Integrating AI-based applications in anatomic pathology workflows

by

Akash Parvatikar

B.E., R.V. College of Engineering, 2016

M.S., University of Pittsburgh, 2018

Submitted to the Graduate Faculty of

the School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Akash Parvatikar

It was defended on

February 25, 2022

and approved by

Robin E.C. Lee PhD, Associate Professor of Computational and Systems Biology, University of Pittsburgh

Min Xu PhD, Associate Professor of Computational Biology, Carnegie Mellon University

Shikhar Uttam PhD, Assistant Professor of Computational and Systems Biology, University of Pittsburgh

Arvind Ramanathan PhD, Computational Scientist, Argonne National Laboratory

Dissertation Director: S. Chakra Chennubhotla PhD, Associate Professor of Computational and Systems Biology, University of Pittsburgh

Integrating AI-based applications in anatomic pathology workflows

Akash Parvatikar, PhD

University of Pittsburgh, 2022

Traditional pathological diagnosis is considered as the gold standard by clinicians. However, this manual practice can be inefficient, error-prone, and highly subjective. To mitigate these issues, digital pathology is gaining traction which has attracted researchers to build black-box AI-based approaches intended to assist anatomic pathology workflows. The success of such approaches is dependent on large-scale generation of pathologist annotated high quality training data which is a serious bottleneck in computational pathology. Additionally, the AI systems must be interpretable and minimize the time-to-decision to achieve clinical adoption and possibly facilitate regulatory agency approvals.

We hypothesize that building computational models of already established anatomic pathology knowledge will alleviate the training data generation bottleneck and develop clinically interpretable models. In addition, implementing computational pathology workflows on the emerging customizable computing AI-based architectures will satisfy high-throughput and minimal time-to-decision requirements.

In this thesis, we tested our hypothesis on differential diagnoses of breast biopsies. We invoke analytical models to provide a quantitative assessment of the structural changes in the breast tissue along a diagnostic continuum triggered by atypia and other malignancies. We further combine the analytical models with a prototype-driven learning strategy to provide interpretability and achieve a superior classification performance in diagnosing breast biopsies over the state-of-the-art methods. To showcase the potential for seamless integration of our computational pathology framework into clinical workflows, we use a next generation high performance AI-based computing architecture to detect histological structures in breast tissue and classify them as high-risk vs low-risk. A key contribution of our framework is in building a communication platform for pathologists and computational scientists to interact and develop AI-based applications and to enhance patient care in a clinical setting.

Table of Contents

List of Tables

List of Figures

Preface

First, I would like to thank my advisor Chakra and without his timely guidance, this dissertation would be impossible to accomplish. His kind, patient, and encouraging demeanor would always give me a ray of hope to do better research. He was always ready to sit with me to either review each paragraph in a manuscript, each slide of a presentation deck, or prepare for an upcoming meeting. I would like to immensely thank him for all the support and undoubtedly, I feel lucky to have worked with such a phenomenal advisor.

Second, my heartfelt thanks to Dr. Jeffrey Fine, the pathologist on our team and a key contributor of my research work. My PhD research started with having long discussions with Dr. Fine to understand the histology images and continued with frequent interactions throughout my journey. I would like to thank him for helping us build our domain knowledge and without his support, this work would be difficult to complete.

Third, I would like to thank the present and past members of my lab, Brian, Burak, Filippo, Om, Rebekah, and Samantha for playing different roles towards the success of my graduate studies. I would like to thank the CPCB directors, administrators and peers who assured that I have a smooth experience by helping me to overcome the hurdles to my progress. Fourth, I would like to extend my special thanks to our collaborator, Arvind from Argonne National Laboratory. The conversations with Arvind especially during the low phase of my career reignited my courage and his belief in me gave me the confidence to progress. Fifth, thank you to other members of my dissertation committee, Robin, Min Xu, and Shikhar for their valuable suggestions.

Finally, I would like to thank my family, mom, dad, brother, and my close friends for having faith in my career decisions and cheering me up during tough times.

1.0 List of abbreviations

DP	Digital pathology
WSI	Whole slide image
FDA	Food and drug administration
WHO	World health organization
AI	Artificial intelligence
ML	Machine learning
TAT	Turnaround time
ROI	Region of interest
$pCAD$	Pathologist's computer-assisted diagnosis
CNN	Convolution neural networks
GNN	Graph-based convolution neural networks
UDH	Usual ductal hyperplasia
CCC	Columnar cell change
PB	Pathological benign
FEA	Flat epithelial atypia
ADH	Atypical ductal hyperplasia
ALH	Atypical lobular hyperplasia
$DCIS$	Ductal carcinoma in-situ
$H\&E$	Hematoxylin and eosin
$SLIC$	Simple linear iterative clustering
MoG	Mixture of Gaussian
KL	Kullback-Leibler
KS	Kolmogorov Smirnov
U	Unary
UB	Unary-binary
UBT	Unary-binary-ternary

G	Global
L	Local
GL	Global and local
TP	True positive
FP	False positive
TN	True negative
FN	False negative
R	Recall
wF	weighted F-score
$HLRBB$	High- and low-risk benign breast lesion dataset
$BRACS$	Breast carcinoma subtyping dataset
LD	Lumen/ductal morphology
ID	Intraductal tissue features
T	Textural features
$T-Pop$	Popular textural properties
$T-Wv$	Complex wavelet-derived textural properties
MAT	Medial axis transform
DCR	Ductal cellular region
$ALCF$	Argonne Leadership Computing Facility
GMS	Glass microscope slides
RDA	Reconfigurable dataflow architecture
DDP	Data distributed parallel training

2.0 Introduction

In the United States 13% of the women develop breast cancer during their lifetime and prognosis is poor for around 3% of them [2]. In diagnosing breast biopsies, pathologists examine the tissue slides under a microscope for recommending an optimal treatment plan for the patient which is considered as the gold standard by clinicians. However, this manual practice is inefficient, error-prone, and highly-subjective. Negative consequences of misdiagnosis can manifest as over-treatment with surgery and long-term drug therapies, or under-treatment with subsequent cancer-related morbidity or mortality. To mitigate these issues, digital pathology (DP) has been slowly gaining traction as whole slide image (WSI) technology has matured and lowered in cost. In 2017, the US Food and Drug Administration (FDA) began approving WSI systems for primary diagnosis [3]. In addition, due to the COVID-19 pandemic, more regulatory relaxation or clarification has resulted in laboratories using WSI systems in new and unconventional ways that permit pathologists to render diagnoses from home [4]. This revolution has attracted researchers to build black-box AI-based approaches to assist anatomic pathology workflows. However, the success of such approaches is dependent on large-scale generation of pathologist annotated high quality training data which is a serious bottleneck in computational pathology. Additionally, the AI systems must be interpretable and minimize the time-to-decision to achieve clinical adoption and possibly facilitate regulatory agency approvals.

In this thesis, we build computational models of already established anatomic pathology knowledge to alleviate the training data generation bottleneck and develop clinically interpretable models. Further, we also implement our computational pathology workflows on the emerging customizable AI-based compute architectures which satisfies high-throughput data processing and achieves the desired turnaround time (TAT) requirements.

In this chapter, we discuss the challenges faced by traditional pathology practice, rise of digital pathology, related work in the domain of digital and computational pathology, contributions of this thesis, and list of publications.

2.1 Challenges faced by traditional pathology practice

In traditional pathology practice, tissue samples are procured, cut and dyed, and delivered to a pathologist or a team of pathologists for analysis. Once a diagnosis has been made, there may be a review process where the slides are sent to additional pathologists for further analysis. This process causes delays while the sample is being moved and analyzed, which leads to suspended patient care [5]. A schematic of pathology workflow is shown in Figure 1, where we sub-categorized the actions that are performed in a pathology practice under three main categories; Pre-Diagnosis, Diagnosis, and Post-Diagnosis. We assert that ML tools can improve the standards for these actions by assisting the pathologists or clinicians. For example, an intelligent software tool can be used to sort new pathology cases by analyzing the digitized slides prior to diagnosis; retrospectively, it can rank the cases according to their severity or complexity. As a consequence, it will allow better case assignment among the anatomical pathologists to improve diagnostic efficiency.

In addition to inefficiencies within current analog routines of pathology, external developments are also concerning. The cancer cases are expected to rise with an increase in the aging population (increase from 1.7 million cases in 2012 to 2.3 million in 2025) [6]. As the number of cancer cases grow, the forecasted shortage of pathologists is alarming (declining from 5.7 to 3.7 per 100,000 people between 2010 and 2030) [7]. Today, major pathology practices are having problems with understaffing and increased workloads. This is even more problematic in areas that are traditionally underserved (e.g., rural areas, community hospitals, etc.) [7].

2.2 Rise of digital pathology

Recent advances in digital imaging technology and computing power have paved the way for a shift in the pathology workflow. Digital pathology (DP) field has been slowly gaining traction as whole slide image (WSI) technology has matured and lowered in cost. In 2017, the US Food and Drug Administration (FDA) began approving WSI systems for

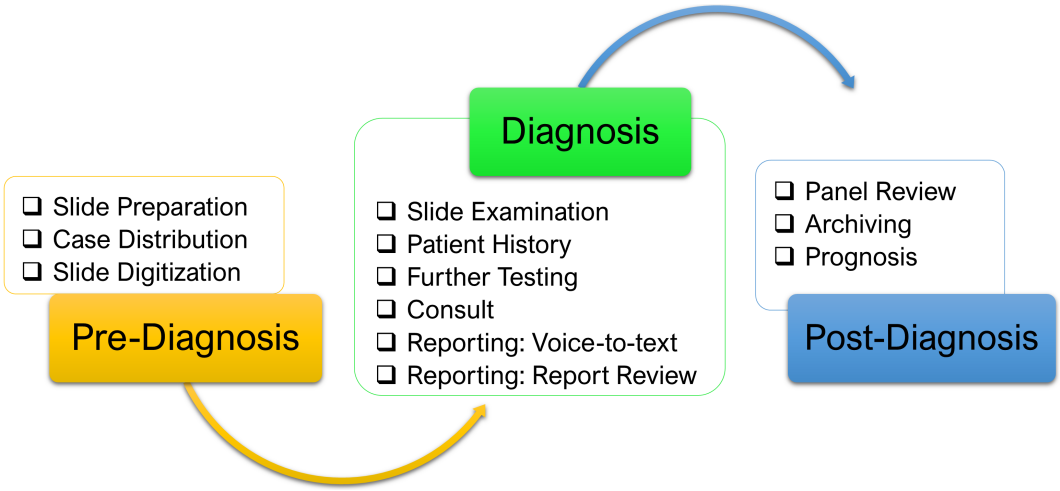

Figure 1: A generalized pathology workflow with detailed sub-categorized actions, in where AI can help elevate the current standards of a pathology practice to be more efficient and accurate.

primary diagnosis [3]. While dissection and specimen cutting still remain a manual process, various types of staining and tissue processing can be automated to be more consistent and less time-consuming [8, 9]. Additionally, the field of digital pathology (DP) seeks to solve some of the issues plaguing analogue slide sharing by enabling the high-resolution scanning and distribution of cases at rapid speeds, saving pathologists' time [10] while not affecting performance [11]. The need for DP solutions especially surfaced during the recent COVID-19 pandemic, which has posed significant challenges to the pathology profession where pathology departments seek to offer remote functionality to their staff [4, 12].

The increase in the pathologists' workload coupled with the growth of the digital pathology market, encourage significant opportunities for computational tools for anatomical pathologists and cancer pathologists in particular. DP enables the digitization of histological images, opening up many possible benefits. Having access to computing tools can provide pathologists with more quantifiable data relevant to risk assessment [13]. DP has proven to be successful for use in teleconferencing [14], allowing for simultaneous viewing of whole case files by multiple pathologists. This allows doctors the ability to virtually discuss a case

with a pathologist who may be specialized in a field relative to the patient, such as breast pathology [13]. Additionally, DP makes the process of getting a second opinion much easier and faster. The digitization of case data also makes cross-site patient data synchronization much easier [13]. Further, digitization allows for easy saving of cases for use in future studies, as well as for educational purposes. Overall, this push towards DP will lead to a greater level of pathologist involvement in patient care [8, 15].

While providing all of the aforementioned benefits, digitization further allows for the usage of intelligent computer systems to aid in the diagnostic process. This paves the way for computational pathology; using computers to process high-dimensional data, such as images or medical records, to improve health care [16]. While this field is still young and growing, we are already seeing a surge of successful applications of various computational methods to pathological data to aid in the diagnostic process. One specific framework that modern computing power allows us to use is machine learning, wherein a computer system learns to optimize for a specific task, such as region-of-interest (ROI) detection, classification, etc.

2.2.1 Traditional image processing approaches in digital pathology

There are a couple of high-quality open-source tools available online that have been used for basic image processing on WSIs or snapshots of tissue regions. Since these applications are defined as open-source, they can be freely accessed, used, and shared by anyone. The first application pathologists can easily access and use is the popular ImageJ software. It is a Java-based image processing program developed as a collaboration between the National Institutes of Health (NIH) and Laboratory for Optical and Computational Instrumentation at the University of Wisconsin [17]. It can be downloaded through their website (https://imagej.net/), where additional tutorials, use cases, and documentation is also provided. Alternatively, the "Fiji" distribution of ImageJ has the most comprehensive set of capabilities for histopathological image processing and is also available open-source (https://fiji.sc/).

The second open-source application for analyzing pathology images is QuPath [18].

