., Millman K., van der Walt S., Gommers R., Virtanen P., Cournapeau D., Wieser E., Taylor J., Berg S., Smith N., et al. (2020). “Array Programming with NumPy”. In: *Nature* 585 (7825), pp. 357–362.

[Hecht et al., 2020] Hecht H., Sarhan M., and Popovici V. (2020). “Disentangled Autoencoder for Cross-Stain Feature Extraction in Pathology Image Analysis”. In: *Applied Sciences* 10 (18), p. 6427.

[Horiguchi et al., 2020] Horiguchi S., Ikami D., and Aizawa K. (2020). “Significance of Softmax-based Features in Comparison to Distance Metric Learning-based Features”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (5), pp. 1279–1285.

[Huang et al., 2020] Huang Q., Yamada M., Tian Y., Singh D., Yin D., and Chang Y. (2020). “GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks”. In: *arXiv:2001.06216* 11320, 113200K.

[Idos et al., 2020] Idos G., Kwok J., Bonthala N., Kysh L., Gruber S., and Qu C. (2020). “The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis”. In: *Scientific Reports* 10 (1), pp. 1–14.

[Jaume et al., 2020] Jaume G., Pati P., Rodriguez A., Florinda F., Scognamiglio G., Anniciello A., Thiran J., Goksel O., and Gabrani M. (2020). “Towards Explainable Graph Representations in Digital Pathology”. In: *International Conference on Machine Learning (ICML) Workshops*. Vol. 11320, 113200K.

[Javed et al., 2020] Javed S., Mahmood A., Fraz M., Koohbanani N., Benes K., Tsang Y., Hewitt K., Epstein D., Snead D., and Rajpoot N. (2020). “Cellular Community Detection for Tissue Phenotyping in Colorectal Cancer Histology Images”. In: *Medical Image Analysis* 63, p. 101696.

[Kumar et al., 2020] Kumar N., Verma R., Anand D., Zhou Y., Onder O., Tsougenis E., Chen H., Heng P., Li J., Hu Z., et al. (2020). “A Multi-organ Nucleus Segmentation Challenge”. In: *IEEE Transactions on Medical Imaging* 39 (5), pp. 1380–1391.

[Levy et al., 2020] Levy J., Haudenschild C., Barwick C., Christensen B., and Vaickus L. (2020). “Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks”. In: *Proceedings of the Pacific Symposium*. Vol. 11320, pp. 285–296.

[Liu et al., 2020] Liu D., Zhang D., Song Y., Zhang F., O’Donnell L., Huang H., Chen M., and Cai W. (2020). “Unsupervised Instance Segmentation in Microscopy Images via Panoptic Domain Adaptation and Task Re-Weighting”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 4243–4252.

[Lu et al., 2020] Lu W., Graham S., Bilal M., Rajpoot N., and Minhas F. (2020). “Capturing Cellular Topology in Multi-Gigapixel Pathology Images”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Vol. 11320, pp. 260–261.

[Luo et al., 2020] Luo D., Cheng W., Xu D., Yu W., Zong B., Chen H., and Zhang X. (2020). "Parameterized Explainer for Graph Neural Network". In: *Advances in Neural Information Processing Systems (NeurIPS*. Vol. 11320, 113200K.

[Marcolini et al., 2020] Marcolini A., Arbitrio E., and Bussola N. (2020). *histolab*. `https://github.com/histolab/histolab`.

[Mormont et al., 2020] Mormont R., Geurts P., and Marée R. (2020). "Multi-task Pre-training of Deep Neural Networks for Digital Pathology". In: *IEEE Journal of Biomedical and Health Informatics* 25 (2), pp. 412–421.

[Nguyen et al., 2020] Nguyen A. and Martinez M. (2020). "On Quantitative Aspects of Model Interpretability". In: *arXiv:2007.07584*. Vol. 11320, 113200K.

[Parvatikar et al., 2020] Parvatikar A., Choudhary O., Ramanathan A., Navolotskaia O., Carter G., Tosun A., Fine J., and Chennubhotla S. (2020). "Modeling Histological Patterns for Differential Diagnosis of Atypical Breast Lesions". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 11320, pp. 550–560.

[Pati et al., 2020] Pati P., Jaume G., Fernandes L., Foncubierta A., Feroce F., Anniciello A., Scognamiglio G., Brancati N., Riccio D., Di Bonito M., et al. (2020). "HACT-Net: A Hierarchical Cell-to-Tissue Graph Neural Network for Histopathological Image Classification". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshops*. Vol. 11320, pp. 208–219.

[Pinckaers et al., 2020] Pinckaers H., Ginneken B., and Litjens G. (2020). "Streaming Convolutional Neural Networks for End-To-End Learning with Multi-Megapixel Images". In: *IEEE Transactions on Medical Imaging* 39 (5), pp. 1306–1315.

[Raju et al., 2020] Raju A., Yao J., Haq M., Jonnagaddala J., and Huang J. (2020). "Graph Attention Multi-instance Learning for Accurate Colorectal Cancer Staging". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 529–539.

[Rathore et al., 2020] Rathore S., Niazi T., Iftikhar M., and Chaddad A. (2020). "Glioma Grading via Analysis of Digital Pathology Images using Machine Learning". In: *Cancers* 12 (3), p. 578.

[Salvatorelli et al., 2020] Salvatorelli L., Puzzo L., Vecchio G., Caltabiano R., Virzì V., and Magro G. (2020). "Ductal Carcinoma In Situ of the Breast: An Update with Emphasis on Radiological and Morphological Features as Predictive Prognostic Factors". In: *Cancers* 12 (3), p. 609.

[Shaban et al., 2020] Shaban M., Awan R., Fraz M., Azam A., Tsang Y., Snead D., and Rajpoot N. (2020). "Context-Aware Convolutional Neural Network for Grading of Colorectal Cancer Histology Images". In: *IEEE Transactions on Medical Imaging* 39 (7), pp. 2395–2405.

[Siegel et al., 2020] Siegel R., Miller K., and Jemal A. (2020). "Cancer Statistics, 2020". In: *CA: A Cancer Journal for Clinicians* 70, pp. 7–30.

[Silva-Rodrìguez et al., 2020] Silva-Rodrìguez J., Colomer A., Sales M., Molina R., and Naranjo V. (2020). "Going Deeper through the Gleason Scoring Scale: An Automatic End-To-End System for Histology Prostate Grading and Cribriform Pattern Detection". In: *Computer Methods and Programs in Biomedicine* 195 (105637), 113200K.

[Ström et al., 2020] Ström P., Kartasalo K., Olsson H., Solorzano L., Delahunt B., Berney D., Bostwick D., Evans A., Grignon D., Humphrey P., et al. (2020). "Artificial Intelligence for Diagnosis and Grading of Prostate Cancer in Biopsies: A Population-based, Diagnostic Study". In: *The Lancet Oncology* 21 (2), pp. 222–232.

[Sureka et al., 2020] Sureka M., Patil A., Anand D., and Sethi A. (2020). "Visualization for Histopathology Images using Graph Convolutional Neural Networks". In: *IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*. Vol. 11320, pp. 331–335.

[Teh et al., 2020] Teh E. and Taylor G. (2020). "Learning With Less Data via Weakly Labeled Patch Classification in Digital Pathology". In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, pp. 471–475.

[Thagaard et al., 2020] Thagaard J., Hauberg S., Vegt B. van der, Ebstrup T., Hansen J., and Dahl A. (2020). "Can You Trust Predictive Uncertainty under Real Dataset Shifts in Digital Pathology?" In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 824–833.

[Tschuchnig et al., 2020] Tschuchnig M., Oostingh G., and Gadermayr M. (2020). "Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential". In: *Patterns* 1 (6), p. 100089.

[Valkonen et al., 2020] Valkonen M., Isola J., Ylinen O., Muhonen V., Saxlin A., Tolonen T., Nykter M., and Ruusuvuori P. (2020). "Cytokeratin-Supervised Deep Learning for Automatic Recognition of Epithelial Cells in Breast Cancers Stained for ER, PR, and Ki-67". In: *IEEE Transactions on Medical Imaging* 39 (2), pp. 534–542.

[Vu et al., 2020] Vu M. and Thai M. (2020). "PGM-explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks". In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, 113200K.

[Wu et al., 2020] Wu Z., Pan S., Chen F., Long G., Zhang C., and Philip S. (2020). "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 32 (1), pp. 4–24.

[Yang et al., 2020] Yang L., Ghosh R., Franklin J., Chen S., You C., Narayan R., Melcher M., and Liphardt J. (2020). "NuSeT: A Deep Learning Tool for Reliably Separating and Analyzing Crowded Cells". In: *PLoS Computational Biology* 11320, 113200K.

[Yao et al., 2020] Yao J., Zhu X., Jonnagaddala J., Hawkins N., and Huang J. (2020). "Whole Slide Images Based Cancer Survival Prediction using Attention Guided Deep Multiple Instance Learning Networks". In: *Medical Image Analysis* 65, p. 101789.

[Yuan et al., 2020a] Yuan H., Yu H., Gui S., and Ji S. (2020a). "Explainability in Graph Neural Networks: A Taxonomic Survey". In: *arXiv:2012.15445* 11320, 113200K.

[Yuan et al., 2020b] Yuan H., Tang J., Hu X., and Ji S. (2020b). "XGNN: Towards Model-Level Explanations of Graph Neural Networks". In: *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Vol. 11320, pp. 430–438.

[Zhao et al., 2020a] Zhao Y., Yang F., Fang Y., Liu H., Zhou N., Zhang J., Sun J., Yang S., Menze B., Fan X., et al. (2020a). "Predicting Lymph Node Metastasis using Histopathological Images based on Multiple Instance Learning with Deep Graph Convolution". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 4837–4846.

[Zhao et al., 2020b] Zhao Y., Yang F., Fang Y., Liu H., Zhou N., Zhang J., Sun J., Yang S., Menze B., Fan X., and Yao J. (2020b). "Predicting Lymph Node Metastasis using Histopathological Images based on Multiple Instance Learning with Deep Graph Convolution". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 4837–4846.

[Zhou et al., 2020] Zhou J., Cui G., Hu S., Zhang Z., Yang C., Liu Z., Wang L., Li C., and Sun M. (2020). "Graph Neural Networks: A Review of Methods and Applications". In: *AI Open* 1, pp. 57–81.

[Ahn et al., 2019] Ahn J., Cho S., and Kwak S. (2019). "Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 2204–2213.

[Anand et al., 2019] Anand D., Gadiya S., and Sethi A. (2019). "Histographs: Graphs in Histopathology". In: *SPIE Medical Imaging: Digital Pathology*. Vol. 11320, 113200O.

[Aresta et al., 2019] Aresta G., Araújo T., Kwok S., Chennamsetty S., Safwan M., Alex V., Marami B., Prastawa M., Chan M., Donovan M., et al. (2019). "BACH: Grand Challenge on Breast Cancer Histology Images". In: *Medical image analysis* 56, pp. 122–139.

[Baldassarre et al., 2019] Baldassarre F. and Azizpour H. (2019). "Explainability Techniques for Graph Convolutional Networks". In: *International Conference on Machine Learning (ICML) Workshops*. Vol. 11320, 113200K.

[Bera et al., 2019] Bera K., Schalper K., Rimm D., Velcheti V., and Madabhushi A. (2019). "Artificial Intelligence in Digital Pathology — New Tools for Diagnosis and Precision Oncology". In: *Nature Reviews Clinical Oncology* 16, pp. 703–715.

[Binder et al., 2019] Binder T., Tantaoui E., Pati P., Catena R., Set-Aghayan A., and Gabrani M. (2019). "Multi-organ Gland Segmentation using Deep Learning". In: *Frontiers in Medicine* 6, p. 173.

[Bug et al., 2019] Bug D., Gräbel P., Feuerhake F., Oswald E., Schüler J., and Merhof D. (2019). "Supervised and Unsupervised Cell-Nuclei Detection in Immunohistology". In: 11320, 113200K.

[Byfield et al., 2019] Byfield P. et al. (2019). *StainTools*. `https://github.com/Peter554/StainTools`.

[Campanella et al., 2019] Campanella G., Hanna M., Geneslaw L., Miraflor A., Silva V., Busam K., Brogi E., Reuter V., Klimstra D., and Fuchs T. (2019). "Clinical-Grade Computational Pathology using Weakly Supervised Deep Learning on Whole Slide Images". In: *Nature Medicine* 25 (8), pp. 1301–1309.

[Carse et al., 2019] Carse J. and McKenna S. (2019). "Active Learning for Patch-Based Digital Pathology Using Convolutional Neural Networks to Reduce Annotation Costs". In: *European Congress on Digital Pathology*. Vol. 11320, pp. 20–27.