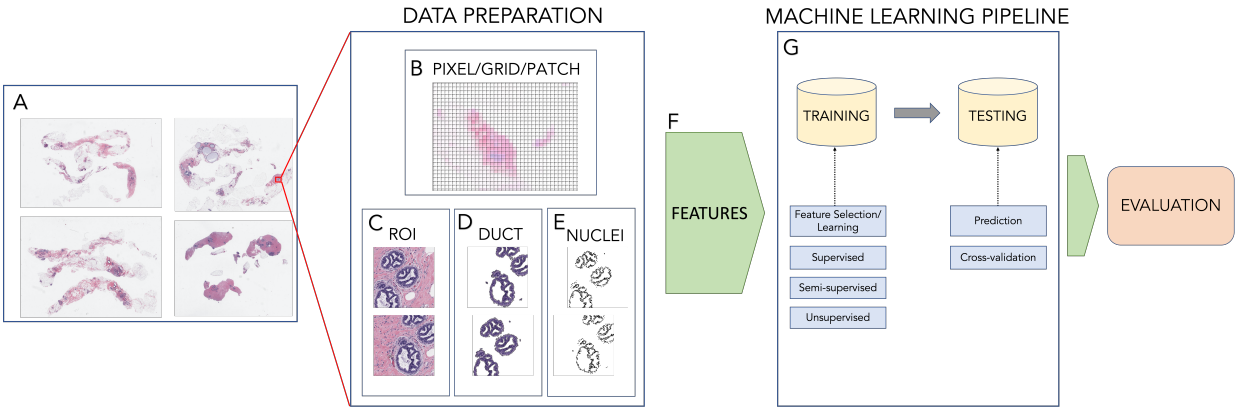

Figure 2: A typical machine learning pipeline for digital histopathological image analysis.

QuPath was developed at the Queen's University Belfast and it provides more capable tools embedded for tissue microarray analysis and for common pathology problems (such as automatic cell detection) (https://qupath.github.io/). A third well-known software is the Cell Profiler (https://cellprofiler.org/). It was developed by Broad Institute of MIT and Harvard and it enables pathologists and scientists to analyze cells in digital histopathology images [19].

2.2.2 Artificial intelligence in digital pathology

Enthusiasm for DP has attracted researchers in building computational pathology (CP) tools that aim to assist pathologists in diagnoses. An innovative framework called *pathologists' computer-assisted diagnosis* (pCAD) was projected as a result of the emerging trends in DP which motivated the application of AI towards visual assessment of the tissue slides [20]. We envision that ML tools can improve the standards for these actions by assisting the pathologists or clinicians. For example, an intelligent software tool can be used to sort new pathology cases by analyzing the digitized slides prior to diagnosis, it can rank the cases according to their severity or complexity, in result it will allow better distribution of cases among the pathologists in a practice to improve efficiency. Most of the ML tools in DP space is related to image analysis as visual assessment of tissue slides is the key for diagnosis.

Fig. 2 illustrates an ML-based generalized workflow for histopathological image analysis on a sample set of WSIs belonging to the breast organ (Fig. 2A). The tissue regions of WSIs are further segmented into one or more categories suitable for subsequent analysis (Fig. 2B-E). Few researchers have built algorithms for diagnostic inference from localized regions of interests (ROIs) (Fig. 2C) and some have worked with analyzing morphological properties of duct and nuclei (Fig. 2D-E). Several studies have also been published which uses pixels or patch based information from WSIs as illustrated in Fig. 2B. The pre-processing of WSIs is followed by feature extraction which widely varies across different studies (Fig. 2F). Several attempts have been made to automate the extraction of qualitative features which are frequently used by pathologists for clinical diagnosis. Some of these features include cell size, shape, and spatial distribution. Additionally, efforts have been undertaken in extracting morphological and pixel intensity features such as gray level co-occurrence matrix (GLCM). In contrast to extracting hand-crafted features, popular deep learning algorithms powered by convolutional neural networks (CNNs) works on the images (ROIs or WSIs) directly to extract features with the help of different filters.

Feature extraction is followed by data training which is shown in Fig. 2G. Depending upon the availability of labels, training stage falls under three broad categories: unsupervised, semi-supervised, and supervised. In order to choose the optimal hyperparameters for obtaining the best classification model, cross-validation is carried out on the training set. Next, this model is deployed on the testing set to obtain class predictions. The classification performance is quantitatively assessed through evaluation metrics (Fig. 2H). Some of the commonly used evaluation metrics are: *accuracy, precision, recall, weighted F-score, area-under-the-curve, ROC score*, etc.

2.3 Related work

Breast cancer: Some of the popular methods previously explored for classifying breast lesions are based on deep-learning (DL) architectures such as fully-convolutional networks [21], aggregating patch-level information to label breast lesion images using traditional CNNs [22],

and multi-purpose networks such as Y-Net [23], a modified U-Net architecture capable of performing segmentation and classification. An alternate approach to diagnosing breast lesion images is characterizing the "appearance" or texture using descriptors such as Local Binary Pattern (LBP), Grey Level Co-Occurences Matrices (GLCM), and Gabor filters [24–27]. In [25], the authors used GLCM, Graph Run Length Matrix (GRLM), and Euler number features to detect invasive breast cancer. In [24], authors analyzed the texture features such as GLCM, LBP, Histogram of Oriented Gradients (HOG), and several others on Breast Cancer Histopathological Image Classification data (BreakHis) [28]. Araújo, Teresa, et al. used a CNN architecture to perform a 4-class (benign, atypia, DCIS and invasive) and 2-class classifications achieving accuracies of 77.8% and 83.3% respectively using a Support Vector Machine classifier on the features extracted from a deep-learning framework [29]. They conducted their study on the the dataset containing a total of 269 high-resolution images with a pixel size of 0.42 $\mu m \times 0.42$ μm [30]. Later in 2018, Nazeri, K. et.al., used a two-stage CNN on the same dataset which was class balanced achieving an overall accuracy of 95% on the four-way classification task [31]. The authors used the first network to extract salient features of image patches and a second network to perform classification on the entire image. More recently, Lu, Ming Y., et al. and others designed a DL framework using a weakly-supervised strategy to address multi-class classification problem and tested their methods on 3-class TGCA Kidney dataset, 2-class Non-small Cell Lung Carcinoma (Adenocarcinoma and Squamous Cell Carcinoma subtypes) and Breast cancer metastasis detection on CAMELYON16 and CAMELYON17 dataset [32]. They adopted a clustering-constrained attention based multiple instance learning (CLAM) to analyze WSIs while offering high-throughput and interpretability using attention maps. A slide-level classification AUC of around 0.95 was achieved on the combined datasets used for breast cancer detection. All these methods perform pixel-level analysis and fail to account for the spatial organization and interactions between biological entities emphasized by the pathologists during diagnosis.

More recently, graph-based convolutional neural networks (GNNs) have addressed this limitation of pixel-level analysis by constructing multi-scale graph topologies to model spatial interactions among histological structures for characterizing breast tissue images [1, 33–35] and colorectal cancer images [36, 37]. In contrast to applying sophisticated DL techniques,

there have been several studies which focused on extracting histologically relevant morphological features [38, 39]. The authors in [39] performed diagnostic classification by building feature representations from the structural alterations of the breast ducts. In [38], authors computed cytological and architectural features of ductal cells to perform diagnostic classification. However, the principal advantage of our method is the ease of visual interpretability to the classification outcome which is not provided by a majority of *black-box* deep learning models [1, 35]. Further, this form of interpretability is crucial in applications such as clinical diagnosis, wherein the pathologists are required to *trust* the AI system prior to launching it in a clinical setting.

Prostate Cancer: Prostate pathology follows a classification system based on Gleason grading whose scores are decided upon the severity of tissue under inspection. In 2019, Garcia G. et. al. and team focused on gland classification by capturing patterns associated with Gleason grades (2 and 3) that suffers from maximum inter-pathologist variability [40]. For the first time, a classification workflow was built using hand-crafted features from morphological, textural, fractal dimension, and contextual information present in the glands to come in unison [41]. They do a comparative study based on using classical ML algorithms on hand-crafted features and deep-learning approach by implementing a modified VGG19 architecture on 35 WSIs. In the 3-class classification task of identifying false glands (artefacts), benign glands, and Gleason grade 3 glands, a non-linear Support Vector Machine implemented on carefully designed and selected hand-crafted features marginally outperformed the deep-learning approach to achieve an accuracy of 0.876 ± 0.026.

Lung Cancer: Another field that has attracted the attention of latest technological advancements of ML in DP is lung cancer pathology. Non-small cell lung cancer consists of two most common subtypes: Adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). In 2018, Coudray, Nicolas, et al. extended the usability of deep learning in histopathology when they successfully predicted 6 commonly mutated genes in LUAD in addition to performing usual classification of an image as cancer vs no-cancer [42]. First, a deep-learning model based on *Inception V3* was deployed to perform tumor classification of lung on 1,634 WSIs obtained from Genomic Commons database [43]. The classification task of cancer vs no-cancer achieved a state-of-the-art AUC of ~ 0.99 and binary classification of tumor type

(LUAD vs LUSC) achieved an AUC of 0.97. Interestingly, $\sim 83\%$ of the 54 TCGA images incorrectly classified by at least one pathologist was corrctly classified by the deep-learning model. Second, a novel approach was undertaken to predict gene mutations in LUAD slides by modifying the network architecture which scientifically proved that gene mutations would affect the tumor cells pattern on a lung cancer WSI. Among the six mutated genes that were predicted using image data (EGFR, STK11, FAT1, SETBP1, KRAS, and TP53), STK11 prediction achived the best performance of ~ 0.85 AUC.

Later in 2019, Wei, Jason W., et al. used ResNet architecture to classify histologic patterns observed in LUAD which is sometimes challenging making differential diagnosis subjective [44,45]. Following the guidelines published by the WHO in 2015, the authors built a CNN model to identify *lepidic, acinar, papillary, micropapillary, and solid* patterns. Three pathologists on team annotated 4161 ROI crops for training, 1068 patches for development, and 422 WSIs for testing to have at least one of the five histologic patterns or belong to benign case. This was the first study which conceptualized an automated classification of LUAD patterns which is beneficial to assist the pathologists for decision-making. Further, among the spectrum of discordances across all predominant patterns, 39.5% disagreement was observed between lepidic and acinar patterns which are assigned as low-grade and intermediate-grade respectively. In the classification task of identifying histologic patterns, the deep-learning model achieved an average kappa score of 0.525 which was greater than 3 pathologist's individual performances of 0.454, 0.515, and 0.514 respectively.

2.4 Thesis contributions

Pathological diagnosis is considered as the gold standard by clinicians. However, this manual pathology practice can be inefficient, error-prone and highly subjective. To mitigate these issues, digital pathology is gaining traction which has attracted researchers to build black-box AI-based approaches intended to assist anatomic pathology workflows. The success of such approaches is dependent on large-scale generation of pathologist annotated high quality training data which is a serious bottleneck in computational pathology. My doctoral

work is motivated by the fact that the success of AI-based computational pathology applications must be interpretable, minimize the time-to-decision and can integrate into anatomic pathology workflows to achieve clinical adoption and possibly facilitate regulatory agency approvals. The overarching goal of this thesis is to build computational models of already established anatomic pathology knowledge to alleviate the training data generation bottleneck and develop clinically interpretable models. Additionally, we demonstrate a proof-of-concept study of integrating AI-based applications in anatomic pathology workflows on the emerging customizable AI-based architectures which satisfies high-throughput and achieves required turnaround time.

2.4.1 Outline

Chapter 2: This chapter presents an approach to build analytical models to capture tissue features that aid in the differential diagnosis of breast biopsies and evaluates the inferential power of these hand-crafted features. These features are assembled following guidelines in the WHO classification of the tumors of the breast (an essential reference for pathologists, clinicians, and researchers) and in consultation with pathologists on team.

Chapter 3: This chapter presents a prototype-driven machine learning framework for the differential diagnosis of breast biopsies which is amenable to clinical interpretability.

Chapter 4: This chapter extends the strategies outlined in chapter 3 on a broad spectrum of breast biopsies. It also presents a new approach for the automatic selection of class-specific prototypes, analytical modeling of additional tissue features, and an improved prototype-driven ML framework to further enhance the diagnostic classification performance.

Chapter 5: This chapter presents a proof-of-concept study to integrate AI-based anatomic pathology applications in clinical settings by training our computational pathology pipelines on remote customizable high-performance AI-enabled compute architectures provided by state-of-the-art data centers and applying the pipelines on edge devices for real-time clinical applications. For demonstration, our pipeline detects histological structures in breast tissue and classifies them into two diagnostic categories, high-risk and low-risk.

Chapter 6: This chapter summarizes the key contributions and findings of my thesis and discusses the clinical impact.

2.4.2 List of publications

Digital Pathology

- **Parvatikar, A.**, Falkenstein, B., Ramanathan, A., Tosun, A. B., Fine, J. L. & Chennubhotla, S. C. (2022, June). A prototype-driven computational pathology pipeline based on analytical modeling of histological structures for differential diagnoses of breast biopsies. In *Computer Vision and Pattern Recognition Conference* - under review
- **Parvatikar, A.**, Choudhary, O., Ramanathan, A., Jenkins, R., Navolotskaia, O., Carter, G., Tosun, A. B., Fine, J.L. & Chennubhotla, S. C. (2021, September). Prototypical Models for Classifying High-Risk Atypical Breast Lesions. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 143-152). Springer, Cham.
- **Parvatikar, A.**, Choudhary, O., Ramanathan, A., Navolotskaia, O., Carter, G., Tosun, A. B., Fine, J.L., Chennubhotla, S. C. (2020, October). Modeling Histological Patterns for Differential Diagnosis of Atypical Breast Lesions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 550-560). Springer, Cham.)

Molecular Biophysics

- Ramanathan, A., Ma, H., **Parvatikar, A.**, Chennubhotla, S. C. (2021). Artificial intelligence techniques for integrative structural biology of intrinsically disordered proteins. Current Opinion in Structural Biology, 66, 216-224.
- Ramanathan, A., **Parvatikar, A.**, Chennubhotla, S. C., Mei, Y., Sinha, S. C. (2020). Transient Unfolding and Long-Range Interactions in Viral BCL2 M11 Enable Binding to the BECN1 BH3 Domain.

3.0 Modeling tissue features for differential diagnosis of breast biopsies

3.1 Chapter summary

The goal of this thesis chapter is to build analytical models for a dictionary of tissue features that aid in the differential diagnosis of atypical breast biopsies and evaluate the inferential power of these hand-crafted features. Diagnosis of high-risk atypical breast biopsies is challenging and remains a critical component of breast cancer screening, presenting even for experienced pathologists a more difficult classification problem than the binary detection task of cancer *vs* not-cancer. Following guidelines in the WHO classification of the tumors of the breast (an essential reference for pathologists, clinicians and researchers) and in consultation with our team of breast sub-specialists ($N = 3$), we assembled a visual dictionary of sixteen tissue features (e.g., cribriform, picket-fence - confined within a duct), a subset that pathologists frequently use in making complex diagnostic decisions of atypical breast biopsies. We invoke parametric models for each feature using a mix of *unary*, *binary* and *ternary* features that account for morphological and architectural tissue properties. We use 1441 ductal regions of interest (ROIs) extracted automatically from 93 whole slide images (WSIs) with a computational pathology pipeline. We collected diagnostic labels for all of the ROIs: normal and columnar cell changes (CCC) as low-risk benign lesions (=1124), and flat epithelial atypia (FEA) and atypical ductal hyperplasia (ADH) as high-risk benign lesions (=317). An example ROI for each of the diagnostic category is shown in Fig 3. We generate likelihood maps for each tissue feature across a given ROI and integrate this information to determine a diagnostic label of high- or low-risk. This method has comparable classification accuracies to the pool of breast pathology sub-specialists. Further, this approach enables a deeper understanding of the discordance among pathologists in diagnosing atypical breast biopsies.

3.2 Introduction

3.2.1 Background

Benign breast biopsy diagnoses account for approximately a million cases annually [46]. The patients are subjected to additional screening procedures depending upon the relative risk associated with the diagnostic subtypes of the benign biopsies (e.g., high-risk is associated with atypical hyperplasia) [47, 48]. Over half of the patients diagnosed with atypical hyperplasia, which is histologically further classified into atypical ductal hyperplasia (ADH) and atypical lobular hyperplasia (ALH), contract breast cancer within 10 years of screening, thereby demanding an accurate diagnosis of these precursor lesions.

On the contrary, a recent clinical study showed significant levels of disagreement in differential diagnosis of cases with atypia (48 - 56%) resulting in *overinterpretation* (subjecting patients to unnecessary medical procedures) and *underinterpretation* (subjecting patients to no treatment) [49]. The underlying difficulty in classifying atypia from benign lesions stems from the fact that diagnostically relevant histopathological patterns overlap in the spectrum of low- to high-risk lesions, complicating the decision-making process (Fig. 4). In the interest of patient management, it is convenient to stratify patients into "low- / high-risk" categories based on their histological evidence and associated risk-factor [48].

3.2.2 Previous work

Previously, we have approached this problem in an unsupervised manner by simply encoding cytological properties of nuclear atypia and integrating them with the spatial distribution of the nuclei in relationship to stroma and lumen components of breast tissue (i.e., architectural patterns) [38]. Measured in terms of *recall* of high-risk lesions, *the classification performance reported here (0.76) is a significant improvement over our previous approach (0.69).* Although there are studies in the machine classification of breast tumors [50–53], many of these do not include diagnostically challenging ADH cases nor provide directions for a *computational* understanding of the structural changes in the breast tissue triggered by atypia and other malignancies. To the best of our knowledge, our work in analytically

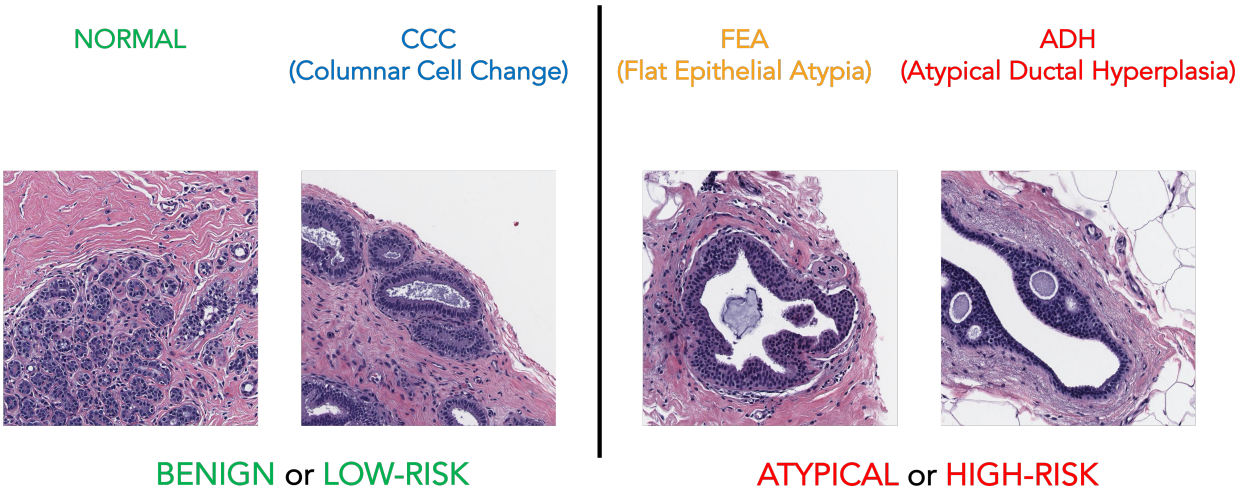

Figure 3: Examples of benign breast lesions.

modeling a visual pattern dictionary that traditionally defines the standards on tumor classification/ nomenclature for pathologists worldwide is the first of its kind.

3.2.3 Contributions

Following guidelines in the WHO classification of the tumors of the breast [54] (an essential reference for pathologists, clinicians and researchers) and in consultation with our team of breast pathology sub-specialists ($N = 3$), we assembled a visual dictionary of a *subset* of histological patterns/ tissue features that aid pathologists in undertaking differential diagnoses of atypical breast biopsies (Fig. 1). Further, we built analytical models for each tissue feature using a mix of unary, binary and ternary features that account for cytological (nuclear shape and orientation, lumen shape), architectural (intraductal), and spatial-extent details of low- and high-risk lesions.

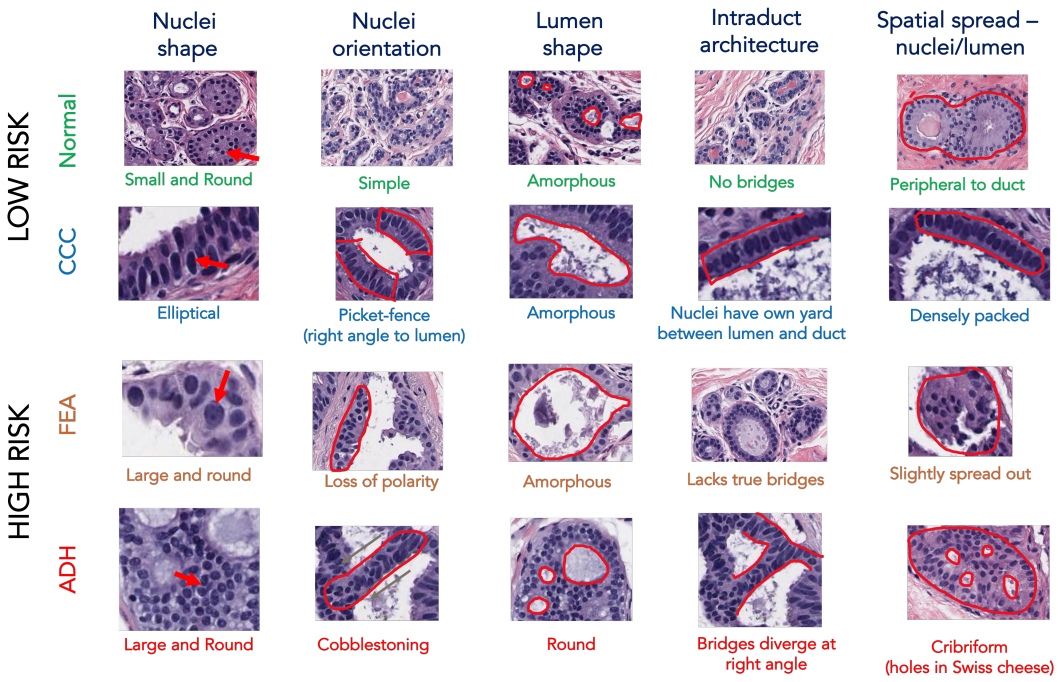

Figure 4: An organization of tissue features frequently used by the pathologists for diagnosis of atypical breast biopsies on the basis of nuclei shape, orientation of the nuclei, shape of the lumen, intra-ductal architecture, and spatial spread of the nuclei and lumen. The frequency of occurrence of these tissue features is different for low- and high-risk categories. Hence, the diagnosis starts becoming subjective leading to discordance among pathologists.

3.3 Methods

3.3.1 Segmenting ducts, lumen and nuclei

We designed a new algorithm for segmenting ducts, lumen and nuclei on large scale WSIs. To start with, WSI images stored in RGB format are color deconvolved into their respective stain intensities namely, hematoxylin and eosin by using the color deconvolution plugin in ImageJ [55]. The stain colors are further normalized with a reference dataset to standardize color variations for downstream processing. To ease the computational burden

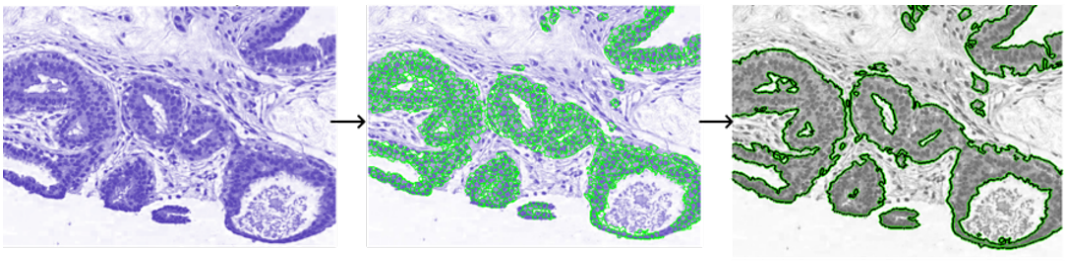

Figure 5: An overview of the duct segmentation strategy: (left) Color deconvolved and stain normalized hematoxylin stain image from a H&E sample. (middle) We segment this image into superpixels. Using machine learning (SVM), we predict the stain-labels of the superpixels and the superpixel-pairs that lie inside a duct. In our model, a duct is defined by superpixel-pairs that are moderate to heavy stained, and are predicted to lie inside a duct. The predicted superpixels are shown in green and are overlaid on the original image. (right) A region-based active contour segmentation is run to separate foreground (ducts) from the background (rest of the image) based on the means of the hematoxylin stain in the two regions. This segmentation is based on Chan-Vese segmentation algorithm.

of detecting ducts in a WSI, we built a Gaussian pyramid of the hematoxylin intensity of the WSI. The hematoxylin intensity image at the coarsest level of the pyramid is broken into non-overlapping superpixels (area = 300 pixels), which are sets of connected pixels with similar intensity values, using simple linear iterative clustering (SLIC) algorithm [56]. The innovative steps of our algorithm are in assigning probabilities for the presence of a duct given a pair of nearby superpixels ("context-ML") and further identifying all those superpixels that are "moderate-to-heavily" stained as the ones inside a duct ("stain-ML"). Using the superpixels identified as initial guesses, we perform a region-based active contour segmentation [57] that separates foreground (ducts/lumen) from the background (rest of the image). For hematoxylin and eosin stained images, the cost-function for the active contour is driven by the difference in the mean of the hematoxylin stain in the foreground and background regions. For example, two superpixels that have a high probability of being inside a duct have roughly the same stain ("moderate to heavy stain") and their boundaries

are merged iteratively by the active contour optimization. Often ducts appear as "clusters" and to segment these we run the region-based active contour on the *probability* map returned by the context and stain-ML models. The probability maps impute non-zero probabilities to ducts and regions bridging them, and a region-based active contour model run on the probability map is more successful in delineating a cluster of ducts.

To identify lumen, we use context- and stain-based ML models to select image regions that are not part of the ducts – non-tissue areas on the WSI, connective tissue areas and lumen. We perform connected-component analysis to select and exclude large components, likely to correspond to non-tissue and connective tissue areas. The remaining components highlight lumen regions that lie inside ducts and are verified visually in our training images. To identify and segment nuclei inside a duct, we first select parts of image lying inside a duct, then use ImageJ to threshold intensities and finally run watershed to delineate the nuclear boundaries.

3.3.2 Building analytical models of tissue features

We invoke parametric models for tissue features using a mix of *unary*, *binary* and *ternary* features as shown in Fig. 6. The colorbars over each feature in Fig. 6 indicate the lesion where the feature is most likely to be found, e.g., large and round nuclei are often found in high-risk lesions, small and elliptical nuclei in low-risk lesions, and cribriform feature is exclusive to ADH lesion.