[Chan et al., 2019] Chan L., Hosseini M., Rowsell C., Plataniotis K., and Damaskinos S. (2019). "HistoSegNet: Semantic Segmentation of Histological Tissue Type in Whole Slide Images". In: *IEEE International Conference on Computer Vision (ICCV)*. Vol. 11320, pp. 10661–10670.

[Courtiol et al., 2019] Courtiol P., Maussion C., Moarii M., Pronier E., Pilcer S., Sefta M., Manceron P., Toldo S., Zaslavskiy M., Le Stang N., et al. (2019). "Deep Learning-based Classification of Mesothelioma Improves Prediction of Patient Outcome". In: *Nature Medicine* 25 (10), pp. 1519–1525.

[Dehmamy et al., 2019] Dehmamy N., Barabàsi A., and Yu R. (2019). "Understanding the Representation Power of Graph Neural Networks in Learning Graph Topology". In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, pp. 15413–15423.

[Dercksen et al., 2019] Dercksen K., Bulten W., and Litjens G. (2019). "Dealing with Label Scarcity in Computational Pathology: A Use Case in Prostate Cancer Classification". In: *SPIE Medical Imaging: Digital Pathology*. Vol. 11320, 113200K.

[Fort et al., 2019] Fort A., Hu H., and Lakshminarayanan B. (2019). "Deep Ensembles: A Loss Landscape Perspective". In: *Advances in Neural Information Processing Systems (NeurIPS)* 11320, 113200K.

[Foucart et al., 2019] Foucart A., Debeir O., and Decaestecker C. (2019). "SNOW: Semi-supervised, NOisy and/or Weak Data for Deep Learning in Digital Pathology". In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, pp. 1869–1872.

[Graham et al., 2019a] Graham S., Vu Q., Raza S., Azam A., Tsang Y., Kwak J., and Rajpoot N. (2019a). "Hover-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-tissue Histology Images". In: *Medical Image Analysis* 58, p. 101563.

[Graham et al., 2019b] Graham S., Chen H., Gamper J., Dou Q., Heng P., Snead D., Tsang Y., and Rajpoot N. (2019b). "MILD-Net: Minimal Information Loss Dilated Network for Gland Instance Segmentation in Colon Histology Images". In: *Medical Image Analysis* 52, pp. 199–211.

[Grunkin et al., 2019] Grunkin M. and Hansen J. (2019). *Assessment of Staining Quality*. US Patent 10,209,165.

[Hanna et al., 2019] Hanna M., Reuter V., Hameed M., Tan L., Chiang S., Sigel C., Hollmann T., Giri D., Samboy J., Moradel C., et al. (2019). "Whole Slide Imaging Equivalency and Efficiency Study: Experience at a Large Academic Center". In: *Modern Pathology* 32 (7), pp. 916–928.

[Hou et al., 2019] Hou L., Agarwal A., Samaras D., Kurc T., Gupta R., and Saltz J. (2019). "Robust Histopathology Image Analysis: To Label or to Synthesize?" In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 8533–8542.

[Johnson et al., 2019] Johnson J. and Khoshgoftaar T. (2019). "Survey on Deep Learning with Class Imbalance". In: *Journal of Big Data* 6 (27), 113200K.

[Kather et al., 2019] Kather J., Krisam J., Charoentong P., Luedde T., Herpel E., Weis C., Gaiser T., Marx A., Valous N., Ferber D., et al. (2019). "Predicting Survival from Colorectal Cancer Histology Slides using Deep Learning: A Retrospective Multicenter Study". In: *PLOS Medicine* 11320, 113200K.

[Kaya et al., 2019] Kaya M. and Bilge H. (2019). "Deep Metric Learning: A Survey". In: *Symmetry* 11, 113200K.

[Khan et al., 2019] Khan U., Stürenberg C., Gencoglu O., Sandeman K., Heikkinen T., Rannikko A., and Mirtti T. (2019). "Improving Prostate Cancer Detection with Breast Histopathology Images". In: *European Congress on Digital Pathology*. Vol. 11320, pp. 91–99.

[Li et al., 2019a] Li C., Wang X., Liu W., Latecki L., Wang B., and Huang J. (2019a). "Weakly Supervised Mitosis Detection in Breast Histopathology Images using Concentric Loss". In: *Medical Image Analysis* 53, pp. 165–178.

[Li et al., 2019b] Li X., Radulovic M., Kanjer K., and Plataniotis K. (2019b). "Discriminative Pattern Mining for Breast Cancer Histopathology Image Classification via Fully Convolutional Autoencoder". In: *IEEE Access* 7, pp. 36433–36445.

[Li et al., 2019c] Li X., Li W., and Tao R. (2019c). "Staged Detection–Identification Framework for Cell Nuclei in Histopathology Images". In: *IEEE Transactions on Instrumentation and Measurement* 69 (1), 113200K.

[Lutnick et al., 2019] Lutnick B., Ginley B., Govind D., McGarry S., LaViolette P., Yacoub R., Jain S., Tomaszewski J., Jen K., and Sarder P. (2019). "Iterative Annotation to Ease Neural Network Training: Specialized Machine Learning in Medical Image Analysis". In: *Nature Machine Intelligence* 11320, pp. 112–119.

[Mahmood et al., 2019] Mahmood F., Borders D., Chen R., McKay G., Salimian K., Baras A., and Durr N. (2019). "Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images". In: *IEEE Transactions on Medical Imaging* 39 (11), pp. 3257–3267.

[Mao et al., 2019] Mao X., Su Z., Tan P., Chow J., and Wang Y. (2019). "Is Discriminator a Good Feature Extractor?" In: *arXiv:1912.00789*. Vol. 11320, 113200K.

[Mercan et al., 2019a] Mercan C., Aksoy S., Mercan E., Shapiro L., Weaver D., and Elmore J. (2019a). "From Patch-level to ROI-level Deep Feature Representations for Breast Histopathology Classification". In: *SPIE Medical Imaging: Digital Pathology*. Vol. 10956, 109560H.

[Mercan et al., 2019b] Mercan E., Mehta S., Bartlett J., Shapiro L., Weaver D., and Elmore J. (2019b). "Assessment of Machine Learning of Breast Pathology Structures for Automated Differentiation of Breast Cancer and High-Risk Proliferative Lesions". In: *JAMA Network Open* 2 (8), 113200K.

[Muhammad et al., 2019] Muhammad H., Sigel C., Campanella G., Boerner T., Pak L., Büttner S., IJzermans J., Koerkamp B., Doukas M., Jarnagin W., et al. (2019). "Unsupervised Subtyping of Cholangiocarcinoma using a Deep Clustering Convolutional Autoencoder". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 604–612.

[Nagpal et al., 2019] Nagpal K., Foote D., Liu Y., Chen P., Wulczyn E., Tan F., Olson N., Smith J., Mohtashamian A., Wren J., et al. (2019). “Development and Validation of a Deep Learning Algorithm for Improving Gleason Scoring of Prostate Cancer”. In: *NPJ Digital Medicine* 2 (1), pp. 1–10.

[Nauhria et al., 2019] Nauhria S. and Hangfu L. (2019). “Virtual Microscopy Enhances the Reliability and Validity in Histopathology Curriculum: Practical Guidelines”. In: *MedEdPublish* 1, 113200K.

[Naylor et al., 2019] Naylor P., Lae M., Reyal F., and Walter T. (2019). “Segmentation of Nuclei in Histopathology Images by Deep Regression of the Distance Map”. In: *IEEE Transactions on Medical Imaging* 38 (2), pp. 448–459.

[Parwani, 2019] Parwani A. (2019). “Next Generation Diagnostic Pathology: Use of Digital Pathology and Artificial Intelligence Tools to Augment a Pathological Diagnosis”. In: *Diagnostic Pathology* 14 (138), 113200K.

[Paszke et al., 2019] Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., et al. (2019). “PyTorch: An Imperative Style, High-performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32, pp. 8026–8037.

[Pope et al., 2019] Pope P., Kolouri S., Rostami M., Martin C., and Hoffmann H. (2019). “Explainability Methods for Graph Convolutional Neural Networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 10764–10773.

[Qaiser et al., 2019a] Qaiser T., Tsang Y., Taniyama D., Sakamoto N., Nakane K., Epstein D., and Rajpoot N. (2019a). “Fast and Accurate Tumor Segmentation of Histology Images using Persistent Homology and Deep Convolutional Features”. In: *Medical Image Analysis* 55, pp. 1–14.

[Qaiser et al., 2019b] Qaiser T. and Rajpoot N. (2019b). “Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring”. In: *IEEE Transactions on Medical Imaging* 38 (11), pp. 2620–2631.

[Qu et al., 2019] Qu H., Wu P., Huang Q., Yi J., Riedlinger G., De S., and Metaxas D. (2019). “Weakly Supervised Deep Nuclei Segmentation using Points Annotation in Histopathology Images”. In: *Medical Imaging with Deep Learning (MIDL)*. Vol. 11320, pp. 390–400.

[Raghu et al., 2019] Raghu M., Zhang C., Kleinberg J., and Bengio S. (2019). “Transfusion: Understanding Transfer Learning for Medical Imaging”. In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, 113200K.

[Ren et al., 2019] Ren J., Hacihaliloglu I., Singer E., Foran D., and Qi X. (2019). “Unsupervised Domain Adaptation for Classification of Histopathology Whole-Slide Images”. In: *Frontiers in Bioengineering and Biotechnology* 7, p. 102.

[Roy et al., 2019] Roy K., Banik D., Bhattacharjee D., and Nasipuri M. (2019). “Patch-Based System for Classification of Breast Histology Images using Deep Learning”. In: *Computerized Medical Imaging and Graphics* 71, pp. 90–103.

[Schwarzenberg et al., 2019] Schwarzenberg R., Huebner M., Harbecke D., Alt C., and Hennig L. (2019). “Layerwise Relevance Visualization in Convolutional Text Graph Classifiers”. In: *Empirical Methods in Natural Language Processing (EMNLP) Workshops* 11320, pp. 58–62.

[Serag et al., 2019] Serag A., Ion-Margineanu A., Qureshi H., McMillan R., Martin M., Diamond J., O’Reilly P., and Hamilton P. (2019). “Translational AI and Deep Learning in Diagnostic Pathology”. In: *Frontiers in Medicine* 6, p. 185.

[Shaban et al., 2019] Shaban M., Khurram S., Fraz M., Alsubaie N., Masood I., Mushtaq S., Hassan M., Loya A., and Rajpoot N. (2019). “A Novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes Predicts Disease Free Survival in Oral Squamous Cell Carcinoma”. In: *Scientific Reports* 9 (1), pp. 1–13.

[Ström et al., 2019] Ström P., Kartasalo K., Olsson H., Solorzano L., Delahunt B., Berney D., Bostwick D., Evans A., Grignon D., Humphrey P., et al. (2019). “Pathologist-level Grading of Prostate Biopsies with Artificial Intelligence”. In: *Bioinformatics* 11320, 113200K.

[Tabibu et al., 2019] Tabibu S., Vinod P., and Jawahar C. (2019). “Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning”. In: *Scientific Reports* 9 (1), pp. 1–9.

[Tan et al., 2019] Tan P., Ellis I., Allison K., Brogi E., Fox S., Lakhani S., Lazar A., Morris E., Sahin A., Salgado R., et al. (2019). “The 2019 World Health Organization Classification of Tumours of the Breast”. In: *International Agency for Research on Cancer* 11320, 113200K.

[Tellez et al., 2019] Tellez D., Litjens G., Bandi P., Bulten W., Bokhorst J., Ciompi F., and Laak J. (2019). “Quantifying the Effects of Data Augmentation and Stain Color Normalization in Convolutional Neural Networks for Computational Pathology”. In: *Medical Image Analysis* 58, 113200K.

[Tokunaga et al., 2019] Tokunaga H., Teramoto Y., Yoshizawa A., and Bise R. (2019). “Adaptive Weighting Multi-Field-Of-View CNN for Semantic Segmentation in Pathology”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 12597–12606.

[Tschannen et al., 2019] Tschannen M., Djolonga J., Rubenstein P., Gelly S., and Lucic M. (2019). “On Mutual Information Maximization for Representation Learning”. In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[Veta et al., 2019] Veta M., Heng Y., Stathonikos N., Bejnordi B., Beca F., Wollmann T., Rohr K., Shah M., Wang D., Rousson M., et al. (2019). “Predicting Breast Tumor Proliferation from Whole-Slide Images: The TUPAC16 Challenge”. In: *Medical Image Analysis* 54, pp. 111–121.

[Wang et al., 2019a] Wang J., Chen R., Lu M., Baras A., and Mahmood F. (2019a). “Weakly Supervised Prostate TMA Classification via Graph Convolutional Networks”. In: *IEEE International Symposium on Biomedical Imaging (ISBI)* 11320, pp. 239–243.