A schematic representation of a subset of the local visual dictionary comprising of 16 tissue features is shown in Fig. 7. This dictionary is formed upon consulting the expert pathologists and studying pathologist's guiding references such as the WHO classification of tumors of the breast [54]. The elements of this dictionary are organized based on the nuclei size and shape (*small*, *large*, *round*, *elliptical*, *large-round*, *small-elliptical*), and spatial spread of the nuclei (*crowded*, *spaced*). These morphometric patterns are used to identify additional tissue features such as *spaced-large*, *crowded-small*, *spaced-small*, *crowded-elliptical*, *spaced-round*, and *large-round-spaced*. Further, the orientation of the nuclei and shape of the lumen is useful to identify higher-order patterns such as *picket-fence* and *cribriform*.

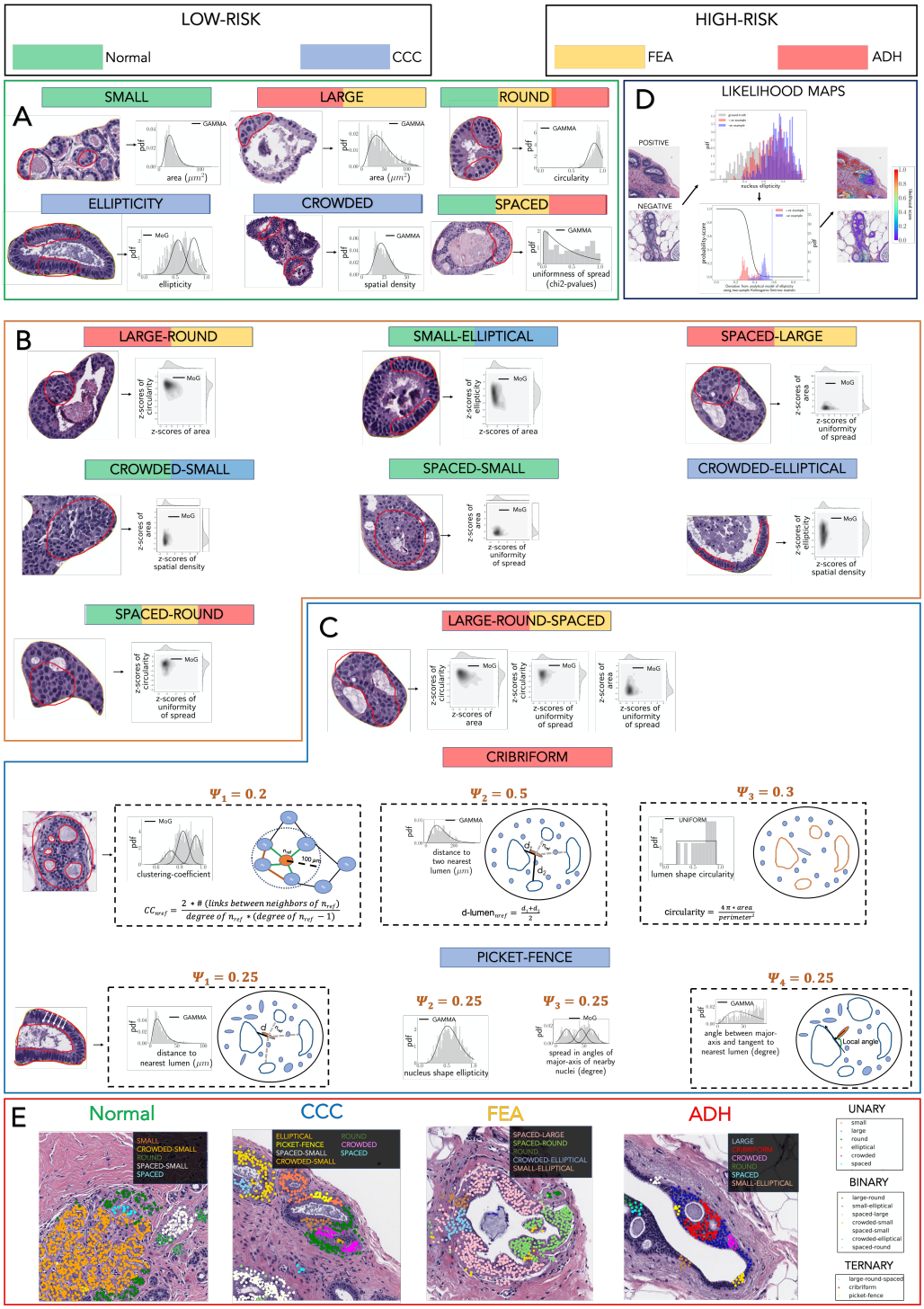

Figure 6: Analytical models of tissue features in the form of (A) unary, (B) binary and (C) ternary features. (D) Computing likelihood scores to reveal (E) dominant features in representative images of low- and high-risk biopsies.

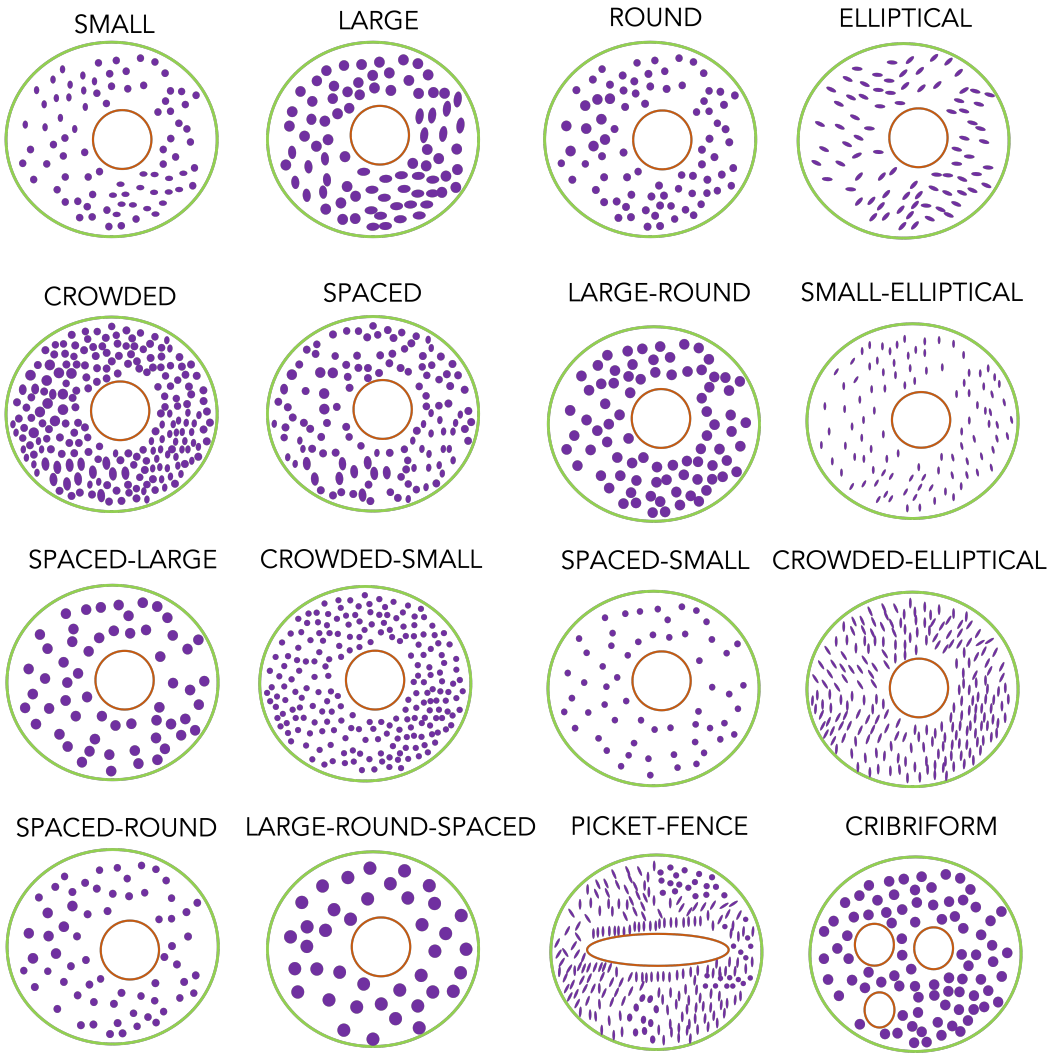

Figure 7: Schematic representation of 16 tissue features useful to diagnose atypical breast biopsies.

The frequency of occurrence of these features is different for each diagnostic category which is a major contribution to discordance among pathologists [49]. Since the measurements of tissue features are amorphous, we take the approach of building a quantitative model and tune the parameters of this model to *match* with the consensus diagnosis of tissue features (*templates*).

Unary Features: In consultation with the breast pathologists on our team, we selected

a spectrum of morphological features on the basis of size, shape, and spatial spread around each nucleus. Nuclear size (quantified using area) is known to provide diagnostic cues in pathological grading [58–61], with groups of small and large nuclei having a propensity to belong to low-risk and high-risk lesions respectively [62]. To build analytical models of *small* and *large*, we first construct a histogram of nuclear areas obtained from an ensemble of ROIs showing prototypical example regions within a duct containing small and large nuclei (Fig. 6A) and model this histogram with a Gamma distribution.

Next, nuclear shape has been identified as diagnostically meaningful, e.g., CCC lesion shows dominant elliptical nuclei [63]. We quantitate this feature with *roundness* measured as $(4\pi \times \text{area})/\text{perimeter}^2$ and *ellipticity* given by the ratio of length of minor-axis to the length of major-axis. Roundness ranges from 0 (irregular star-like appearance) to 1 (perfect circle), while ellipticity characterizes the "flatness" of an object with lower values denoting highly elliptical nuclei (Fig. 6A). In each case, because of the intrinsic heterogeneity of these measurements, we consider a spatial neighborhood around each nucleus, and model the distributions of roundness with a Gamma distribution and ellipticity with a 2-component mixture of Gaussians (MoG) model (Fig. 6A).

Finally, several studies have shown that studying the spatial organization of nuclei provides insights into the abnormalities of cells which might eventually lead to malignancy. For instance, the nuclei arrangement in a CCC lesion frequently exhibits crowding and/or overlapping [64, 65].

However, for cases belonging to high-risk atypical lesions (FEA and ADH) the nuclei tends to be uniform and evenly-spaced [64, 66]. To quantify "crowding" around each nucleus, its average distance to 10 nearest nuclei is computed. An analytical model of *crowdedness* is constructed by considering local ROIs within a duct where clusters of nuclei show significant crowding behavior and then computing its spatial density. To capture evenly spaced/ uniform dispersion pattern around a nucleus, we start by placing a regular grid of size 3×3 centered at a reference nucleus and measure the density of 20 neighboring nuclei by counting the population of nuclei in each grid cell as described in [67]. We then compare this observed population against expected number of nuclei under the *complete spatial randomness* hypothesis which asserts the occurrence of points (here nuclei) within grids in a

random fashion analogous to a Poisson point process using a χ^2-test statistic and acquiring the corresponding p-value using the χ^2 distribution table. Larger the p-value, greater is the likelihood of observing a uniform/ evenly spaced dispersion of nuclei around the reference nucleus.

Binary Features: Although, the unary features show some inferential strength (indicated by the color bars on top of each feature in Fig. 6), a pathologist typically makes an informed decision by paying attention to the pairwise combinations of such features. For instance, a CCC lesion (low-risk) exhibits crowded and elliptical nuclei arrangement. A high-risk lesion tends to display a greater likelihood of large-round, spaced-large, and spaced-round nuclei. A lesion showing majority regions of small nuclei coupled with crowded and/ or spaced behavior is representative of a normal duct. In our study, we considered 7 such binary features obtained from pairwise combinations of unary features which is shown in panel Fig. 6B. We take z-scores for each unary feature, and model the joint distribution of z-scores from the feature pair with a two-component, two-dimensional mixture of Gaussian distribution.

Ternary Features: Some of the diagnostically relevant tissue features are best represented by a combination of more than two unary features. *I. Large-Round-Spaced*: We take z-scores from each feature, i.e., large, round and spaced, and build a three-component, three-dimensional mixture of Gaussian model using ground-truth examples. *II. Cribriform*: This feature is characterized by polarization of epithelial cells within spaces formed by "almost" circular multiple lumen (> 2) which are 5-6 cells wide and whose appearance closely resembles to "holes in Swiss cheese". This complex architectural feature can be identified by analytically modeling three sub-features: *clustering coefficient*, distance of the nucleus from two nearest lumen, and circularity of the lumen (computed using ImageJ) adjacent to the nucleus. The polarization of epithelial cells around lumen is characterized by clustering-coefficient and is computed by following the method described in [68] and is illustrated in the second row of Fig. 6C. A group of nuclei occupying the spacing between two lumen has a tendency to show cribriform pattern around them. Thus, we measure the average distance between each nucleus to the nearest two lumen and model its distribution using gamma function (see middle row of Fig. 6C). The final likelihood for cribriform feature is obtained

from the weighted sum of the likelihood scores of sub-features. We performed grid search on the mixing coefficients to learn that the likelihood scores from the three sub-features should be mixed in the proportion of 0.2, 0.5, and 0.3 respectively. *III. Picket-Fence*: This spatial arrangement is recognized from a group of crowded elliptical nuclei oriented perpendicular to the basement membrane (lumen). The analytical model of this higher-order visual feature can be obtained by constructing parametric models of four simple sub-features: distance of a nucleus to nearest lumen, nuclear ellipticity, a spread in the angle of major-axis of 10 nearby nuclei, and its local angle with respect to the basement membrane as shown in the last row of Fig. 1C. Since, each sub-feature contributes equally to observing this ternary feature, we chose to assign a mixing coefficient of 0.25 in combining the likelihood scores from the four sub-features to determine the presence of a picket-fence pattern.