[Wang et al., 2019b] Wang M., Yu L., Zheng D., Gan Q., Gai Y., Ye Z., Li M., Zhou J., Huang Q., Ma C., et al. (2019b). “Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs”. In: *SPIE Medical Imaging: Digital Pathology*. Vol. 11320, 113200K.

[Wang et al., 2019c] Wang S., Yang D., Rong R., Zhan X., and Xiao G. (2019c). “Pathology Image Analysis Using Segmentation Deep Learning Algorithms”. In: *The American Journal of Pathology* 189 (9), pp. 1686–1698.

[Wang et al., 2019d] Wang S., Zhu Y., Yu L., Chen H., Lin H., Wan X., Fan X., and Heng P. (2019d). “RMDL: Recalibrated Multi-instance Deep Learning for Whole Slide Gastric Image Classification”. In: *Medical Image Analysis* 58, p. 101549.

[Wang et al., 2019e] Wang X., Chen H., Gan C., Lin H., Dou Q., Tsougenis E., Huang Q., Cai M., and Heng P. (2019e). “Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis”. In: *IEEE Transactions on Cybernetics* 50 (9), pp. 3950–3962.

[Wei et al., 2019] Wei J., Tafe L., Linnik Y., Vaickus L., Tomita N., and Hassanpour S. (2019). “Pathologist-level Classification of Histologic Patterns on Resected Lung Adenocarcinoma Slides with Deep Neural Networks”. In: *Scientific Reports* 9 (1), pp. 1–8.

[Wu et al., 2019] Wu J., Zhong J., Chen E., Zhang J., Jay J., and Yu L. (2019). “Weakly-and Semi-Supervised Graph CNN for Identifying Basal Cell Carcinoma on Pathological Images”. In: *International Workshop on Graph Learning in Medical Imaging*. Vol. 11320, pp. 112–119.

[Xie et al., 2019] Xie J., Liu R., Luttrell J., and Zhang C. (2019). “Deep Learning based Analysis of Histopathological Images of Breast Cancer”. In: *Frontiers in Genetics* 10, p. 80.

[Xing et al., 2019] Xing F., Xie Y., Shi X., Chen P., Zhang Z., and Yang L. (2019). “Towards Pixel-to-pixel Deep Nucleus Detection in Microscopy Images”. In: *BMC Bioinformatics* 20 (1), pp. 1–16.

[Xu et al., 2019a] Xu G., Song Z., Sun Z., Ku C., Yang Z., Liu C., Wang S., Ma J., and Xu W. (2019a). “CAMEL: A Weakly Supervised Learning Framework for Histopathology Image Segmentation”. In: *IEEE International Conference on Computer Vision (ICCV)*. Vol. 11320, pp. 10681–10690.

[Xu et al., 2019b] Xu K., Hu W., Leskovec J., and Jegelka S. (2019b). “How Powerful are Graph Neural Networks?” In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[Yamamoto et al., 2019] Yamamoto Y., Tsuzuki T., Akatsuka J., Ueki M., Morikawa H., Numata Y., Takahara T., Tsuyuki T., Tsutsumi K., Nakazawa R., et al. (2019). “Automated Acquisition of Explainable Knowledge from Unannotated Histopathology Images”. In: *Nature Communications* 10 (1), pp. 1–9.

[Yan et al., 2019] Yan C., Cai C., Xie J., Fu Y., Shuai H., Fan X., and Xu J. (2019). “Prior Consistent CNN with Multi-Task Learning for Colon Image Classification”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshops*. Vol. 11320, 113200K.

[Ye et al., 2019] Ye H., Wang D., Li J., Zhu S., and Zhu C. (2019). “Improving Histopathological Image Segmentation and Classification using Graph Convolution Network”. In: *International Conference on Computing and Pattern Recognition*. Vol. 11320, pp. 192–198.

[Yi et al., 2019] Yi X., Walia E., and Babyn P. (2019). “Generative Adversarial Network in Medical Imaging: A Review”. In: *Medical Image Analysis* 58 (101552), 113200K.

[Ying et al., 2019] Ying R., Bourgeois D., You J., Zitnik M., and Leskovec J. (2019). “GNNExplainer: Generating Explanations for Graph Neural Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32, p. 9240.

[Zhang et al., 2019] Zhang Z., Chen P., McGough M., Xing F., and Wang C. (2019). “Pathologist-level Interpretable Whole-Slide Cancer Diagnosis with Deep Learning”. In: *Nature Machine Intelligence* 1, pp. 236–245.

[Zheng et al., 2019] Zheng Y., Jiang B., Shi J., Zhang H., and Xie F. (2019). “Encoding Histopathological WSIs using GNN for Scalable Diagnostically Relevant Regions Retrieval”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 550–558.

[Zhou et al., 2019a] Zhou Y., Graham S., Koohbanani N., Shaban M., Heng P., and Rajpoot N. (2019a). “CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images”. In: *IEEE International Conference on Computer Vision (ICCV) Workshops*. Vol. 11320, 113200K.

[Zhou et al., 2019b] Zhou Y., Onder O., Dou Q., Tsougenis E., Chen H., and Heng P. (2019b). "CIA-Net: Robust Nuclei Instance Segmentation with Contour-Aware Information Aggregation". In: *Information Processing in Medical Imaging (IPMI)*. Vol. 11320, pp. 682–693.

[Akbar et al., 2018] Akbar S. and Martel A. (2018). "Cluster-based Learning from Weakly Labeled Bags in Digital Pathology". In: *arXiv:1812.00884*. Vol. 11320, 113200K.

[Akram et al., 2018] Akram S., Qaiser T., Graham S., Kannala J., Heikkilä J., and Rajpoot N. (2018). "Leveraging Unlabeled Whole-Slide-Images for Mitosis Detection". In: *Computational Pathology and Ophthalmic Medical Image Analysis*. Vol. 11320, pp. 69–77.

[Alialy et al., 2018] Alialy R., Tavakkol S., Tavakkol E., Ghorbani-Aghbologhi A., Ghaffarieh A., Kim S., and Shahabi C. (2018). "A Review on the Applications of Crowd Sourcing in Human Pathology". In: *Journal of Pathology Informatics* 9 (2), 113200K.

[Aresta et al., 2018] Aresta G., Araújo T., Eloy C., Polònia A., and Aguiar P. (2018). *ICIAR 2018 Grand Challenge on Breast Cancer Histology images*. `https://iciar2018-challenge.grand-challenge.org/Home/`.

[Arvaniti et al., 2018] Arvaniti E., Fricker K., Moret M., Rupp N., Hermanns T., Fankhauser C., Wey N., Wild P., Rüschoff J., and Claassen M. (2018). "Automated Gleason Grading of Prostate Cancer Tissue Microarrays via Deep Learning". In: *Scientific Reports* 8 (12054), 113200K.

[Awan et al., 2018] Awan R. and Rajpoot N. (2018). "Deep Autoencoder Features for Registration of Histology Images". In: *Annual Conference on Medical Image Understanding and Analysis*. Vol. 11320, pp. 371–378.

[Bandi et al., 2018] Bandi P., Geessink O., Manson Q., Van Dijk M., Balkenhol M., Hermsen M., Bejnordi B., Lee B., Paeng K., Zhong A., et al. (2018). "From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The Camelyon17 Challenge". In: *IEEE Transactions on Medical Imaging* 38 (2), pp. 550–560.

[Bardou et al., 2018] Bardou D., Zhang K., and Ahmad S. (2018). "Classification of Breast Cancer Based on Histology Images using Convolutional Neural Networks". In: *IEEE Access* 6 (1), pp. 24680–24693.

[BenTaieb et al., 2018] BenTaieb A. and Hamarneh G. (2018). "Predicting Cancer with a Recurrent Visual Attention Model for Histopathology Images". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 129–137.

[Binder et al., 2018] Binder A., Bockmayr M., Hagele M., Wienert S., Heim D., Hellweg K., Stenzinger A., Parlow L., Budczies J., Goeppert B., et al. (2018). "Towards Computational Fluorescence Microscopy: Machine Learning-based Integrated Prediction of Morphological and Molecular Tumor Profiles". In: *arXiv:1805.11178* 11320, 113200K.

[Bokhorst et al., 2018] Bokhorst J., Pinckaers H., van Zwam P., Nagtegaal I., van der Laak J., and Ciompi F. (2018). "Learning from Sparsely Annotated Data for Semantic Segmentation in Histopathology Images". In: *Medical Imaging with Deep Learning (MIDL)*. Vol. 11320, 113200K.

[Brancati et al., 2018] Brancati N., Frucci M., and Riccio D. (2018). "Multi-classification of Breast Cancer Histology Images by Using a Fine-Tuning Strategy". In: *International Conference Image Analysis and Recognition (ICIAR)*. Vol. 11320, pp. 771–778.

[Bychkov et al., 2018] Bychkov D., Linder N., Turkki R., Nordling S., Kovanen P., Verrill C., Walliander M., Lundin M., Haglund C., and Lundin J. (2018). "Deep Learning based Tissue Analysis Predicts Outcome in Colorectal Cancer". In: *Scientific Reports* 8 (3395), 113200K.

[Chattopadhay et al., 2018] Chattopadhay A., Sarkar A., Howlader P., and Balasubramanian V. (2018). "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Vol. 11320, pp. 839–847.

[Chen et al., 2018] Chen R., Li X., Grosse R., and Duvenaud D. (2018). "Isolating Sources of Disentanglement in Variational Autoencoders". In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, 113200K.

[Chennamsetty et al., 2018] Chennamsetty S., Safwan M., and Alex V. (2018). "Classification of Breast Cancer Histology Image using Ensemble of Pre-Trained Neural Networks". In: *International Conference Image Analysis and Recognition (ICIAR)*. Vol. 11320, pp. 804–811.

[Coudray et al., 2018] Coudray N., Ocampo P., Sakellaropoulos T., Narula N., Snuderl M., Fenyö D., Moreira A., Razavian N., and Tsirigos A. (2018). "Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning". In: *Nature Medicine* 24 (10), pp. 1559–1567.

[Couture et al., 2018] Couture H., Williams L., Geradts J., Nyante S., Butler E., Marron J., Perou C., Troester M., and Niethammer M. (2018). "Image Analysis with Deep Learning to Predict Breast Cancer Grade, ER Status, Histologic Subtype, and Intrinsic Subtype". In: *NPJ Breast Cancer* 4 (1), pp. 1–8.

[Dong et al., 2018] Dong N., Kampffmeyer M., Liang X., Wang Z., Dai W., and Xing E. (2018). "Reinforced Auto-Zoom Net: Towards Accurate and Fast Breast Cancer Segmentation in Whole-Slide Images". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Vol. 11320, pp. 317–325.

[Finlayson et al., 2018] Finlayson S., Lee H., Kohane I., and Oakden-Rayner L. (2018). “Towards Generative Adversarial Networks as a New Paradigm for Radiology Education”. In: *arXiv:1812.01547* 11320, 113200K.

[Frid-Adar et al., 2018] Frid-Adar M., Diamant I., Klang E., Amitai M., Goldberger J., and Greenspan H. (2018). “GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification”. In: *Neurocomputing* 321, pp. 321–331.

[Gadermayr et al., 2018] Gadermayr M., Gupta L., Klinkhammer B., Boor P., and Merhof D. (2018). “Unsupervisedly Training Gans for Segmenting Digital Pathology with Automatically Generated Annotations”. In: *Medical Imaging with Deep Learning (MIDL)*. Vol. 11320, 113200K.

[Graziani et al., 2018] Graziani M., Andrearczyk V., and Müller H. (2018). “Regression Concept Vectors for Bidirectional Explanations in Histopathology”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshops*. Vol. 11320, pp. 124–132.

[Hamad et al., 2018] Hamad A., Ersoy I., and Bunyak F. (2018). “Improving Nuclei Classification Performance in H&E Stained Tissue Images Using Fully Convolutional Regression Network and Convolutional Neural Network”. In: *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. Vol. 11320, 113200K.

[Höfener et al., 2018] Höfener H., Homeyer A., Weiss N., Molin J., Lundström C., and Hahn H. (2018). “Deep Learning Nuclei Detection: A Simple Approach Can Deliver State-of-the-art Results”. In: *Computerized Medical Imaging and Graphics* 70, pp. 43–52.

[Hoffman et al., 2018] Hoffman R., Mueller S., Klein G., and Litman J. (2018). “Metrics for Explainable AI: Challenges and Prospects”. In: *arXiv:1812.04608*. Vol. 11320, 113200K.

[Ilse et al., 2018] Ilse M., Tomczak J., and Welling M. (2018). “Attention-based Deep Multiple Instance Learning”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, pp. 2127–2136.

[Jampani et al., 2018] Jampani V., Sun D., Liu M., Yang M., and Kautz J. (2018). “Superpixel Sampling Networks”. In: *European Conference on Computer Vision (ECCV)*. Vol. 11320, pp. 352–368.

[Kashyap et al., 2018] Kashyap A., Jain M., Shukla S., and Andley M. (2018). “Role of Nuclear Morphometry in Breast Cancer and its Correlation with Cytomorphological Grading of Breast Cancer: A Study of 64 Cases”. In: *Journal of Cytology* 35 (1), pp. 41–45.

[Kim et al., 2018] Kim B., Wattenberg M., Gilmer J., Cai C., Wexler J., Viegas F., and Sayres R. (2018). “Interpretability beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. In: *International Conference on Machine Learning (ICML)* 11320, pp. 2668–2677.

[Komura et al., 2018] Komura D. and Ishikawa S. (2018). “Machine Learning Methods for Histopathological Image Analysis”. In: *Computational and Structural Biotechnology Journal* 11320, pp. 34–42.

[Kumar et al., 2018] Kumar A., Sarawagi S., and Jain U. (2018). “Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, pp. 2805–2814.

[Li et al., 2018a] Li C., Wang X., Liu W., and Latecki L. (2018a). “DeepMitosis: Mitosis Detection via Deep Detection, Verification and Segmentation Networks”. In: *Medical Image Analysis* 45, pp. 121–133.

[Li et al., 2018b] Li J., Speier W., Ho K., Sarma K., Gertych A., Knudsen B., and Arnold C. (2018b). “An EM-based Semi-supervised Deep Learning Approach for Semantic Segmentation of Histopathological Images from Radical Prostatectomies”. In: *Computerized Medical Imaging and Graphics* 69, pp. 125–133.

[Li et al., 2018c] Li R., Yao J., Zhu X., Li Y., and Huang J. (2018c). “Graph CNN for Survival Analysis on Whole Slide Pathological Images”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 174–182.

[Li et al., 2018d] Li Y., Tian X., Shen X., and Tao D. (2018d). “Classification and Representation Joint Learning via Deep Networks”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Vol. 11320, pp. 2215–2221.

[Li et al., 2018e] Li Z., Hu Z., Xu J., Tan T., Chen H., Duan Z., Liu P., Tang J., Cai G., Ouyang Q., et al. (2018e). “Computer-aided Diagnosis of Lung Carcinoma using Deep Learning: A Pilot Study”. In: *arXiv:1803.05471* 11320, 113200K.

[Liang et al., 2018] Liang Q., Nan Y., Coppola G., Zou K., Sun W., Zhang D., Wang Y., and Yu G. (2018). “Weakly Supervised Biomedical Image Segmentation by Reiterative Learning”. In: *IEEE Journal of Biomedical and Health Informatics* 23 (3), pp. 1205–1214.

[Marami et al., 2018] Marami B., Prastawa M., Chan M., Donovan M., Fernandez G., and Zeineh J. (2018). “Ensemble Network for Region Identification in Breast Histopathology Slides”. In: *International Conference Image Analysis and Recognition (ICIAR)*. Vol. 11320, pp. 861–868.

[Mehta et al., 2018] Mehta S., Mercan E., Bartlett J., Weaver D., Elmore J., and Shapiro L. (2018). “Learning to Segment Breast Biopsy Whole Slide Images”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Vol. 11320, pp. 663–672.

[Mercan et al., 2018] Mercan E., Mehta S., Bartlett J., Weaver D., Elmore J., and Shapiro L. (2018). “Automated Diagnosis of Breast Cancer and Pre-invasive Lesions on Digital Whole Slide Images”. In: *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. Vol. 11320, pp. 60–68.

[Mobadersany et al., 2018] Mobadersany P., Yousefi S., Amgad M., Gutman D., Barnholtz-Sloan J., Vega J., Brat D., and Cooper L. (2018). “Predicting Cancer Outcomes from Histology and Genomics using Convolutional Networks”. In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 115 (13), pp. 2970–2979.

[Morris et al., 2018] Morris C., Ritzert M., Fey M., Hamilton W., Lenssen J., Rattan G., and Grohe M. (2018). “Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks”. In: *Association for the Advancement of Artificial Intelligence (AAAI)*. Vol. 33, pp. 4602–4609.

[Pantanowitz et al., 2018] Pantanowitz L., Sharma A., Carter A., Kurc T., Sussman A., and Saltz J. (2018). “Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives”. In: *Journal of Pathology Informatics* 9 (40), 113200K.

[Pati et al., 2018] Pati P., Andani S., Pediaditis M., Viana M., Ruschoff J., Wild P., and Gabrani M. (2018). “Deep Positive-Unlabeled Learning for Region of Interest Localization in Breast Tissue Images”. In: *SPIE Medical Imaging: Digital Pathology*. Vol. 10581, p. 1058107.

[Peikari et al., 2018] Peikari M., Salama S., Nofech-Mozes S., and Martel A. (2018). “A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification”. In: *Scientific Reports* 8 (7193), 113200K.

[Qaiser et al., 2018] Qaiser T., Mukherjee A., Reddy Pb C., Munugoti S., Tallam V., Pitk"aaho T., Lehtim"aki T., Naughton T., Berseth M., Pedraza A., et al. (2018). “Her2 Challenge Contest: A Detailed Assessment of Automated Her2 Scoring Algorithms in Whole Slide Images of Breast Cancer Tissues”. In: *Histopathology* 72 (2), pp. 227–238.

[Rajbongshi et al., 2018] Rajbongshi N., Bora K., Nath D., Das A., and Mahanta L. (2018). “Analysis of Morphological Features of Benign and Malignant Breast Cell Extracted From FNAC Microscopic Image Using the Pearsonian System of Curves”. In: *Journal of Cytology* 35 (2), pp. 99–104.

[Sagun et al., 2018] Sagun L. and Arias R. (2018). “Digital Pathology: An Innovative Approach to Medical Education”. In: *Philippine Journal of Pathology* 11320, 113200K.

[Sandler et al., 2018] Sandler M., Howard A., Zhu M., Zhmoginov A., and Chen L. (2018). “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 4510–4520.

[Schwaller et al., 2018] Schwaller P., Gaudin T., Lanyi D., Bekas C., and Laino T. (2018). ““Found in Translation”: Predicting Outcomes of Complex Organic Chemistry Reactions using Neural Sequence-to-sequence Models”. In: *Chemical Science* 9 (28), pp. 6091–6098.

[Shao et al., 2018] Shao W., Sun L., and Zhang D. (2018). “Deep Activate Learning for Nuclei Classification in Pathology Images”. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, pp. 199–202.

[Sharma et al., 2018] Sharma M. and Miyamoto H. (2018). “Percent Gleason Pattern 4 in Stratifying the Prognosis of Patients with Intermediate-risk Prostate Cancer”. In: *Translational Andrology and Urology* 7 (4), 113200K.

[Sirinukunwattana et al., 2018] Sirinukunwattana K., Alham N., Verrill C., and Rittscher J. (2018). “Improving Whole Slide Segmentation through Visual Context - A Systematic Study”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, 113200K.

[Sornapudi et al., 2018] Sornapudi S., Stanley R., Stoecker W., Almubarak H., Long R., Antani S., Thoma G., Zuna R., and Frazier S. (2018). “Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels”. In: *Journal of Pathology Informatics* 9, 113200K.

[Stanisavljevic et al., 2018] Stanisavljevic M., Anghel A., Papandreou N., Andani S., Pati P., Rüschoff J., Wild P., Gabrani M., and Pozidis H. (2018). “A Fast and Scalable Pipeline for Stain Normalization of Whole-Slide Images in Histopathology”. In: *European Conference on Computer Vision (ECCV) Workshops*. Vol. 11320, pp. 424–436.

[Tellez et al., 2018] Tellez D., Balkenhol M., Karssemeijer N., Litjens G., van der Laak J., and Ciompi F. (2018). “H and E Stain Augmentation Improves Generalization of Convolutional Networks for Histopathological Mitosis Detection”. In: *SPIE Medical Imaging: Digital Pathology*. Vol. 10581, 105810Z.

[Tizhoosh et al., 2018] Tizhoosh H. and Pantanowitz L. (2018). “Artificial Intelligence and Digital Pathology: Challenges and Opportunities”. In: *Journal of Pathology Informatics* 38 (9), 113200K.

[Veličković et al., 2018] Veličković P., Cucurull G., Casanova A., Romero A., Liò P., and Bengio Y. (2018). “Graph Attention Networks”. In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[Wen et al., 2018] Wen S., Kurc T., Hou L., Saltz J., Gupta R., Batiste R., Zhao T., Nguyen V., Samaras D., and Zhu W. (2018). “Comparison of Different Classifiers with Active Learning to Support Quality Control in Nucleus Segmentation in Pathology Images”. In: *AMIA Summits on Translational Science Proceedings* 2018, p. 227.

[Wilson et al., 2018] Wilson M., Fleming K., Kuti M., Looi L., Lago N., and Ru K. (2018). “Access to Pathology and Laboratory Medicine Services: A Crucial Gap”. In: *The Lancet* 391 (10133), pp. 1927–1938.

[Xie et al., 2018] Xie Y., Xing F., Shi X., Kong X., Su H., and Yang L. (2018). “Efficient and Robust Cell Detection: A Structured Regression Approach”. In: *Medical Image Analysis* 44, pp. 245–254.

[Xu et al., 2018] Xu K., Li C., Tian Y., Sonobe T., Kawarabayashi K., and Jegelka S. (2018). “Representation Learning on Graphs with Jumping Knowledge Networks”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, 113200K.

[Ying et al., 2018] Ying R., Morris C., Hamilton W., You J., Ren X., and Leskovec J. (2018). “Hierarchical Graph Representation Learning with Differentiable Pooling”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 11320, pp. 4800–4810.

[Zhou et al., 2018] Zhou Y., Dou Q., Chen H., Qin J., and Heng P. (2018). “SFCN-OPI: Detection and Fine-Grained Classification of Nuclei Using Sibling FCN with Objectness Prior Interaction”. In: *Association for the Advancement of Artificial Intelligence (AAAI)*. Vol. 32. 1, 113200K.

[Arar et al., 2017] Arar N., Pati P., Kashyap A., Fomitcheva A., Goksel O., Kaigala G., and Gabrani M. (2017). “Computational Immunohistochemistry: Recipes for Standardization of Immunostaining”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 48–55.

[Bankhead et al., 2017] Bankhead P., Loughrey M., Fernàndez J., Dombrowski Y., McArt D., Dunne P., McQuaid S., Gray R., Murray L., Coleman H., James J., Salto-Tellez M., and Hamilton P. (2017). “QuPath: Open Source Software for Digital Pathology Image Analysis”. In: *Scientific Reports* 7 (1), pp. 1–7.

[Bejnordi et al., 2017a] Bejnordi B., Zuidhof G., Balkenhol M., Hermsen M., Bult P., van Ginneken B., Karssemeijer N., Litjens G., and van der Laak J. (2017a). “Context-aware Stacked Convolutional Neural Networks for Classification of Breast Carcinomas in Whole-Slide Histopathology Images”. In: *Journal of Medical Imaging* 4 (4), 113200K.

[Bejnordi et al., 2017b] Bejnordi B., Veta M., Van Diest P., Van Ginneken B., Karssemeijer N., Litjens G., Van Der Laak J., Hermsen M., Manson Q., Balkenhol M., et al. (2017b). “Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer”. In: *JAMA* 318 (22), pp. 2199–2210.

[Chang et al., 2017] Chang H., Learned-Miller E., and McCallum A. (2017). “Active Bias: Training More Accurate Neural Networks by Emphasizing High Variance Samples”. In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, 113200K.

[Cruz-Roa et al., 2017] Cruz-Roa A., Gilmore H., Basavanhally A., Feldman M., Ganesan S., Shih N., Tomaszewski J., González F., and Madabhushi A. (2017). “Accurate and Reproducible Invasive Breast Cancer Detection in Whole-Slide Images: A Deep Learning Approach for Quantifying Tumor Extent”. In: *Scientific Reports* 7 (1), pp. 1–14.

[Dhurandhar et al., 2017] Dhurandhar A., Iyengar V., Luss R., and Shanmugam K. (2017). “A Formal Framework to Characterize Interpretability of Procedures”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, 113200K.

[Donahue et al., 2017] Donahue J., Krähenbühl P., and Darrell T. (2017). “Adversarial Feature Learning”. In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[Doshi-Velez et al., 2017] Doshi-Velez F. and Kim B. (2017). “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv:1702.08608*. Vol. 11320, 113200K.