3.3.3 Computing likelihood scores of tissue features

As discussed in the previous section, the analytical models of the tissue features are probability distributions. For example, a cytological feature like nuclear ellipticity for a given nucleus inside an ROI can be assigned a probability value under the mixture of Gaussian model for the template ($\mathcal{G}_t$) image derived in Fig. 6A. However, accurate measurements of ellipticity values are greatly influenced by the precision with which nuclei boundaries are segmented. This naturally leads to heterogeneity in the estimates of ellipticity. To account for this heterogeneity, we chose to compare the neighborhood around the reference nucleus to the neighborhood in the template image. In particular, we model the ellipticity values in the neighborhood of the reference nucleus with a new mixture of Gaussian model ($\mathcal{G}_n$) and then compare model parameters of $\mathcal{G}_n$ with $\mathcal{G}_t$. We used two different distance measures for comparing the model parameters: *Kullback-Leibler divergence* for mixture of Gaussians and *two-sample Kolmogorov Smirnov test* for unimodal Gamma distributions. Small distances imply greater evidence for the presence of the tissue feature. We turn the distances into a likelihood score by an inverted S-function as shown in Fig. 6D. This process is carried out in a similar fashion for every feature present in the visual dictionary.

3.3.4 Preliminary strategy for differential diagnosis

We adopt a non-linear strategy here, similar to what expert pathologists do, in that we find sub-regions within ROI by non-maxima suppression (threshold value of 0.85 on the likelihood scores) where the evidence for one or more of the unary, binary or ternary feature is dominating. Fig. 6E provides a visual illustration of the likelihood maps of dominant tissue features in representative images of low- and high-risk biopsies. Low-risk breast biopsies show dominant islands of *round*, *small*, *spaced*, and *spaced-small* in a normal ROI and *elliptical*, *round*, *spaced-small*, *crowded-small*, and *picket-fence* neighborhoods in a CCC ROI. In comparison, high-risk biopsies show dominant regions of *spaced-large*, and *spaced-round* in a FEA labeled ROI and compelling strengths for *large* and *cribriform* features along with traces of *crowded* and *spaced* in ADH labeled ROI. These features validate the canonical forms shown in Fig. 6A-C.

Having identified dominant unary, binary and ternary feature regions, we use 3 descriptive statistics: median value of the likelihood scores of all the nuclei found in each sub-region, median number of nuclei found in each sub-region and the number of sub-regions.

This is calculated for each one of the unary, binary and ternary features (total = 16), thereby obtaining a 48 column feature vector for a single image. We computed feature vectors for all 1441 labeled duct ROIs which resulted in 834×48 size feature map used to train the classifier and 607×48 data matrix for testing. To analyze the benefit of including binary and ternary features we further slice the 48 column feature vector to be suitable for three scenarios: unary (U) only, unary and binary (U-B), and unary, binary, and ternary features (U-B-T). Due to inherent training and testing class imbalance, which reflects the real-world prevalence statistics of atypical lesions, we upsampled high-risk examples using SMOTE technique [69].

Prior to classifying the lesions, we pay close attention to the presence of cribriform feature, a symbolic visual primitive of ADH (a high-risk) category [66, 70, 71]. ROIs predicted to show cribriform feature are classified as high-risk, if the number of nuclei forming the cribriform sub-region is greater than 8 (hyperparameter optimized over the training data). The reduced dataset, devoid of cribriform, is tested for each of the scenarios (U, U-B, and

Table 1: Performance measures with U, U-B and U-B-T feature sets and comparison with other baseline strategies (including majority classification and average single expert pathologist assessment) and deep-learning models.

	Baseline		Comparisons			U	U-B	U-B-T
Models	Majority	Expert	Lenet	Overfeat	Alexnet	LR		
Recall	0	**0.77**	0.23	0.31	0.4	0.56	0.59	**0.76**
Specificity			0.88	0.84	0.86	0.64	0.69	0.63
TN			475	451	462	345	373	336
FN			54	48	42	31	29	17

U-B-T) with logistic regression (LR), support vector machine (SVM), random forest (RF), and gradient boosted classifier algorithms. The best model was chosen by optimizing the parameters using GridSearchCV based on precision, recall, and F-scores and then performed a 10-fold stratified cross-validation to check for overfitting. In optimizing the hyperparameters, the operating point was selected to value recall over precision reflecting the clinical decision objective where a false negative outcome is penalized higher than a false positive.

3.4 Results

3.4.1 Dataset

We used 1441 ductal ROIs extracted automatically from a computational pathology pipeline (see Section 3.3.1) from 93 WSIs which were scanned at $0.5\mu m$/pixel resolution at 20× magnification captured using Aperio ScanScope XT microscope. Among these, the *training* set constituting 834 ROIs were diagnostically labeled by a single sub-specialist pathologist (SP1), while a consensus diagnosis was achieved for the remaining 607 *testing* set ROIs with a pool of 3 breast pathology sub-specialists (SP1, SP2, and SP3). The diagnostic labels include: "Normal", "CCC", "FEA", or "ADH", which were further regrouped into

two classes: low-risk (Normal and CCC) and high-risk (FEA and ADH). While the training set comprised of 587 low-risk and 247 high-risk examples, the test set included 537 low-risk and only 70 high-risk cases, leading to the issue of class-imbalance and the choice of *recall* of high-risk lesions as a performance metric for the classification strategy. We are reporting recall to emphasize correct detection of high-risk lesions, as the consequence of misdiagnoses (false negative) implies increased chance of developing cancer for lack of providing early treatment. The concordance among the 3 pathologists in labeling the test set was moderate (Fleiss' kappa score of ≈ 0.55 [38]).

3.4.2 Classification performance

Table 1 shows the outcome of the differential diagnosis strategy that we implemented using the three feature sets: U, U-B, and U-B-T. The average performance of the three pathologists informs the baseline with single expert pathologist [38]. We tested with Logistic Regression (LR), Random Forest, and SVM with SMOTE and cross-validation parameter scanning. LR performed the best. SVM and Random Forest misclassified high-risk images containing large/round/spaced nuclei (a high-risk feature, see Fig. 6) as low-risk. This resulted in lower recall compared to LR, which was successful in capturing these features. Additionally, we tested approaches with deep learning: Lenet [72], Alexnet [73], and Overfeat [74]. For training deep learning networks the ROIs obtained from duct segmentation were downscaled to 512×512 and the class imbalance was handled by performing data augmentation through rotations and reflections. Further, these class-balanced batches were trained using 3 networks for 3,000 epochs.

3.5 Discussion

We find progressive improvement in the performance from U to U-B to U-B-T feature sets, achieving highest recall of 0.76 which outperforms the majority classification (obtained by assigning all cases to the majority label of low-risk, thereby having a recall of 0) and has

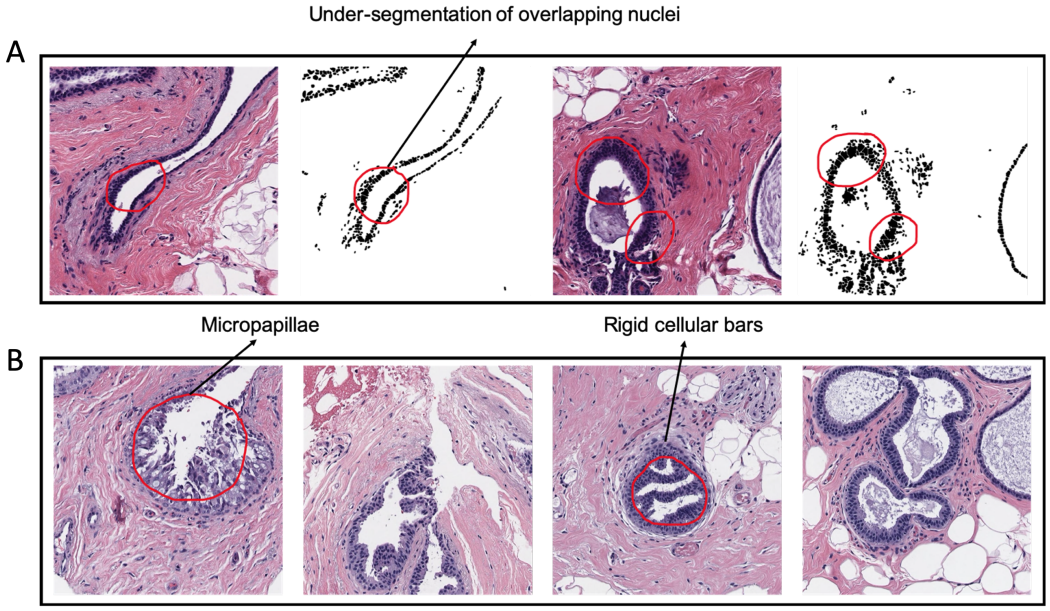

Figure 8: (A) **False positives**: The nuclei segmentation procedure sometimes fails to perfectly segment overlapping and/or heavily crowded nuclei leading to over-segmentation and thus classifying the above images as high-risk. (B) **False negatives**: The non-inclusion of additional distinguishing characteristic of ADH (micropapillae and rigid cellular bars) which is shown above led to wrongly classifying above images as low-risk.

a comparable performance to the assessment made by single breast pathology sub-specialist (SP1). Our approach with ∼150 parameters is readily amenable to explainability which cannot be delivered by current deep learning (DL) methods (∼10-50 million parameters and large training data). To the best of our knowledge, there are no widely reported DL methods for borderline of atypical breast biopsies, but an abundance of these algorithms for cancer vs no-cancer datasets. To further promote research in the use of DL for borderline cases, we chose to continue working with the same set of networks as used in our previous work [38], with one exception of incorporating improved duct segmentation component. The average computation time to obtain likelihood scores and return a diagnostic label is 1 minute for an image with 1000 nuclei on a single 2.4 GHz processor. Further, we analyzed ML model misclassifications generated from our method which is illustrated in Fig. 8. In some low-risk

examples, the accurate identification of class-specific tissue feature (e.g. small, crowded-small) is missed due to the under-segmentation of overlapping nuclei resulting in a wrong classification (false positive). However, we observed that U-B-T features (best recall) misclassified 24% of the high-risk images as low-risk (false negative). Upon investigation, we found that majority of the wrongly classified images had rigid cellular bars and micropapillae (club-shaped lumina) architecture, two additional distinguishing characteristics of ADH [54] not included in the dictionary for the present study. We anticipate that successful inclusion and analytical modeling of additional tissue features will improve the classification performance. Further, this approach lacked a good learning strategy to infer the diagnostic label from the clusters of strong histologically relevant features from the ductal ROIs which is the major focus of the next chapter.

4.0 Prototypical models for classifying high-risk atypical breast biopsies

4.1 Chapter summary

As described in the previous chapter, high-risk atypical breast biopsies are a notoriously difficult dilemma for pathologists who diagnose breast biopsies in breast cancer screening programs. To address the limitations of the learning strategy presented in chapter 3, here we reframe the computational diagnosis of atypical breast biopsies as a problem of prototype recognition on the basis that pathologists mentally relate current tissue features to previously encountered features during their routine diagnostic work. In an unsupervised manner, we investigate the relative importance of ductal (global) and intraductal features (local) in a set of pre-selected prototypical ducts in classifying atypical breast biopsies. We conducted experiments to test this strategy on subgroups of breast biopsies that are a major source of inter-observer variability; these are benign, columnar cell changes, epithelial atypia, and atypical ductal hyperplasia in order of increasing cancer risk. Our model provides clinically relevant explanations to its recommendations, thus it is intrinsically interpretable, which is a major contribution of this work. Our experiments also show state-of-the-art performance in recall compared to the latest deep-learning based graph neural networks (GNNs).

4.2 Introduction

4.2.1 Background

As elaborated in 5.2, breast cancer screening and early detection can help reduce the incidence and mortality rates [2]. Although effective, screening relies on accurate pathological diagnoses of breast biopsies for more than one million women per year in the US [49, 75]. Most benign and malignant biopsy diagnoses are straightforward, but a subset are a significant source of disagreement between pathologists and are particularly troublesome for

clinicians. Pathologists are expected to triage their patients' biopsies rapidly and accurately, and they have routines for difficult or ambiguous cases (e.g., second-opinion consults, additional stains). Still, disagreement remains an issue; while the literature suggests that diagnosis should be straightforward if diagnostic rules are followed [76], concordance remains elusive in real world diagnosis, reported in one study as low as 48% [49].

4.2.2 Previous work

Although there have been numerous efforts in using prototypes for scene recognition [77–79], to date, this idea has not been explored to classify breast biopsies. One of the first studies to detect high-risk breast biopsies was proposed in [38] which was based on encoding cytological and architectural properties of cells within the ducts. The work in [39] used structural alterations of the ducts as features to classify breast lesions into benign, atypia, ductal carcinoma in-situ (DCIS), and invasive. A different approach was proposed in [80], where the authors used analytical models to find clusters within ROIs with strong histologically relevant features. However, their approach lacked a good learning strategy to infer the diagnostic label from these clusters. Further, two recent studies approached this problem using attention-based networks to generate global representation of breast biopsy images [81] and biological entity-based graph neural networks (GNNs) [33] (also tested as a baseline method). Both methods were tested on an unbalanced dataset like ours and both reported low performance measures in detecting high-risk breast biopsies.