[Garcià-Arteaga et al., 2017] Garcià-Arteaga J., Corredor G., Wang X., Velcheti V., Madabhushi A., and Romero E. (2017). “A Lymphocyte Spatial Distribution Graph-Based Method for Automated Classification of Recurrence Risk on Lung Cancer Images”. In: *International Symposium on Medical Information Processing and Analysis*. Vol. 10572, 113200K.

[Gilmer et al., 2017] Gilmer J., Schoenholz S., Riley P., Vinyals O., and Dahl G. (2017). “Neural Message Passing for Quantum Chemistry”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, pp. 1263–1272.

[Guo et al., 2017] Guo C., Pleiss G., Sun Y., and Weinberger K. (2017). “On Calibration of Modern Neural Networks”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, pp. 1321–1330.

[Hamilton et al., 2017] Hamilton W., Ying Z., and Leskovec J. (2017). “Inductive Representation Learning on Large Graphs”. In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, pp. 1024–1034.

[Han et al., 2017] Han Z., Wei B., Zheng Y., Yin Y., Li K., and Li S. (2017). “Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model”. In: *Scientific Reports* 7 (4172), 113200K.

[Hermans et al., 2017] Hermans A., Beyer L., and Leibe B. (2017). “In Defense of the Triplet Loss for Person Re-Identification”. In: *arXiv:1703.07737v4*. Vol. 11320, 113200K.

[Holzinger et al., 2017] Holzinger A., Malle B., Kieseberg P., Roth P., Müller H., Reihs R., and Zatloukal K. (2017). "Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology". In: *arXiv:1712.06657*. Vol. 11320, 113200K.

[Huang et al., 2017] Huang G., Liu Z., Maaten L., and Weinberger K. (2017). "Densely Connected Convolutional Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 2261–2269.

[Jia et al., 2017] Jia Z., Huang X., Chang E., and Xu Y. (2017). "Constrained Deep Weak Supervision for Histopathology Image Segmentation". In: *IEEE Transactions on Medical Imaging* 36 (11), pp. 2376–2388.

[Kainz et al., 2017] Kainz P., Pfeiffer M., and Urschler M. (2017). "Segmentation and Classification of Colon Glands with Deep Convolutional Neural Networks and Total Variation Regularization". In: *PeerJ* 5, 113200K.

[Kendall et al., 2017] Kendall A. and Yarin G. (2017). "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, 113200K.

[Kipf et al., 2017] Kipf T. and Welling M. (2017). "Semi-Supervised Classification with Graph Convolutional Networks". In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[Korbar et al., 2017] Korbar B., Olofson A., Miraflor A., Nicka C., Suriawinata M., Torresani L., Suriawinata A., and Hassanpour S. (2017). "Looking under the Hood Deep Neural Network Visualization to Interpret Whole Slide Image Analysis Outcomes for Colorectal Polyps". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Vol. 11320, pp. 69–75.

[Kumar et al., 2017] Kumar N., Verma R., Sharma S., Bhargava S., Vahadane A., and Sethi A. (2017). "A Dataset and A Technique for Generalized Nuclear Segmentation for Computational Pathology". In: *IEEE Transactions on Medical Imaging* 36 (7), pp. 1550–1560.

[Lakshminarayanan et al., 2017] Lakshminarayanan B., Pritzel A., and Blundell C. (2017). "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 30, 113200K.

[Litjens et al., 2017] Litjens G., Kooi T., Bejnordi B., Setio A., Ciompi F., Ghafoorian M., Van Der Laak J., Van Ginneken B., and Sánchez C. (2017). "A Survey on Deep Learning in Medical Image Analysis". In: *Medical Image Analysis* 42, pp. 60–88.

[Liu et al., 2017a] Liu X., Kumar B., You J., and Jia P. (2017a). "Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Vol. 11320, pp. 522–531.

[Liu et al., 2017b] Liu Y., Gadepalli K., Norouzi M., Dahl G., Kohlberger T., Boyko A., Venugopalan S., Timofeev A., Nelson P., Corrado G., et al. (2017b). "Detecting Cancer Metastases on Gigapixel Pathology Images". In: *arXiv:1703.02442*. Vol. 11320, 113200K.

[Mukhopadhyay et al., 2017] Mukhopadhyay S., Feldman M., Abels E., Ashfaq R., Beltaifa S., Cacciabeve N., Cathro H., Cheng L., Cooper K., Dickey G., et al. (2017). "Whole Slide Imaging versus Microscopy for Primary Diagnosis in Surgical Pathology: A Multicenter Blinded Randomized Noninferiority Study of 1992 Cases (Pivotal Study)". In: *The American Journal of Surgical Pathology* 42 (1), pp. 39–52.

[Nguyen et al., 2017] Nguyen L., Tosun A., Fine J., Taylor D., and Chennubhotla S. (2017). "Architectural Patterns for Differential Diagnosis of Proliferative Breast Lesions from Histopathological Images". In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, pp. 152–155.

[Samek et al., 2017] Samek W., Binder A., Montavon G., Lapuschkin S., and Muller K. (2017). "Evaluating the Visualization of What a Deep Neural Network has Learned". In: *IEEE Transactions on Neural Networks and Learning Systems* 28, pp. 2660–2673.

[Schlegl et al., 2017] Schlegl T., Seeböck P., Waldstein S., Schmidt-Erfurth U., and Langs G. (2017). "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery". In: *International Conference on Information Processing in Medical Imaging (IPMI)*. Vol. 11320, pp. 146–157.

[Selvaraju et al., 2017] Selvaraju R., Cogswell M., Das A., Vedantam R., and Batra D. (2017). "Grad-CAM : Visual Explanations from Deep Networks". In: *International Conference on Computer Vision (ICCV)*. Vol. 11320, pp. 618–626.

[Sharma et al., 2017a] Sharma H., Zerbe N., Boger C., Wienert S., Hellwich O., and Hufnagl P. (2017a). "A Comparative Study of Cell Nuclei Attributed Relational Graphs for Knowledge Description and Categorization in Histopathological Gastric Cancer Whole Slide Images". In: *IEEE Symposium on Computer-Based Medical Systems (CBMS)*. Vol. 11320, pp. 61–66.

[Sharma et al., 2017b] Sharma H., Zerbe N., Klempert I., Hellwich O., and Hufnagl P. (2017b). "Deep Convolutional Neural Networks for Automatic Classification of Gastric Carcinoma using Whole Slide Images in Digital Histopathology". In: *Computerized Medical Imaging and Graphics* 61, pp. 2–13.

[Sirinukunwattana et al., 2017] Sirinukunwattana K., Pluim J., Chen H., Qi X., Heng P., Guo Y., Wang L., Matuszewski B., Bruni E., Sanchez U., et al. (2017). "Gland Segmentation in Colon Histology Images: The Glas Challenge Contest". In: 35, pp. 489–502.

[Vandenberghe et al., 2017] Vandenberghe M., Scott M., Scorer P., Söderberg M., Balcerzak D., and Barker C. (2017). "Relevance of Deep Learning to Facilitate the Diagnosis of HER2 Status in Breast Cancer". In: *Scientific Reports* 7 (1), pp. 1–11.

[Wang et al., 2017] Wang X., He K., and Gupta A. (2017). "Transitive Invariance for Self-Supervised Visual Representation Learning". In: *IEEE International Conference on Computer Vision (ICCV)*. Vol. 11320, pp. 1329–1338.

[Wu et al., 2017] Wu C., Manmatha R., Smola A., and Krähenbühl P. (2017). "Sampling Matters in Deep Embedding Learning". In: *IEEE International Conference on Computer Vision (ICCV)*. Vol. 11320, pp. 2840–2848.

[Xu et al., 2017a] Xu Y., Li Y., Wang Y., Liu M., Fan Y., Lai M., and Chang E. (2017a). "Gland Instance Segmentation using Deep Multichannel Neural Networks". In: *IEEE Transactions on Biomedical Engineering* 64 (12), pp. 2901–2912.

[Xu et al., 2017b] Xu Y., Jia Z., Wang L., Ai Y., Zhang F., Lai M., and Chang E. (2017b). "Large Scale Tissue Histopathology Image Classification, Segmentation, and Visualization via Deep Convolutional Activation Features". In: *BMC Bioinformatics* 18 (281), pp. 1–17.

[Yang et al., 2017] Yang L., Zhang Y., Chen J., Zhang S., and Chen D. (2017). "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 399–407.

[Yu et al., 2017] Yu Y., Gong Z., Zhong P., and Shan J. (2017). "Unsupervised Representation Learning with Deep Convolutional Neural Network for Remote Sensing Images". In: *International Conference on Image and Graphics*. Vol. 11320, pp. 97–108.

[Zhang et al., 2017] Zhang L., Sonka M., Lu L., Summers R., and Yao J. (2017). "Combining Fully Convolutional Networks and Graph-based Approach for Automated Segmentation of Cervical Cell Nuclei". In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, pp. 406–409.

[Zhong et al., 2017] Zhong Q., Guo T., Rechsteiner M., Rüschoff J., Rupp N., Fankhauser C., Saba K., Mortezavi A., Poyet C., Hermanns T., Zhu Y., Moch H., Aebersold R., and Wild P. (2017). "A Curated Collection of Tissue Microarray Images and Clinical Outcome Data of Prostate Cancer Patients". In: 4 (170014), 113200K.

[Zhu et al., 2017] Zhu X., Yao J., Zhu F., and Huang J. (2017). "WSISA: Making Survival Prediction from Whole Slide Pathology Images". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 7234–7242.

[Zintgraf et al., 2017] Zintgraf L., Cohen T., Adel T., and Welling M. (2017). "Visualizing Deep Neural Network Decisions: Prediction Difference Analysis". In: *International Conference on Learning Representations (ICLR)* 11320, 113200K.

[Allison et al., 2016] Allison K., Rendi M., Peacock S., Morgan T., Elmore J., and Weaver D. (2016). "Histologic Features associated with Diagnostic Agreement in Atypical Ductal Hyperplasia of the Breast: Illustrative Cases from the B-Path Study". In: *Histopathology* 69 (6), pp. 1028–1046.

[Castelvecchi, 2016] Castelvecchi D. (2016). "Can We Open the Black Box of AI?" In: *Nature News* 538 (7623), p. 20.

[Chen et al., 2016] Chen H., Dou Q., Wang X., Qin J., and Heng P. (2016). "Mitosis Detection in Breast Cancer Histology Images via deep cascaded networks". In: *Association for the Advancement of Artificial Intelligence (AAAI)*. Vol. 11320, pp. 1160–1166.

[Choy et al., 2016] Choy B., Pearce S., Anderson B., Shalhav A., Zagaja G., Eggener S., and Paner G. (2016). "Prognostic Significance of Percentage and Architectural Types of Contemporary Gleason Pattern 4 Prostate Cancer in Radical Prostatectomy". In: *The American Journal of Surgical Pathology* 40 (10), pp. 1400–1406.

[Defferrard et al., 2016] Defferrard M., Bresson X., and Vandergheynst P. (2016). "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 11320, pp. 3844–3852.

[Epstein et al., 2016] Epstein J., Egevad L., Amin M., Delahunt B., Srigley J., and Humphrey P. (2016). "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma". In: *The American Journal of Surgical Pathology* 40 (2), pp. 244–252.

[Gal et al., 2016] Gal Y. and Ghahramani Z. (2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *International Conference on Machine Learning (ICML)*. Vol. 11320, pp. 1050–1059.

[Ganin et al., 2016] Ganin Y., Ustinova E., Ajakan H., Germain P., Larochelle H., Laviolette F., Marchand M., and Lempitsky V. (2016). "Domain-adversarial Training of Neural Networks". In: *The Journal of Machine Learning Research* 17 (1), pp. 2096–2030.

[Gordetsky et al., 2016] Gordetsky J. and Epstein J. (2016). “Grading of Prostatic Adenocarcinoma: Current State and Prognostic Implications”. In: *Diagnostic Pathology* 11 (1), pp. 1–8.

[He et al., 2016] He K., Zhang X., Ren S., and Sun J. (2016). “Deep Residual Learning for Image Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 770–778.

[Higgins et al., 2016] Higgins I., Matthey L., Pal A., Burgess C., Glorot X., Botvinick M., Mohamed S., and Lerchner A. (2016). “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework”. In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[Hou et al., 2016] Hou L., Samaras D., Kurc T., Gao Y., Davis J., and Saltz J. (2016). “Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 2424–2433.

[Jiang et al., 2016] Jiang Z., Zheng Y., Tan H., Tang B., and Zhou H. (2016). “Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Vol. 11320, 113200K.

[Kashif et al., 2016] Kashif M., Raza S., Sirinukunwattana K., Arif M., and Rajpoot N. (2016). “Handcrafted Features with Convolutional Neural Networks for Detection of Tumor Cells in Histology Images”. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, pp. 1029–1032.

[Larsen et al., 2016] Larsen A., Sønderby S., Larochelle H., and Winther O. (2016). “Autoencoding Beyond Pixels using a Learned Similarity Metric”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, pp. 1558–1566.

[Larsson et al., 2016] Larsson G., Maire M., and Shakhnarovich G. (2016). “Learning Representations for Automatic Colorization”. In: *European Conference on Computer Vision (ECCV)*. Vol. 11320, pp. 577–593.

[Li et al., 2016] Li W., Manivannan S., Akbar S., Zhang J., Trucco E., and McKenna S. (2016). “Gland Segmentation in Colon Histology Images using Hand-Crafted Features and Convolutional Neural Networks”. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, pp. 1405–1408.

[Madabhushi et al., 2016] Madabhushi A. and Lee G. (2016). “Image Analysis and Machine Learning in Digital Pathology: Challenges and Opportunities”. In: *Medical Image Analysis* 33, pp. 170–175.

[Noroozi et al., 2016] Noroozi M. and Favaro P. (2016). “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles”. In: *European Conference on Computer Vision (ECCV)*. Vol. 11320, pp. 69–84.

[Oh Song et al., 2016] Oh Song H., Xiang Y., Jegelka S., and Savarese S. (2016). “Deep Metric Learning via Lifted Structured Feature Embedding”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 4004–4012.

[Ozkan et al., 2016] Ozkan T., Eruyar A., Cebeci O., Memik O., Ozcan L., and Kuskonmaz I. (2016). “Interobserver Variability in Gleason Histological Grading of Prostate Cancer”. In: *Scandinavian Journal of Urology* 50 (6), pp. 420–424.

[Paeng et al., 2016] Paeng K., Hwang S., Park S., and Kim M. (2016). “A Unified Framework for Tumor Proliferation Score Prediction in Breast Histopathology”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshops*. Vol. 11320, 113200K.

[Pathak et al., 2016] Pathak D., Krahenbuhl P., Donahue J., Darrell T., and Efros A. (2016). “Context Encoders: Feature Learning by Inpainting”. In: *IEEE Conference on Computer vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 2536–2544.

[Ribeiro et al., 2016] Ribeiro M., Singh S., and Guestrin C. (2016). “Why Should I Trust You? Explaining the Predictions of Any Classifier”. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 11320, pp. 1135–1144.

[Romo-Bucheli et al., 2016] Romo-Bucheli D., Janowczyk A., Gilmore H., Romero E., and Madabhushi A. (2016). “Automated Tubule Nuclei Quantification and Correlation with Oncotype DX Risk Categories in ER+ Breast Cancer Whole Slide Images”. In: *Scientific Reports* 6 (1), pp. 1–9.

[Schaumberg et al., 2016] Schaumberg A., Sirintrapun S., Al-Ahmadie H., Schüffler P., and Fuchs T. (2016). “DeepScope: Nonintrusive Whole Slide Saliency Annotation and Prediction from Pathologists at the Microscope”. In: *Computational intelligence methods for bioinformatics and biostatistics* 10477, pp. 42–58.

[Sharma et al., 2016] Sharma H., Zerbe N., Heim D., Wienert S., Lohmann S., Hellwich O., and Hufnagl P. (2016). “Cell Nuclei Attributed Relational Graphs for Efficient Representation and Classification of Gastric Cancer in Digital Histopathology”. In: *SPIE Medical Imaging: Digital Pathology*. Vol. 9791, p. 97910X.

[Sirinukunwattana et al., 2016] Sirinukunwattana K., Raza S., Tsang Y., Snead D., Cree I., and Rajpoot N. (2016). “Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images”. In: *IEEE Transactions on Medical Imaging* 35 (5), pp. 1196–1206.

[Sohn, 2016] Sohn K. (2016). “Improved Deep Metric Learning with Multi-Class N-Pair Loss Objective”. In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, pp. 1857–1865.

[Spanhol et al., 2016a] Spanhol F., Oliveira L., Petitjean C., and Heutte L. (2016a). “A Dataset for Breast Cancer Histopathological Image Classification”. In: *IEEE Transactions on Biomedical Engineering* 63 (7), pp. 1455–1462.

[Spanhol et al., 2016b] Spanhol F., Oliveira L., Petitjean C., and Heutte L. (2016b). “Breast Cancer Histopathological Image Classification using Convolutional Neural Networks”. In: *IEEE International Joint Conference on Neural Networks (IJCNN)*. Vol. 11320, pp. 2560–2567.

[Taylor et al., 2016] Taylor D., Zeaf I., Lovchik R., and Kaigala G. (2016). “Centimeter-scale Surface Interactions using Hydrodynamic Flow Confinements”. In: *Langmuir* 32 (41), pp. 10537–10544.

[Vahadane et al., 2016] Vahadane A., Peng T., Sethi A., Albarqouni S., Wang L., Baust M., Steiger K., Schlitter A., Esposito I., and Navab N. (2016). “Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images”. In: *IEEE Transactions on Medical Imaging* 35 (8), pp. 1962–1971.

[Vyberg et al., 2016] Vyberg M. and Nielsen S. (2016). “Proficiency Testing in Immunohistochemistry—Experiences from Nordic Immunohistochemical Quality Control (Nordiqc)”. In: *Virchows Archiv* 468 (1), pp. 19–29.

[Wang et al., 2016] Wang S., Yao J., Xu Z., and Huang J. (2016). “Subtype Cell Detection with an Accelerated Deep Convolution Neural Network”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 640–648.

[Xie et al., 2016] Xie J., Girshick R., and Farhadi A. (2016). “Unsupervised Deep Embedding for Clustering Analysis”. In: *International Conference on Machine Learning (ICML)*. Vol. 11320, pp. 478–487.

[Xu et al., 2016a] Xu J., Luo X., Wang G., Gilmore H., and Madabhushi A. (2016a). “A Deep Convolutional Neural Network for Segmenting and Classifying Epithelial and Stromal Regions in Histopathological Images”. In: *Neurocomputing* 191, pp. 214–223.

[Xu et al., 2016b] Xu J., Xiang L., Liu Q., Gilmore H., Wu J., Tang J., and Madabhushi A. (2016b). “Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images”. In: *IEEE Transactions on Medical Imaging* 35 (1), pp. 119–130.

[Yang et al., 2016] Yang P. and Yang G. (2016). “Feature Extraction using Dual-Tree Complex Wavelet Transform and Gray Level Co-Occurrence Matrix”. In: *Neurocomputing* 197, pp. 212–220.

[Zagoruyko et al., 2016] Zagoruyko S. and Komodakis N. (2016). “Wide Residual Networks”. In: *British Machine Vision Conference (BMVC)*. Vol. 11320, 113200K.

[Zhang et al., 2016a] Zhang R., Isola P., and Efros A. (2016a). “Colorful Image Colorization”. In: *European Conference on Computer Vision (ECCV)*. Vol. 11320, pp. 649–666.

[Zhang et al., 2016b] Zhang X., Zhou F., Lin Y., and Zhang S. (2016b). “Embedding Label Structures for Fine-Grained Feature Representation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 1114–1123.

[Zhou et al., 2016] Zhou B., Khosla A., Lapedriza A., Oliva A., and Torralba A. (2016). “Learning Deep Features for Discriminative Localization”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 2921–2929.

[Allemani et al., 2015] Allemani C., Weir H., Carreira H., Harewood R., Spika D., Wang X., Bannon F., Ahn J., Johnson C., Bonaventure A., et al. (2015). “Global Surveillance of Cancer Survival 1995-2009: Analysis of Individual Data for 25,676,887 Patients from 279 Population-Based Registries in 67 Countries (CONCORD-2)”. In: *The Lancet* 385 (9972), pp. 977–1010.

[Amin et al., 2015] Amin M., Smith S., Reuter V., Epstein J., Grignon D., Hansel D., Lin O., McKenney J., Montironi R., Paner G., et al. (2015). “Update for the Practicing Pathologist: The International Consultation on Urologic Disease-European Association of Urology Consultation on Bladder Cancer”. In: *Modern Pathology* 28 (5), pp. 612–630.

[Bach et al., 2015] Bach S., Binder A., Montavon G., Klauschen F., Müller K., and Samek W. (2015). “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”. In: *PLoS ONE* 10 (7), 113200K.

[Bejnordi et al., 2015] Bejnordi B., Litjens G., Hermsen M., Karssemeijer N., and Laak J. (2015). “A Multi-Scale Superpixel Classification Approach to the Detection of Regions of Interest in Whole Slide Histopathology Images”. In: *SPIE Medical Imaging: Digital Pathology*. Vol. 9420, 94200H.

[Doersch et al., 2015] Doersch C., Gupta A., and Efros A. (2015). “Unsupervised Visual Representation Learning by Context Prediction”. In: *IEEE International Conference on Computer Vision (ICCV)*. Vol. 11320, pp. 1422–1430.

[Elmore et al., 2015] Elmore J., Longton G., Carney P., Geller B., Onega T., Tosteson A., Nelson H., Pepe M., Allison K., Schnitt S., O'Malley F., and Weaver D. (2015). "Diagnostic Concordance among Pathologists Interpreting Breast Biopsy Specimens". In: *JAMA* 313 (11), pp. 1122–1132.

[He et al., 2015] He K., Zhang X., Ren S., and Sun J. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *IEEE International Conference on Computer Vision (ICCV)*. Vol. 11320, pp. 1026–1034.

[Hinton et al., 2015] Hinton G., Vinyals O., and Dean J. (2015). "Distilling the Knowledge in a Neural Network". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 11320, 113200K.

[Kindermans et al., 2015] Kindermans P., Schutt K. T., Alber M., Muller K., and Dahne S. (2015). "PatternNet and PatternLRP - Improving the Interpretability of Neural Networks". In: *arXiv:1705.05598* 11320, 113200K.

[Kingma et al., 2015] Kingma D. and Ba J. (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[Montavon et al., 2015] Montavon G., Bach S., Binder A., Samek W., and Muller K. (2015). "Explaining NonLinear Classification Decisions with Deep Taylor Decomposition". In: *Pattern Recognition* 65, pp. 211–222.

[Peikari et al., 2015] Peikari M., Zubovits J., Clarke G., and Martel A. (2015). "Clustering Analysis for Semi-Supervised Learning Improves Classification Performance of Digital Pathology". In: *International Workshop on Machine Learning in Medical Imaging*. Vol. 11320, pp. 263–270.

[Ronneberger et al., 2015] Ronneberger O., Fischer P., and Brox T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 234–241.

[Salmo, 2015] Salmo E. (2015). "An Audit of Inter-Observer Variability in Gleason Grading of Prostate Cancer Biopsies: The Experience of Central Pathology Review in the North West of England". In: *Integr Cancer Sci Ther* 2 (2), pp. 104–106.

[Schroff et al., 2015] Schroff F., Kalenichenko D., and Philbin J. (2015). "Facenet: A Unified Embedding for Face Recognition and Clustering". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 815–823.

[Sharma et al., 2015] Sharma H., Zerbe N., Lohmann S., Kayser K., Hellwich O., and Hufnagl P. (2015). "A Review of Graph-Based Methods for Image Analysis in Digital Histopathology". In: *Diagnostic Pathology*. Vol. 1. 1, 113200K.

[Song et al., 2015] Song Y., Zhang L., Chen S., Ni D., Lei B., and Wang T. (2015). "Accurate Segmentation of Cervical Cytoplasm and Nuclei based on Multiscale Convolutional Network and Graph Partitioning". In: *IEEE Transactions on Biomedical Engineering* 62 (10), pp. 2421–2433.

[Torlakovic et al., 2015] Torlakovic E., Nielsen S., Francis G., Garratt J., Gilks B., Goldsmith J., Hornick J., Hyjek E., Ibrahim M., Miller K., et al. (2015). "Standardization of Positive Controls in Diagnostic Immunohistochemistry: Recommendations from the International Ad Hoc Expert Committee". In: *Applied Immunohistochemistry & Molecular Morphology* 23 (1), pp. 1–18.

[Veta et al., 2015] Veta M., Van Diest P., Willems S., Wang H., Madabhushi A., Cruz-Roa A., Gonzalez F., Larsen A., Vestergaard J., Dahl A., et al. (2015). "Assessment of Algorithms for Mitosis Detection in Breast Cancer Histopathology Images". In: *Medical Image Analysis* 20 (1), pp. 237–248.

[Wang et al., 2015] Wang X. and Gupta A. (2015). "Unsupervised Learning of Visual Representations using Videos". In: *IEEE International Conference on Computer Vision (ICCV)*. Vol. 11320, pp. 2794–2802.

[Xie et al., 2015] Xie Y., Xing F., Kong X., Su H., and Yang L. (2015). "Beyond Classification: Structured Regression for Robust Cell Detection using Convolutional Neural Network". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 358–365.