4.2.3 Contributions

In this study, we focus on modeling and differentiating difficult breast biopsy subtypes: atypical ductal hyperplasia (ADH), flat epithelial atypia (FEA), columnar cell changes (CCC), and Normal (including usual ductal hyperplasia (UDH) and very simple non-columnar ducts). Our approach originates from the method that pathologists practice, which is to carefully assess alterations in breast ducts before making diagnostic decisions [2, 54, 82]. Pathologists continually observe tissue features and make decisions supported by the morphology. In doing so, they look at an entire duct (*global*) and features within portions of the duct (*lo-*

cal) striving to generate mental associations with prototypical ducts and/or their parts they previously encountered in training or clinical practice. We propose an end-to-end computational pathology model that can imitate this diagnostic process and provide explanations for inferred labels.

We hypothesize that ductal regions-of-interest (ROIs) having similar global and local features will have similar diagnostic labels and some features are more important than others when making diagnostic decisions. Our approach is related to other prototype-driven image recognition systems that favor visual interpretability [77–79].

To the best of our knowledge, our work is the first one to: (1) use a diverse set of concordant *prototype* images (diagnostic class agreed by all 3 pathologists) for learning, (2) characterize clinically relevant global and local properties in breast histopathology images, and (3) provide explanations by measuring the relative importance of prototype features, global and local, for the differential diagnosis of breast biopsies. We also show that our approach facilitates diagnostic explanations with accuracies comparable to the state-of-the-art methods.

4.3 Methods

4.3.1 Machine learning framework

3.1 Machine learning framework: In this chapter, we develop an end-to-end computational pathology system that models the entire duct (global) and the tissue features occurring within selective portions of the duct (local) with the goal of generating associations with similar ducts and/or parts (prototypical). *We hypothesize that images with one or more ducts having similar global and local features will have similar diagnostic labels and some features are more important than others when making diagnostic decisions.* We will first introduce a composite mapping function to learn the relative importance of global and local features in a prototype set $\mathcal{P}$ for differential diagnoses:

$$h(x;\mathcal{P}) = \sum_{k=1}^{p} \beta_k \left[\exp^{-\lambda_k^G c_k(x)} \times \prod_{j=1}^{m_k} \exp^{-\lambda_{kj}^L f_{kj}(x)} \right]. \quad (1)$$

Here $h(x;\mathcal{P})$ captures the association of a previously unseen image x with a set of prototype images in $\mathcal{P}$. The index k varies over the images in the prototype set $\mathcal{P}$ (size = p), while j indexes over a local feature set (size = m_k) in a given prototype image indexed by k. β_k determines if the resemblance of a previously unseen image x to the prototype k has a positive (β_+) or negative influence (β_-). λ_k^G and λ_{kj}^L indicate the relative importance of global (ductal) and local (intra-ductal) features in the prototype k respectively. The relative importance can be imagined as a distance measure, so we enforce non-negativity constraints on λ_k^G and λ_{kj}^L values. The functions $c_k(x)$ and $f_{kj}(x)$ compute the global and local differences respectively between x and the prototype set $\mathcal{P}$ (more details below). Finally, in formulating $h(x;\mathcal{P})$ we assume that the prototype images are independent and that the global and local information in each prototype can be functionally disentangled into a product form.

Since our goal is to learn the relative importance of global and local features in a prototype set, we solve the following optimization problem:

$$\arg\min_{\beta,\lambda} \mathcal{L}(\beta,\lambda) = \arg\min \sum_{i=1}^{n} \text{CrsEnt}(\sigma(h(x_i)), y_i) + C_\beta||\beta||^2 + C_\lambda|\lambda| \tag{2}$$

using gradient descent. We use cross-entropy loss function (CrsEnt) to penalize misclassifications on the training set $\mathcal{X} = \{x_i\}$ and to obtain $\beta_{\text{optimal}} = \{\beta_k\}$ and $\lambda_{\text{optimal}} = \{\lambda_k^G, \lambda_{kj}^L\}$. We use a $tanh(\sigma)$ activation function on $h(x)$ from Eq. 1. To avoid overfitting, we invoke ℓ_2^2 and ℓ_1 regularization with coefficients C_β and C_λ respectively. Following the intuition that a pathologist might pay no attention to some features, e.g., small-round nuclei do not feature typically in the diagnosis of ADH, we choose ℓ_1 regularization for λ to sparsify the weights.

4.3.2 Encoding global and local descriptions of a duct

The functions $c_k(x)$ and $f_{kj}(x)$ in Eq. 1 compute the global and local differences between input image x and prototype set $\mathcal{P}$, as outlined in the steps below.

Step 1: For a proof-of-concept, we adopt the approach from [80] to build analytical models of 16 diagnostically relevant tissue features (see chapter 3 for more details) following the guidelines presented in the WHO classification of tumors of the breast [54].

Recap of the analytical model of a cribriform feature: Fig. 9 illustrates how to model a histological feature, *cribriform*, that is critical to diagnosing ADH. By considering a spatial neighborhood of $100\mu m$ around each cell in ground-truth annotations of cribriform features in ROIs, the model incorporates three different components: (1) polarization of epithelial cells around lumen inside the ROI; (2) distance of any given nucleus in the ROI to two nearest lumen; and (3) circularity of lumen structure adjacent to a nucleus inside the ROI. For the ROI in Fig. 9A, the analytical models driving these three components are: (1) mixture of Gaussians (MoG) ($\mu_1 = 0.87, \mu_2 = 0.94, \mu_3 = 0.72$, $\sigma_1 = 0.002, \sigma_2 = 0.002, \sigma_3 = 0.003$, $\pi_1 = 0.44, \pi_2 = 0.35, \pi_3 = 0.21$) for modeling the distribution of clustering coefficients [68]; (2) Gamma distribution ($\alpha = 3.11$, $\beta = 34.37$) for modeling distance values to lumen and (3) a uniform distribution ($a = 0.2$, $b = 0.92$) to model the circularity values of nuclei inside the ROI. We further combine these three components with a mixture model, performing grid-search to optimize the mixing coefficients (Fig. 9B), to form the histological feature of cribriform (P_{gt}^{crib}).

We pursue a similar approach to modeling other tissue features using ground-truth ROI annotations: 1. *small*, 2. *large*, 3. *round*, 4. *crowded*, and 5. *spaced*, each modeled as a Gamma distribution; 6. *elliptical*, 7. *large-round*, 8. *small-elliptical*, 9. *spaced-large*, 10. *crowded-small*, 11. *spaced-small*, 12. *crowded-elliptical*, and 13. *spaced-round* each modeled

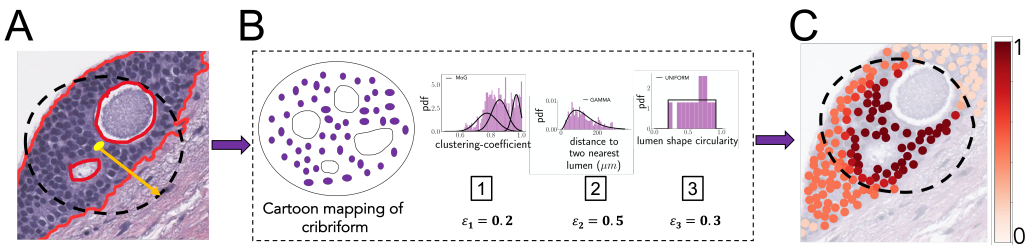

Figure 9: Modeling cribriform pattern in a sample ROI (A) using parametric models for three component patterns in (B) and generating cell-level likelihood scores (C). Ductal region and intra-ductal lumen are outlined in red in (A).

as two-component MoG; and more complex patterns 14. *large-round-spaced*, 15. *picket-fence*, and 16. *cribriform* using a combination of Gamma, MoG, and Uniform distributions. Details on parameter estimation are discussed in 3 [80].

Generating likelihood scores: Next, to compare ground-truth model of any tissue feature P_{gt} with a new model generated from the reference nucleus of an input image (P_{new}), we use two distance measures, 2-sample Kolmogorov-Smirnov test and Kullback-Leibler divergence to compare Gamma and MoG distributions respectively. To map smaller distances that indicate stronger presence of the feature, we compute likelihood scores by applying an inverted S-function on the distances. In Fig. 9C the final likelihood score from evaluating the cribriform feature is a weighted sum of the likelihood scores of the component features. A similar operation is carried out for generating cell-level likelihood scores for the remaining 15 features. The principal advantage of these analytical models is in their ability to handle heterogeneity that emerges from running imprecise low-level image processing routines, such as methods for segmenting nuclei or identifying boundaries of ductal ROIs. The heatmap visualization in Fig. 9C is a mechanism for explaining the model to pathologists, informing where these features are and how strongly they influence the overall diagnosis of a ROI.

Step 2: To encode the global description of a duct, we represent it by a matrix of size $n \times l$ populated with likelihood scores, where n and l refer to the total number of cells and the number of histomorphological patterns respectively ($l = 16$). Additionally, we include the size of the largest duct if the ROI has a cluster of ducts. However, considering only the global information may lead to diagnostic inconsistencies. For example, a duct resembling FEA is better diagnosed as ADH if it contains a local cribriform feature or as a CCC duct if it contains some hyperplasia (further meriting a comparison of local hyperplastic area with models of FEA/ADH).

Step 3: To encode the local description of a duct, we adopt a strategy followed by most expert pathologists. To this extent, for every histomorphological feature, we identify islands within the duct where that particular feature is dominant and consider the largest island for further analysis. We detect feature islands by performing non-maxima suppression on cell-level likelihood scores using a threshold ($= 0.8$) based on cross-validation.

Step 4: Finally, we have the machinery to compute the functions $c_k(x)$ and $f_{kj}(x)$ from

Eq. 1. We define $c_k(x) = \|d(p_k, x)\|$, where a small value of $c_k(x)$ implies high similarity of image x to prototype p_k. We combine two measures to generate d: Kolmogorov-Smirnov test comparing 16-dim probability distributions of cell-level likelihood scores individually between x and p_k and an inverted S-function on the ratio of the duct sizes between x and p_k. This leads to a 17-dim vector d, which is further compressed by its ℓ_2 norm to obtain a single scalar value $c_k(x)$ for every pair of x and p_k. We further simplify the computation of $f_{kj}(x)$ by applying an inverted S-function on the ratio of the largest feature island sizes from the same histological feature between x and p_k, suitably modified to account for islands that are missing in either x or p_k.

4.4 Results

4.4.1 Dataset

We worked on the same dataset described in 3 which consisted of 93 WSIs which were labeled by an expert pathologist on the team to contain at least one ADH ROI. The breast biopsy slides were scanned at $0.5\mu m$/pixel resolution at $20\times$ magnification using the Aperio ScanScope XT (Leica Biosystems) microscope from which 1295 ductal ROI images of size $\approx 1K \times 1K$ pixels were extracted using a duct segmentation algorithm described in [80]. Briefly, the algorithm first breaks down the image into non-overlapping superpixels and then evaluates each superpixel's stain level together with its neighboring superpixels and assigns probabilities of them belonging to a duct. These guesses are then used to perform Chan-Vese region-based active contour segmentation algorithm [57] that separates the foreground (i.e., ducts) from the background.

We collected ground truth annotations of extracted ROIs from 3 breast pathology subspecialists (P1, P2, and P3), who labeled the ROIs with one of the four diagnostic categories: Normal, CCC, FEA, and ADH. The diagnostic concordance for the four categories among P1, P2, and P3 were moderate with a Fleiss' kappa score of ≈ 0.55 [38]. The entire dataset was split into two sets.

i. Prototype set: We formed three prototype sets (PS-1, PS-2, and PS-3) containing ROIs with consensus diagnostic labels from the 3 pathologists having a balanced distribution over the four diagnostic categories. The final set of prototype ROIs were verified by P1 to confirm adequate variability is obtained. The number of aforementioned *islands* are also listed in Table 2.

ii. Train and test set: The training set consists of 754 ROIs labeled by P1 and the test set contains 541 ROIs consensus labeled by P1-P3. The training and test set were separated at WSI level to avoid over-fitting, since ROIs belonging to the same WSI can be correlated histologically. Due to limited number of ROIs belonging to the non-Normal category as seen in Table 2, the ROIs which do not participate in the prototype set were also included in the dataset.

Table 2: Statistics of the atypical breast biopsy ROI dataset

Prototype Set	PS-1	PS-2	PS-3
No. of ROIs	20	20	30
No. of feature islands	84	86	145

Class	NORMAL	CCC	FEA	ADH	Total
Train	420	99	116	119	754
Test	371	105	33	32	541

4.4.2 Model training and evaluation

Our ML model (Eq. 1) is trained to minimize the objective function (Eq. 2) using gradient descent (learning rate $= 1 \times 10^{-4}$ and convergence tolerance $= 1 \times 10^{-3}$). Regularization coefficients C_β and C_λ were initialized to 2. To speed up convergence, we shuffle the training data after each iteration so that successive training examples rarely belong to the same class. Prior to training, the model parameters β and λ were initialized with weights randomly drawn from *LeCun normal* [83]. After each iteration, the parametric values of the objective function ($\mathcal{L}$), error-rate (ϵ), β, and λ are stored. After model convergence, we use β_{optimal} and λ_{optimal} parameters in the mapping function (3) to obtain h_{test}. Fig. 10 illustrates the optimization of our ML framework for a sample classifier. To obtain a diagnostic label from the optimal parameters, we generate prediction probabilities p by first applying a *tanh*

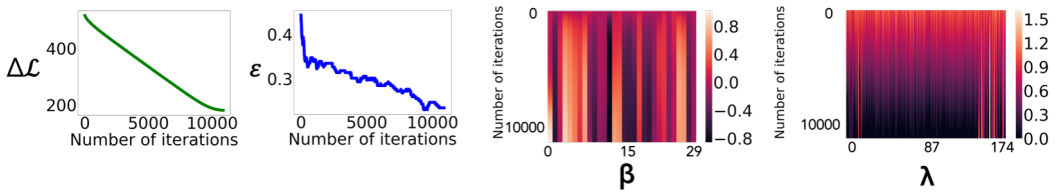

Figure 10: Learning parameters of an ADH-vs-rest classifier with gradient descent. The panel of figures (left to right) shows the values of model parameters: absolute change in the objective function, training error-rate, β, and λ after each iteration for the classifier built using GL3 (global+local) model using the prototype set PS3.

(σ) activation to h_{test} and then projecting it to the positive octant. If $p \geq 0.5$, the diagnostic label is 1 and 0 otherwise.

4.4.3 Baseline models

Following the method laid out in [33], we define two baseline models, B1 and B2, by re-implementing their cell-graph GNNs. We chose GNNs, a recently emerged state-of-the-art technique for encoding spatial organizations, over pixel-based convolutional neural networks (CNNs) as our experiments with CNNs showed poor performances in capturing the spatial context [80]. B1 is obtained by generating a cell-graph topology and cells within each graph are embedded with cytological features as in [33]. To assess the effect of histological features in cell embeddings, we generate B2 by replacing the duct-level cytological features with likelihood scores generated by our method. Finally, B3 is obtained by implementing a logistic regression classifier using the duct-level likelihood scores, following a similar strategy as in [80].

4.4.4 Classification results

For the sake of differential diagnosis of atypical breast biopsies, we implemented several models using global (G), local (L), and both global and local information (GL) from three

prototype sets (PS1-PS3) and compared it with the baseline models (B1-B3) (see Table 3). During the training step of each model, we created a balanced training set by randomly subsampling ROIs from each category so that we have equal number of ROIs for each classification category. To check for statistical significance, for each classification task, we run our ML algorithm on 10 training sets wherein the images are randomly selected and we report the classification scores as the mean and standard deviation over 10 runs (Table 3). The top panel of Table 3 (HR row) compares the classification performance of low-risk (Normal+CCC, −ve class) vs high-risk (FEA+ADH, +ve class) cases. For each diagnostic category (+ve class), we further implemented a different binary classifier for each modeling strategy proposed. The bottom panel of Table 3 (ADH and FEA row) shows the comparative performances of ADH- and FEA-vs-rest diagnostic classification.

Table 3: Diagnostic results from the binary classification task expressed in %

		Baseline			PS-1			PS-2			PS-3		
	Model	B1	B2	B3	G1	L1	GL1	G2	L2	GL2	G3	L3	GL3
HR	**R**	56±6	68±6	62±3	66±4	71±1	73±4	68±4	72±2	68±3	66±7	74±2	69±3
	wF	77±2	82±3	76±1	65±2	61±1	65±1	67±4	61±1	63±2	63±1	64±1	64±1
ADH	**R**	38±8	45±7	56±3	70±7	61±8	78±8	59±13	80±4	71±4	72±6	70±11	68±5
	wF	78±4	86±2	79±1	70±3	64±2	67±5	64±3	62±1	60±6	64±2	67±1	64±1
FEA	**R**	48±12	40±6	35±4	54±6	64±5	68±7	58±6	60±3	63±5	63±6	67±2	62±5
	wF	81±5	82±3	78±1	71±2	65±2	69±3	66±4	66±3	69±3	66±2	69±2	66±3
CCC	**R**	51±7	63±8	60±2	55±8	46±5	53±6	60±5	57±3	68±3	52±6	53±3	54±5
	wF	52±5	63±3	54±1	55±3	50±5	53±5	55±3	55±5	54±3	53±4	51±5	54±3
Normal	**R**	84±4	85±3	78±1	52±29	61±1	66±2	50±14	65±1	60±3	70±26	63±2	61±1
	wF	71±1	78±1	72±1	53±10	64±1	63±2	52±10	68±1	62±1	61±11	66±1	66±2

Performance metrics: For each classification scenario, we use *recall* (R) as the performance metric to focus on the correct detection of positive class, since there is a significant class imbalance (see Table 2) and the consequence of misdiagnosis (false negative) implies

increased chance of developing cancer with lack of providing early treatment. We include *weighted F-measure* (wF) as an additional metric which gives importance to the correct detection of both positive and negative classes [84]. The class specific weights in wF are proportional to the number of positive and negative examples present in the test set.

Classification performance: We highlight the best *recall* performances in Table 3, that are achieved using state-of-the-art baseline models against our method in black and gray boxes, respectively. Our method shows significant improvement ($p < 0.01$) in detecting diagnostically critical high-risk ADH and FEA ROIs compared to the baseline methods (the best average recall achieved is 80% for ADH classifier and 68% for FEA). We also observe that baseline models are performing better on detecting Normal ROIs. This behaviour explains higher weighted F-measure of baseline models in low- vs. high-risk classification, since in the testing set low-risk ROIs are 7-fold more than high-risk ROIs (i.e., baseline models are biased to detect low-risk lesions even when the training set was balanced). It is critical to note that real-life clinical observance of high-risk lesions is also around 15% [54], which is naturally reflected in our testing set, and it is crucial to catch these less-seen high-risk lesions for pre-cancer interventions while being able to provide diagnostic explanations to given recommendations.

4.5 Discussion

The interpretability of our model is depicted in Fig. 11, which shows that our model leverages both global (λ_G) and local (λ_L) information of the ductal ROIs of two prototypical images, I and II, in detecting ADH from one of the experiments using GL3 classifier built using prototype set PS3. Fig. 11-I positively guides in detecting ADH category ($\beta = 0.15$) whereas Fig. 11-II is counterintuitive in detecting ADH lesions ($\beta = -0.47$). Although two of the histological feature islands, large and large-round present within these ROIs overlap, we assert that the absence of complex architectural pattern such as cribriform within Fig. 11-II might have led to a negative influence of this prototype's influence to detect ADH. Although it is possible that a FEA type lesion could be upgraded to ADH pathologically without

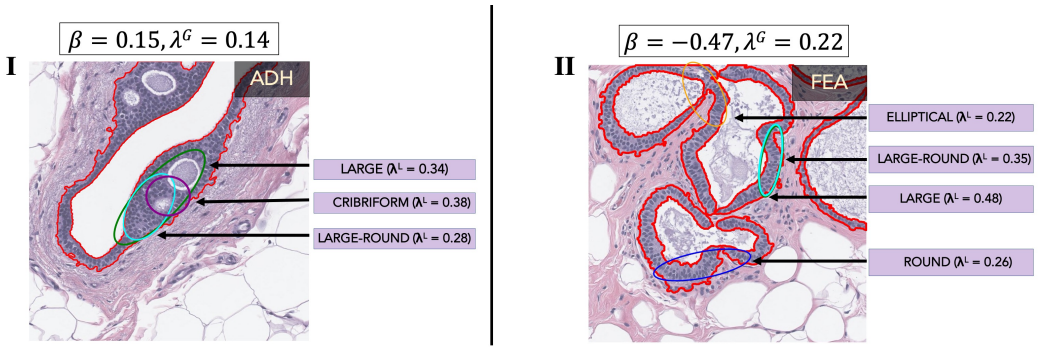

Figure 11: Highlighting the relative importance of the global and local features from different prototypes (I and II) in ADH-vs-rest classifier.

cribriform architecture, this would require thickening of the duct lining to more than 5 cell layers which is uncommon in clinical practice.

Computational Cost: The entire pipeline is implemented in native Python 3.8. Total time required to obtain a diagnostic label with computation of all features for a previously unseen ROI is less than $30s$ on a 64-bit single $3.4GHz$ Intel Xeon processor.

Limitations: (1) Tissue features like duct and lumen morphology, texture properties, etc. are missing; (2) Selection of prototypes was made on the basis of expert visual inspection. There is a need for more sophisticated statistical approaches [85] for prototype selection and (3) for a more detailed ablation study to test the robustness and reliability of our ML framework; (4) To offset the issue of unbalanced datasets, we are collecting expert annotations on additional high-risk lesion biopsies. Most of these limitations are addressed in the next chapter.

5.0 Enhancing the computational pathology framework for the differential diagnoses of a broad spectrum of breast biopsies

5.1 Chapter summary

In chapters 3 and 4, we evaluated our methods on the $1K \times 1K$ breast biopsy ROIs containing one or more ducts. It is well-established that, 90% of all the breast cancer cases originate in the epithelial ducts. Thus, to better understand the morphological abnormalities of the terminal duct lobular units (TDLUs), in this chapter, first, we evaluate our framework on each duct separately and assign duct-level diagnostic labels. Second, we introduce additional tissue features to enrich the breadth of diagnostically relevant feature dictionary. Third, we enhance the prototype-driven computational pathology framework conceptualized in chapter 4 for the challenging task of differentially diagnosing a broad spectrum of breast biopsies. The key components of this framework are: (i) *analytical models* for additional diagnostically relevant tissue features, along with texture-based models; (ii) *automatic class-specific prototype selection* using analytically modeled tissue features; and (iii) improved *prototype-driven machine learning* for differential diagnoses. Fourth, we show significant improvement in the classification performance ($\approx 20\%$) over state-of-the-art methods on two different datasets: high- and low-risk benign breast lesions (HLRBB) (1237 ROIs at $20\times$) and publicly available breast carcinoma subtyping (BRACS) data (4539 ROIs at $40\times$). Finally, our framework provides pathologist-friendly explanations paving the way for better, transparent, and trustworthy diagnostic tools.

5.2 Introduction

5.2.1 Background

Pathologists typically diagnose the breast tissue slides under a microscope by examining: i. lumen and ductal morphology, ii. nuclei size, shape, and spatial arrangement and their combinations, iii. intraductal architecture, and iv. textural properties. We assembled a subset of these tissue features that pathologists frequently use and documented in the standard reference book from WHO on the classification of tumors [54]) in making complex diagnostic decisions as shown in Fig. 12 (textural properties are not depicted). Our goal is to test the inferential power of a prototype-driven computational pathology pipeline based on analytical modeling of these tissue features in differential diagnoses of breast biopsies.

In this chapter, we will consider a broad spectrum of breast biopsies including: (i) invasive breast cancer (IC), (ii) three high-risk benign lesions: ductal carcinoma in-situ (DCIS), atypical ductal hyperplasia (ADH), flat epithelial atypia (FEA), and (iii) three low-risk benign lesions: usual ductal hyperplasia, columnar cell change (CCC) and Normal; where the risk is indicated by the relative chance of developing breast cancer. Infiltrating mammary carcinoma or IC is carcinoma cells infiltrating into the breast stroma and not confined to breast ducts.

DCIS is a carcinoma that is confined to the breast ducts and is not invasive into the stroma. DCIS represents a spectrum of disease ranging from low-grade to high-grade where the cytologic atypia appears malignant. ADH lesion features both hyperplasia (too many cells in a milk duct) and atypia (e.g., cribriform architecture shown in 12); hence considered as pre-cancer lesions. FEA is a lesion that combines the nuclear atypia seen in ADH, but lacks hyperplasia and has simple architecture (no cribriform) sitting between ADH and low-risk benign lesions [86].

CCC is considered to be low-risk, but it shows morphological overlap with FEA ducts. UDH is a benign proliferative breast lesion where there are too many cells within the duct (as can be seen in ADH or DCIS), but the cells are benign appearing and lack atypia. Finally, we have Normal ducts, which are simple non-columnar ducts. Diagnoses of these

transitional benign lesions are problematic and concordance remains elusive in real world diagnosis, reported by one study to be as low as 48% [49]. Negative consequences of this can manifest as over-treatment with surgery and long-term drug therapies, or under-treatment with subsequent cancer-related morbidity or mortality. Even if correctly diagnosed with high-risk lesion, majority of women will retrospectively be considered to have been over-treated; surgical excision to address a 4% current cancer risk means that 96% of patients who undergo these surgeries do not turn out to have cancer after all.

5.2.2 Related Work

Prototype-driven recognition: There have been numerous studies that have discussed the merits of using prototype-driven approaches for image classification tasks [71,77,79,85,87,88]. In [88], the authors used human face image prototypes and built feature representations from comparing the input faces to the prototype sets. Similarly, in [77], the authors discussed the benefits of using prototype-based methods on indoor-scene categorization (67 indoor scenes) which suffers from large intra-class variability, a relatable issue in the domain of histopathology. More recently, a deep learning architecture called ProtoPNet was developed which identifies similar looking parts within an input image to the prototype set for bird species classification and car model identification [78, 79]. ProtoPNet has achieved comparable results to the previous state-of-the-art methods such as MA-CNN [89], B-CNN [90], and RA-CNN [91] on the same dataset. A comprehensive study has been conducted in [85] which discusses the importance of choosing "good" prototype set and its effect on classification performance. The principal advantage of using prototype-driven methods is the ease of visual interpretability to the classification outcome which is not provided by a majority of *black-box* deep learning models. Further, this form of interpretability is crucial in applications such as clinical diagnosis, wherein the pathologists are required to *trust* the AI system prior to launching them in a clinical setting. To the best of our knowledge, the application of prototype-driven visual recognition method for histology image classification is the first of its kind and in this paper, we show that it holds merit for other medical image classification tasks which demand diagnostic explanations.