[Yosinski et al., 2015] Yosinski J., Clune J., Nguyen A. M., Fuchs T. J., and Lipson H. (2015). "Understanding Neural Networks through Deep Visualization". In: *International Conference on Machine Learning (ICML) Workshops* 11320, 113200K.

[Amin et al., 2014] Amin M., Lin D., Gore J., Srigley J., Samaratunga H., Egevad L., Rubin M., Nacey J., Carter H., Klotz L., et al. (2014). "The Critical Role of the Pathologist in Determining Eligibility for Active Surveillance as a Management Option in Patients with Prostate Cancer: Consensus Statement with Recommendations Supported by the College of American Pathologists, International Society of Urological Pathology, Association of Directors of Anatomic and Surgical Pathology, the New Zealand Society of Pathologists, and the Prostate Cancer Foundation". In: *Archives of Pathology and Laboratory Medicine* 138 (10), pp. 1387–1405.

[Andriluka et al., 2014] Andriluka M., Pishchulin L., Gehler P., and Bernt S. (2014). "2D Human Pose Estimation: New Benchmark and State of the Art Analysis". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 3686–3693.

[Gomes et al., 2014] Gomes D., Porto S., Balabram D., and Gobbi H. (2014). "Inter-observer Variability between General Pathologists and A Specialist in Breast Pathology in the Diagnosis of Lobular Neoplasia, Columnar Cell Lesions, Atypical Ductal Hyperplasia and Ductal Carcinoma In Situ of the Breast". In: *Diagnostic Pathology* 9 (121), 113200K.

[Goodfellow et al., 2014] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., and Bengio Y. (2014). "Generative Adversarial Networks". In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, 113200K.

[Howat et al., 2014] Howat W., Lewis A., Jones P., Kampf C., Ponten F., van der Loos C., Gray N., Womack C., and Warford A. (2014). "Antibody Validation of Immunohistochemistry for Biomarker Discovery: Recommendations of a Consortium of Academic and Pharmaceutical Based Histopathology Researchers". In: *Methods* 70 (1), pp. 34–38.

[Hu et al., 2014] Hu S., Xu C., Guan W., Tang Y., and Liu Y. (2014). "Texture Feature Extraction based on Wavelet Transform and Gray-Level Co-Occurrence Matrices Applied to Osteosarcoma Diagnosis". In: *Bio-Medical Materials and Engineering* 24, pp. 129–143.

[Huang et al., 2014] Huang C., Kong M., Zhou M., Rosenkrantz A., Taneja S., Melamed J., and Deng F. (2014). "Gleason Score 3+ 4= 7 Prostate Cancer with Minimal Quantity of Gleason Pattern 4 on Needle Biopsy is Associated With Low-Risk Tumor in Radical Prostatectomy Specimen". In: *The American Journal of Surgical Pathology* 38 (8), pp. 1096–1101.

[Jiang et al., 2014] Jiang L., Meng D., Yu S., Lan Z., Shan S., and Hauptmann A. (2014). "Self-paced Learning with Diversity". In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, 113200K.

[Kingma et al., 2014] Kingma D. and Welling M. (2014). "Auto-encoding Variational Bayes". In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[O'Hurley et al., 2014] O'Hurley G., Sjostedt E., Rahman A., Li B., Kampf C., Ponten F., Gallagher W., and Lindskog C. (2014). "Garbage In, Garbage Out: A Critical Evaluation of Strategies used for Validation of Immunohistochemical Biomarkers". In: *Molecular Oncology* 8 (4), pp. 783–798.

[Roux, 2014] Roux L. (2014). *Mitosis Atypia 14 Grand Challenge*. `https://mitos-atypia-14.grand-challenge.org/dataset/`.

[Seguin et al., 2014] Seguin B., Saab H., Gabrani M., and Estellers V. (2014). "Estimating Pattern Sensitivity to the Printing Process for Varying Dose/Focus Conditions for RET Development in the Sub-22nm Era". In: *SPIE Metrology, Inspection, and Process Control for Microlithography XXVIII*. Vol. 9050, 90500P.

[Simonyan et al., 2014] Simonyan K. and Zisserman A. (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations (ICLR)*. Vol. 11320, 113200K.

[Torlakovic et al., 2014] Torlakovic E., Francis G., Garratt J., Gilks B., Hyjek E., Ibrahim M., Miller R., Nielsen S., Petcu E., Swanson P., et al. (2014). "Standardization of Negative Controls in Diagnostic Immunohistochemistry: Recommendations from the International Ad Hoc Expert Panel". In: *Applied Immunohistochemistry & Molecular Morphology: AIMM/official Publication of the Society for Applied Immunohistochemistry* 22 (4), p. 241.

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

[Wolff et al., 2014] Wolff A., Hammond M., Hicks D., Dowsett M., McShane L., Allison K., Allred D., Bartlett J., Bilous M., Fitzgibbons P., et al. (2014). "Recommendations for Human Epidermal Growth Factor Receptor 2 Testing In Breast Cancer: American Society Of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Update". In: *Archives of Pathology and Laboratory Medicine* 138 (2), pp. 241–256.

[Xu et al., 2014] Xu Y., Zhu J., Chang E., Lai M., and Tu Z. (2014). "Weakly Supervised Histopathology Cancer Image Segmentation and Classification". In: *Medical Image Analysis* 18 (3), pp. 591–604.

[Zeiler et al., 2014] Zeiler M. and Fergus R. (2014). "Visualizing and Understanding Convolutional Networks". In: *European Conference on Computer Vision (ECCV)* 11320, pp. 818–833.

[Ciresan et al., 2013] Ciresan D., Giusti A., Gambardella L., and Schmidhuber J. (2013). "Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 411–418.

[Cruz-Roa et al., 2013] Cruz-Roa A., Ovalle A., John E., Madabhushi A., and González Osorio F. (2013). "A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 11320, pp. 403–410.

[He et al., 2013] He H. and Ma Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. 1st. Vol. 11320. Wiley-IEEE Press, 113200K.

[Krause et al., 2013] Krause J., Stark M., Deng J., and Fei-Fei L. (2013). "3D Object Representations for Fine-Grained Categorization". In: *IEEE International Conference on Computer Vision (ICCV) Workshops*. Vol. 11320, pp. 554–561.

[Lopez et al., 2013] Lopez X., D'Andrea E., Barbot P., Bridoux A., Rorive S., Salmon I., Debeir O., and Decaestecke C. (2013). "An Automated Blur Detection Method for Histological Whole Slide Imaging". In: *PLoS ONE* 8 (12), 113200K.

[Roux et al., 2013] Roux L., Racoceanu D., Lomènie N., Kulikova M., Irshad H., Klossa J., Capron F., Genestie C., Naour G., and Gurcan M. (2013). "Mitosis Detection in Breast Cancer Histological Images an ICPR 2012 Contest". In: *Journal of Pathology Informatics* 4 (8), 113200K.

[Simonyan et al., 2013] Simonyan K., Vedaldi A., and Zisserman A. (2013). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *arXiv:1312.6034* 11320, 113200K.

[Achanta et al., 2012] Achanta R., Shaji A., Smith K., Lucchi A., Fua P., and Süsstrunk S. (2012). "SLIC Superpixels Compared to State-Of-The-Art Superpixel Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (11), pp. 2274–2282.

[Brügmann et al., 2012] Brügmann A., Eld M., Lelkaitis G., Nielsen S., Grunkin M., Hansen J., Foged N., and Vyberg M. (2012). "Digital Image Analysis of Membrane Connectivity is a Robust Measure of HER2 Immunostains". In: *Breast Cancer Research and Treatment* 132 (1), pp. 41–49.

[Ciresan et al., 2012] Ciresan D., Giusti A., Gambardella L., and Schmidhuber J. (2012). "Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images". In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, pp. 2843–2851.

[Hamilton et al., 2012] Hamilton P., Wang Y., and McCullough S. (2012). "Virtual Microscopy and Digital Pathology in Training And Education". In: *APMIS* 120 (4), pp. 305–315.

[Lovchik et al., 2012] Lovchik R., Kaigala G., Georgiadis M., and Delamarche E. (2012). "Micro-immunohistochemistry using a Microfluidic Probe". In: *Lab on a Chip* 12 (6), pp. 1040–1043.

[Pinard et al., 2012] Pinard R., Tedeschi G., Williams C., and Wang D. (2012). *Methods and System for Validating Sample Images for Quantitative Immunoassays*. US Patent 8,160,348.

[Rizzardi et al., 2012] Rizzardi A., Johnson A., Vogel R., Pambuccian S., Henriksen J., Skubitz A., Metzger G., and Schmechel S. (2012). "Quantitative Comparison of Immunohistochemical Staining Measured by Digital Image Analysis Versus Pathologist Visual Scoring". In: *Diagnostic Pathology* 7 (1), pp. 1–10.

[Chang et al., 2011] Chang H., Loss L., and Parvin B. (2011). "Nuclear Segmentation in H&E Sections via Multireference Graph Cut (MRGC)". In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, 113200K.

[Copete et al., 2011] Copete M., Garratt J., Gilks B., Pilavdzic D., Berendt R., Bigras G., Mitchell S., Lining L., Cheung C., and Torlakovic E. (2011). "Inappropriate Calibration and Optimisation of Pan-Keratin (Pan-CK) and Low Molecular Weight Keratin (LMWCK) Immunohistochemistry Tests: Canadian Immunohistochemistry Quality Control (CIQC) Experience". In: *Journal of Clinical Pathology* 64 (3), pp. 220–225.

[Kaigala et al., 2011] Kaigala G., Lovchik R., Drechsler U., and Delamarche E. (2011). "A Vertical Microfluidic Probe". In: *Langmuir* 27 (9), pp. 5686–5693.

[Kumar et al., 2010] Kumar M., Packer B., and Koller D. (2010). "Self-paced Learning for Latent Variable Models". In: *Advances in Neural Information Processing (NeurIPS)*. Vol. 11320, 113200K.

[Marcel et al., 2010] Marcel S. and Rodriguez Y. (2010). "Torchvision the Machine-Vision Package of Torch". In: *ACM International Conference on Multimedia*. Vol. 11320, pp. 1485–1488.

[Thai-Nghe et al., 2010] Thai-Nghe N., Gantner Z., and Schmidt-Thieme L. (2010). "Cost-sensitive Learning Methods for Imbalanced Data". In: *International Joint Conference on Neural Networks (IJCNN)*. Vol. 11320, pp. 1–8.

[Deng et al., 2009] Deng J., Dong W., Socher R., Li L., Kai L., and Li F. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 248–255.

[Dobson et al., 2009] Dobson L., Conway C., Hanley A., Johnson A., Costello S., O'Grady A., Connolly Y., Magee H., O'Shea D., Jeffers M., et al. (2009). "Image Analysis as an Adjunct to Manual HER-2 Immunohistochemical Review: A Diagnostic Tool to Standardize Interpretation". In: *Histopathology* 57 (1), pp. 27–38.

[Jian et al., 2009] Jian M., Liu L., and Guo F. (2009). "Texture Image Classification using Visual Perceptual Texture and Gabor Wavelet Features". In: *Journal of Computers* 4 (8), pp. 763–770.

[Macenko et al., 2009] Macenko M., Niethammer M., Marron J., Borland D., Woosley J., Xiaojun G., Schmitt C., and Thomas N. (2009). "A Method for Normalizing Histology Slides for Quantitative Analysis". In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. Vol. 11320, pp. 1107–1110.

[Masmoudi et al., 2009] Masmoudi H., Hewitt S., Petrick N., Myers K., and Gavrielides M. (2009). "Automated Quantitative Assessment of HER-2/neu Immunohistochemical Expression in Breast Cancer". In: *IEEE Transactions on Medical Imaging* 28 (6), pp. 916–925.

[Scarselli et al., 2009] Scarselli F., Gori M., Tsoi A. C., Hagenbuchner M., and Monfardini G. (2009). "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20 (1), pp. 61–80.

[Wilbur et al., 2009] Wilbur D., Madi K., Colvin R., Duncan L., Faquin W., Ferry J., Frosch M., Houser S., Kradin R., Lauwers G., et al. (2009). "Whole-slide Imaging Digital Pathology as a Platform for Teleconsultation: A Pilot Study using Paired Subspecialist Correlations". In: *Archives of Pathology & Laboratory Medicine* 133 (12), pp. 1949–1953.

[Cosatto et al., 2008] Cosatto E., Miller M., Graf H., and Meyer J. (2008). "Grading Nuclear Pleomorphism on Histological Micrographs". In: *International Conference on Pattern Recognition (ICPR)*. Vol. 11320, pp. 1–4.

[Van der Maaten et al., 2008] Van der Maaten L. and Hinton G. (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (11), 113200K.

[Von Wasielewski et al., 2008] Von Wasielewski R., Hasselmann S., Rüschoff J., Fisseler-Eckhoff A., and Kreipe H. (2008). "Proficiency Testing of Immunohistochemical Biomarker Assays in Breast Cancer". In: *Virchows Archiv* 453 (6), pp. 537–543.

[Yaziji et al., 2008] Yaziji H., Taylor C., Goldstein N., Dabbs D., Hammond E., Hewlett B., Floyd A., Barry T., Martin A., Badve S., et al. (2008). "Consensus Recommendations on Estrogen Receptor Testing in Breast Cancer by Immunohistochemistry". In: *Applied Immunohistochemistry & Molecular Morphology* 16 (6), pp. 513–520.

[Cheng et al., 2007] Cheng L., Davidson D., Lin H., and Koch M. (2007). "Percentage of Gleason Pattern 4 and 5 Predicts Survival After Radical Prostatectomy". In: *Cancer* 110 (9), pp. 1967–1972.

[Goldstein et al., 2007] Goldstein N., Hewitt S., Taylor C., Yaziji H., Hicks D., et al. (2007). "Recommendations for Improved Standardization of Immunohistochemistry". In: *Applied Immunohistochemistry & Molecular Morphology* 15 (2), pp. 124–133.

[Han et al., 2007] Han J. and Ma K. (2007). "Rotation-invariant and Scale-invariant Gabor Features for Texture Image Retrieval". In: *Journal of Image and Vision Computing* 25, pp. 1474–1481.

[Kong et al., 2007] Kong J., Shimada H., Boyer K., Saltz J., and Gurcan M. (2007). "Image Analysis for Automated Assessment of Grade of Neuroblastic Differentiation". In: *IEEE International Symposium on Biomedical Imaging (ISI)*. Vol. 11320, pp. 61–64.

[Wolff et al., 2007] Wolff A., Hammond M., Schwartz J., Hagerty K., Allred D., Cote R., Dowsett M., Fitzgibbons P., Hanna W., Langer A., et al. (2007). "American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer". In: *Archives of Pathology & Laboratory Medicine* 131 (1), pp. 18–43.

[Hadsell et al., 2006] Hadsell R., Chopra S., and LeCun Y. (2006). "Dimensionality Reduction by Learning an Invariant Mapping". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 11320, pp. 1735–1742.

[Araùjo et al., 2005] Araùjo T., Aresta G., Castro E., Rouco J., Aguiar P., Eloy C., Polónia A., and Campilho A. (2005). "Classification of Breast Cancer Histology Images using Convolutional Neural Networks". In: *PLoS One* 12 (6), 113200K.

[Epstein et al., 2005] Epstein J., Allsbrook Jr W., Amin M., Egevad L., et al. (2005). "The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma". In: *The American Journal of Surgical Pathology* 29 (9), pp. 1228–1242.

[Demir et al., 2004] Demir C., Yener B., and Gultekin S. (2004). "The Cell Graphs of Cancer". In: *Bioinformatics* 20, pp. 145–151.

[Lowe, 2004] Lowe D. (2004). "Distinctive Image Features from Scale-invariant Keypoints". In: *International Journal of Computer Vision* 60 (2), pp. 91–110.

[Reinhard et al., 2001] Reinhard E., Adhikhmin M., Gooch B., and Shirley P. (2001). "Color Transfer Between Images". In: *IEEE Computer Graphics and Applications* 21 (5), pp. 34–41.

[Taylor, 2000] Taylor C. (2000). "The Total Test Approach to Standardization of Immunohistochemistry". In: *Archives of Pathology & Laboratory Medicine* 124 (7), pp. 945–951.

[Francis et al., 1997] Francis K. and Palsson B. (1997). "Effective Intercellular Communication Distances are Determined by the Relative Time Constants for Cyto/Chemokine Secretion and Diffusion". In: *Proceedings of the National Academy of Sciences* 94 (23), pp. 12258–12262.

[Barber et al., 1996] Barber B., Dobkin D., and Huhdanpaa H. (1996). "The Quickhull Algorithm for Convex Hulls". In: *ACM Transactions on Mathematical Software* 22 (4), pp. 469–483.

[Livens et al., 1996] Livens S., Scheunders P., Van de Wouwer G., Van Dyck D., Smets H., Winkelmans J., and Bogaerts W. (1996). "A Texture Analysis Approach to Corrosion Image Classification". In: *Microscopy, Microanalysis, Microstructures* 7 (2), pp. 143–152.

[Potjer, 1996] Potjer F. (1996). “Region Adjacency Graphs and Connected Morphological Operators”. In: *Mathematical Morphology and Its Applications to Image and Signal Processing* 5, pp. 111–118.

[Cybenko, 1989] Cybenko G. (1989). “Approximation by Superpositions of a Sigmoidal Function”. In: *Mathematics of Control, Signals and Systems* 2 (4), pp. 303–314.

[Weininger, 1988] Weininger D. (1988). “SMILES, A Chemical Language and Information System. 1. Introduction To Methodology And Encoding Rules”. In: *Journal of Chemical Information and Computer Sciences* 28 (1), pp. 31–36.

[Farin, 1986] Farin G. (1986). “Triangular Bernstein-Bezier Patches”. In: *Computer Aided Geometric Design* 3 (2), pp. 83–127.

[Alfeld, 1984] Alfeld P. (1984). “A Trivariate Clough-Tocher Scheme for Tetrahedral Data”. In: *Computer Aided Geometric Design* 1 (2), pp. 169–181.

[DeGroot et al., 1983] DeGroot M. and Fienberg S. (1983). “The Comparison and Evaluation of Forecasters”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 32 (1-2), pp. 12–22.

[Davies et al., 1979] Davies D. L. and Bouldin D. W. (1979). “A Cluster Separation Measure”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (2), 113200K.

[Haralick et al., 1973] Haralick R., Shanmugam K., and Dinstein I. (1973). “Texural Features for Image Classification”. In: *IEEE Transaction on Systems, Man and Cybernatics* 3 (6), pp. 610–621.

[Weisfeiler et al., 1968] Weisfeiler B. and Lehman A. A. (1968). “A Reduction of a Graph to a Canonical Form and an Algebra Arising during this Reduction”. In: *Nauchno-Technicheskaya Informatsia*. Vol. 2. 9, pp. 12–16.

List of Own Publications

Articles in peer-reviewed journals

Pati† P., Jaume† G., Foncubierta-Rodríguez A., Feroce F., Anniciello A., Scognamiglio G., Brancati N., Fiche M., Dubruc E., Riccio D., Bonito M., Pietro G., Botti G., Thiran J., Frucci M., Goksel O., and Gabrani M. (2022). “Hierarchical Graph Representations in Digital Pathology”. In: *Medical Image Analysis* 75, p. 102264.

Pati, P., Foncubierta-Rodríguez A., Goksel O., and Gabrani M. (2021c). “Reducing Annotation Effort in Digital Pathology: A Co-Representation Learning Framework for Classification Tasks”. In: *Medical Image Analysis* 67, p. 101859.

Arar† N., **Pati† P.**, Kashyap A., Khartchenko A., Goksel O., Kaigala G., and Gabrani M. (2019). “High-Quality Immunohistochemical Stains through Computational Assay Parameter Optimization”. In: *IEEE Transactions on Biomedical Engineering* 66 (10), pp. 2952–2963.

Binder T., Tantaoui E., **Pati, P.**, Catena R., Set-Aghayan A., and Gabrani M. (2019). “Multi-organ Gland Segmentation using Deep Learning”. In: *Frontiers in Medicine* 6, p. 173.

Kashyap A., Khartchenko A., **Pati, P.**, Gabrani M., Schraml P., and Kaigala G. (2019). “Quantitative Microimmunohistochemistry for the Grading of Immunostains on Tumour Tissues”. In: *Nature Biomedical Engineering* 3 (6), pp. 478–490.

Conference contributions

Anklin† V., **Pati† P.**, Jaume† G., Bozorgtabar B., Foncubierta-Rodríguez A., Thiran J., Sibony M., Gabrani M., and Goksel O. (2021). “Learning Whole-Slide Segmentation from Inexact and Incomplete Labels using Tissue Graphs”. In: *Medical Image Computing and Computer Assisted Interventions (MICCAI)*, pp. 636–646.

Jaume† G., **Pati† P.**, Anklin V., Foncubierta-Rodríguez A., and Gabrani M. (2021a). “HistoCartography: A Toolkit for Graph Analytics in Digital Pathology”. In: *Medical Image Computing and Computer Assisted Interventions (MICCAI) Workshop on Computational Pathology (COMPAY)*. Vol. 156. pp. 117–128. [Best Software Paper Award].

Jaume† G., **Pati† P.**, Bozorgtabar B., Foncubierta-Rodríguez A., Feroce F., Anniciello A., Rau T., Thiran J., Gabrani M., and Goksel O. (2021b). “Quantifying Explainers of Graph Neural Networks in Computational Pathology”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8106–8116.

Jaume† G., **Pati† P.**, Foncubierta-Rodriguez A., Feroce F., Scognamiglio G., Anniciello A., Thiran J., Goksel O., and Gabrani M. (2020). “Towards Explainable Graph Representations in Digital Pathology”. In: *International Conference on Machine Learning (ICML) Workshop on Computational Biology (CompoBio)*. [Best Paper Award].

Pati P., Foncubierta-Rodríguez A., Goksel O., and Gabrani M. (2020). “Mitosis Detection under Limited Annotation: A Joint Learning Approach”. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 486–489.

Pati† P., Jaume† G., Fernandes L., Foncubierta-Rodríguez A., Feroce F., Anniciello A., Scognamiglio G., Brancati N., Riccio D., Di Bonito M., De Pietro G., Botti G., Goksel O., Thiran J., Frucci M., and Gabrani M. (2020). “HACT-Net: A Hierarchical Cell-to-Tissue Graph Neural Network for Histopathological Image Classification”. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI) Workshop on Graphs in Biomedical Image Analysis (GRAIL)*. pp. 208–219. [Best Paper Award].

Pati P., Catena R., Goksel O., and Gabrani M. (2019). “A Deep Learning Framework for Context-Aware Mitotic Activity Estimation in Whole Slide Images”. In: *SPIE Medical Imaging: Digital Pathology*. Vol. 10956, p. 1095609.

Pati P., Andani S., Pediaditis M., Viana M., Rüschoff J., Wild P., and Gabrani M. (2018). “Deep Positive-Unlabeled Learning for Region of Interest Localization in Breast Tissue Images”. In: *SPIE Medical Imaging: Digital Pathology*. Vol. 10581, p. 1058107.

Arar† N., **Pati† P.**, Kashyap A., Khartchenko A., Goksel O., Kaigala G., and Gabrani M. (2017). “Computational Immunohistochemistry: Recipes for Standardization of Immunostaining”. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 48–55.

---

† denotes equal contribution in a shared first authorship

Preprints

Brancati N., Anniciello A., **Pati, P.**, Riccio D., Scognamiglio G., Jaume G., De Pietro G., Di Bonito M., Foncubierta A., Botti G., Gabrani M., Feroce F., and Frucci M. (2021). *BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images*. arXiv:2111.04740.

Patents

Pati, P., Jaume G., Foncubierta A., and Gabrani M. (2021a). “Interpretation of Whole-slide Images in Digital Pathology”. [Filed].

Pati, P., Jaume G., Thandiackal K., Foncubierta A., and Gabrani M. (2021b). “Processing Multimodal Images of Tissue for Medical Evaluation”. [Filed].

Arar N., **Pati, P.**, Gabrani M., Kaigala G., Kashyap A., and Khartchenko A. (2019). “Tissue Staining Quality Determination”. US Patent 10,706,535.

Kaigala G., Khartchenko A., Kashyap A., Gabrani M., and **Pati, P.** (2019). “Biomarker Quantification in a Tissue Sample”. US Patent 20,190,286,790.

Curriculum Vitae

Personal data

Name	Pushpak Pati
Date of Birth	July 17, 1991
Place of Birth	Cuttack, Odisha, India
Citizen of	India

Education

2018 – 2021	*Doctoral Studies* ETH Zürich Computer Vision Laboratory Zürich, Switzerland
2015 – 2017	*Master of Science* ETH Zürich Electrical Engineering and Information Technology Zürich, Switzerland
2009 – 2013	*Bachelor of Technology* National Institute of Technology Electronics and Instrumentation Engineering Rourkela, India

Experience

2018 – 2021	Pre-doc Researcher IBM Research, Zürich, Switzerland
2013 – 2015	Software Engineer Microsoft, Hyderabad, India
2012 – 2012	Research Intern Indian Statistical Institute, Kolkata, India