5.2.3 Contributions

We demonstrate the inferential power of our prototype-driven computational pathology pipeline based on analytical modeling of tissue features for differential diagnoses of breast biopsies. The key components of this framework are: (i) *analytical models* for a subset of diagnostically relevant tissue features, along with texture-based models (Fig. 12), that the pathologists frequently use and documented in the standard reference book from WHO on the classification of tumors [54]; (ii) *automatic class-specific prototype selection* using analytically modeled tissue features; and (iii) *prototype-driven ML* for differential diagnoses. We show significant improvement ($\sim 20\%$) over the state-of-the-art methods [1] in the classification performance on two different datasets: high- and low-risk benign breast lesions (HLRBB) (1237 ROIs at 20×) and publicly available breast carcinoma subtyping (BRACS) data (4539 ROIs at 40×). Our framework provides pathologist-friendly explanations paving the way for better, transparent, and trustworthy diagnostic tools.

5.3 Methods

5.3.1 Enhanced machine learning framework

To aid pathologists' routine diagnostic workflow, we present a computational pathology-based diagnostic framework which reflects a pathologist's cognitive process and provides explanations to the classifier outcome. A key element of our framework is to learn the relative importance of lumen/ductal morphology (LD), intraductal tissue features (ID) and textural features (T) (see Fig. 12) from a set of prototypical images to obtain a diagnostic label. In doing so, we assert that the assignment of relative importances to LD, ID, and T features is driven by similar looking ducts (*prototypes*) which were previously encountered during pathology training or clinical practice. To achieve this in the context of differential diagnosis, we introduce a mathematical formulation to learn the contributing power of LD, ID, and T features within a prototype set $\mathcal{P}$ in association with an input image x using a mapping function $m : x \rightarrow \mathbb{R}$ defined as,

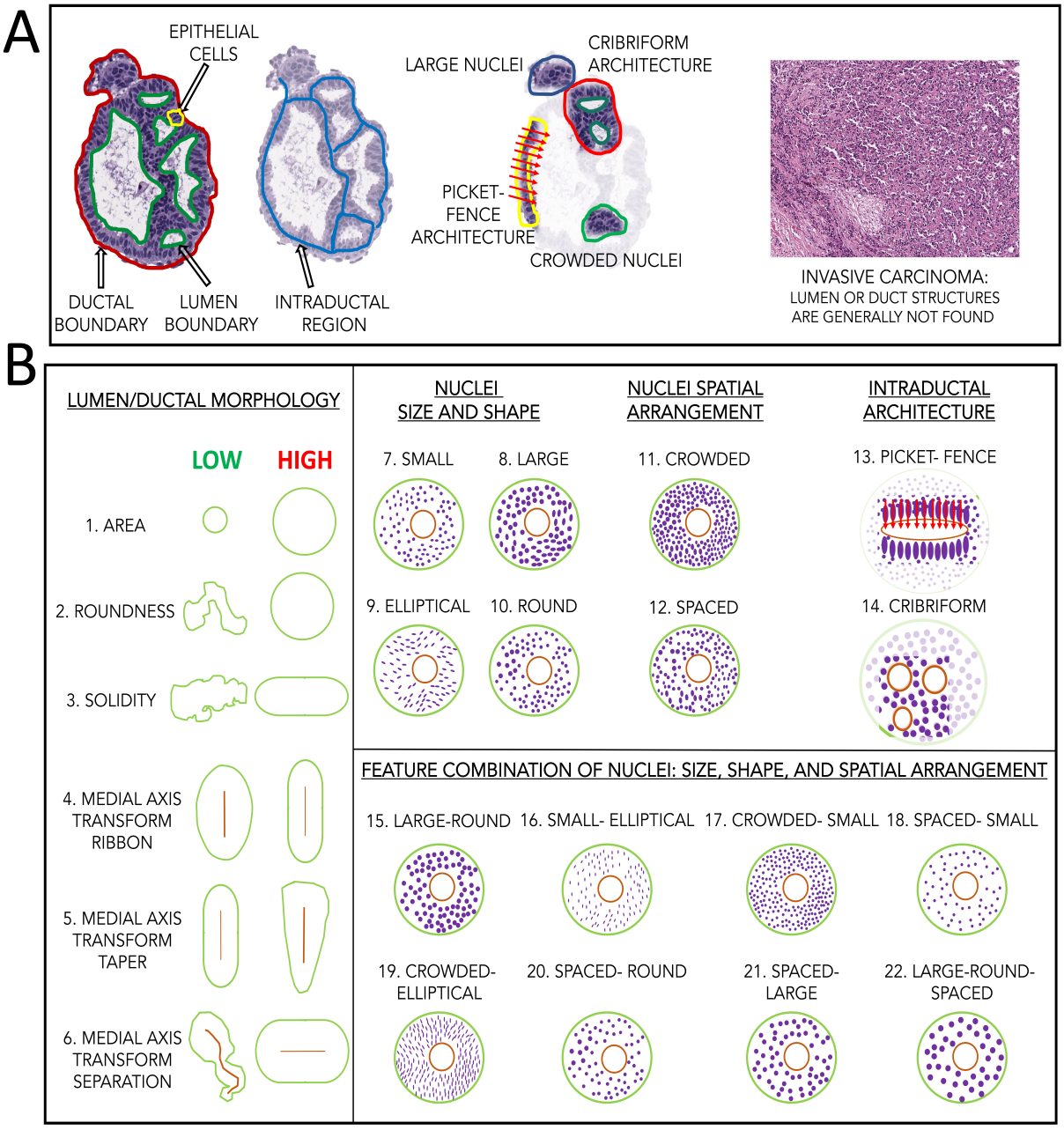

Figure 12: (A) Tissue features outlined: duct (red), lumen (green), epithelial cells (yellow), and intraductal regions (blue). (B) A schematic representation of the tissue features analytically modeled in this study. In the absence of duct and lumen structures (e.g., invasive carcinoma), we invoke texture-based models.

$$m(x;\mathcal{P}) = \sum_{k=1}^{p} \beta_k \exp^{-\sum_{l=1}^{h} \lambda_k^l f_k^l(x)}. \tag{3}$$

The mapping function m is a result of combining the associations of x to all prototype images $\mathcal{P}$ (with k = 1 ... p) by analytically comparing the LD, ID, and T information for a total of h features in each of the k prototypes to obtain information differences $f_k^l(x)$. β (size $p \times 1$) and λ (size $p \times h$) are the tuning parameters of our ML model. β_k indicates if the feature similarity of prototype p to x positively guides (β_+) in predicting a diagnostic label or otherwise (β_-). λ_k^l highlights the importance of l^{th} feature present within k^{th} prototype image. During the process of estimating the model parameters, we enforce non-negativity constraints on λ to reflect the distance measurements of function $f_k^l(x)$ between input x and prototype k for feature l (see 5.3.5 for more details). Further, if a particular feature is absent within the prototype image, we assign $\lambda = 0$ and remove it from the optimization step.

To learn β_{optimal} and λ_{optimal}, we use the association function m to define our diagnostic classification objective as,

$$\begin{aligned} \arg\min_{\beta,\lambda} \mathcal{L}(\beta,\lambda) = \arg\min \sum_{i=1}^{n} \text{CE}(\sigma(m(x_i)), y_i) \\ + C_\beta ||\beta||^2 + C_\lambda |\lambda| \end{aligned} \tag{4}$$

and solve it using gradient descent based optimization strategy. To penalize training misclassifications, we use cross-entropy loss function (CE) and $||\beta||^2$, $|\lambda|$ as model regularizers to account for overfitting with regularization constants C_β and C_λ. Further, we use $\tanh(\sigma)$ activation function on m. The ℓ_1 regularization used for λ parameter sparsifies the feature importances resonating with the pathologist's judgement of paying minimal attention to some patterns over others.

The gradient of objective function $\mathcal{L}$ with respect to β, λ are computed as:

$$\frac{\partial L}{\partial \beta_k} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma\left(m(x_i)\right) - y_i}{\sigma\left(m(x_i)\right)\left(1-\sigma\left(m(x_i)\right)\right)} \times$$
$$\left[1-\tanh^2(m(x_i))\right] \times \exp^{-\sum_{l=1}^{h} \lambda_k^l f_k^l(x_i)} + 2C_\beta \beta_k, \tag{5}$$
$$\frac{\partial L}{\partial \lambda_k^l} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\sigma\left(m(x_i)\right) - y_i}{\sigma\left(m(x_i)\right)\left(1-\sigma\left(m(x_i)\right)\right)} \times$$
$$\left[1-\tanh^2(m(x_i))\right] \times \beta_k f_k^l(x_i) \exp^{-\sum_{l=1}^{h} \lambda_k^l f_k^l(x_i)} + C_\lambda \frac{\lambda}{|\lambda|}. \tag{6}$$

5.3.2 Encoding lumen and ductal morphology (LD)

To capture the lumen and ductal morphology, we estimate i) LD-MORPH based on ductal morphological properties such as area, roundness, and solidity computed using ImageJ software from binary duct masks and ii) LD-MAT by characterizing the shape of the ductal cellular region or DCR (the region between duct boundary and lumen) (see Fig. 12A) based on the derived scores from the medial axis transform (MAT) (see Fig. 12B).

The MAT-based skeletonization along with its derived scores have been shown to have strong ties to the human visual system and how it processes and categorizes shapes [92–94]. This is likely due to the low-dimensionality of the skeleton, as well as its stability across small perturbations [95]. The MAT scores also facilitate scene classification in contour images (when only the boundaries of shapes in the image are present) [96]. We hypothesize that this property will allow the MAT algorithm to closely mimic the process of a pathologist analyzing the shapes of the histological structures found in the TDLU. To describe LD-MAT, first, we construct skeletons of the binary masks of DCR using the *skeletonize* plugin in ImageJ software [55]. Next, following the methods presented in [96], to make the binary skeletons amenable for characterizing structures, we derive three scores for each pixel on the skeleton: ribbon, taper, and separation, where each score is useful in recognizing different local symmetry patterns [97].

To describe each score, consider s to be a point on the skeleton and $R(s)$ to be the radius function which maps s to a value which is the radius of the maximally inscribed

disk centered at s. Thus, $R(s)$ is the distance from skeleton point s to its corresponding, equidistant points on the shape boundary. All scores are computed over a window of size $k = 16$ skeleton points, centered at the point s for which we want to compute the scores.

1. Ribbon score: Captures the degree of parallelism of the surrounding contours using the gradient of $R(s)$. Ribbon score is high when the gradient is low, that is, when the surrounding contours are close to parallel. Thus, we can expect a high ribbon score for DCRs of elongated ducts with no hyperplasia, as the DCR is smooth and mostly parallel (see Fig. 12B). The ribbon score is modeled using the equation, $R'(s)/\max_{s_i \in [1,k]} R'(s_i)$.

2. Taper score: Captures the rate of change of the gradient of $R(s)$. Taper score is high when $R(s)$ changes at a constant rate, such as in the shape of a funnel or railroad tracks stretching towards the horizon. Taper score will be moderate in examples with parallel contours, and high in regions where contours approach each other to a tip (see Fig. 12B). We can expect a high taper score for structures which are more circular or oblong, such as in normal ducts. The equation used to quantify taper score is $R''(s)/\max_{s_i \in [1,k]} R''(s_i)$.

3. Separation score: Captures the degree of separation between the contours, and increases with distance. Separation score will be high for ducts which are large (typically observed in ADH and DCIS ducts where cells crowd the lumen), and low for smaller ducts (Normal). To quantify separation score, we use the equation $1 - ((2 * k - 1))/\text{trapz}(S))$, where trapz() is the trapezoidal numerical integral with unit spacing and S is the set of radius values for the window centered at s.

For each duct, we obtain a matrix LD-MORPH (size 1×3) populated with three morphology properties and LD-MAT (size $s_{dcr} \times 3$) populated with MAT scores ($\in [0, 1]$) for the DCR skeletons containing s_{dcr} skeleton pixels respectively. Note, all three MAT derived scores are rotation and scale invariant. Thus, we capture the complete lumen and ductal information using 6 features: area, roundness, solidity, MAT-ribbon, MAT-taper, and MAT-separation scores (Fig. 12B).

5.3.3 Encoding intraductal tissue features (ID)

To model the intraductal information based on nuclei size, shape, spatial arrangement and architecture, we build analytical models for sixteen tissue features (Fig. 12B) upon consulting the expert pathologists on our team and studying standard pathologist guiding references such as the WHO classification of tumors of the breast [54]. The idea of modeling the features was first conceptualized in [80]. A detailed description of the strategy used to model the tissue features is provided in chapter 3. The tissue features are organized based on the nuclei size and shape (*small*, *large*, *round*, *elliptical*, *large-round*, *small-elliptical*), and spatial arrangement of the cells (*crowded*, *spaced*). These are used to identify additional features such as *spaced-large*, *crowded-small*, *spaced-small*, *crowded-elliptical*, *spaced-round*, and *large-round-spaced*. Further, the orientation of the cells and shape of the lumen is computed to identify higher-order features such as *picket-fence* and *cribriform*. The frequency of occurrence of these tissue features is different for each diagnostic category which is a major contribution to discordance among pathologists [49]. We build a quantitative model and tune the parameters of this model to *match* with the consensus diagnosis of *template* features.

Quantifying ID: The measurement of univariate features such as small and large is done using cell areas, while round and elliptical are quantified using shape statistics such as roundness $= 4\pi \times$ area/perimeter^2 and ellipticity $=$ length$_{\mathrm{minor-axis}}$/length$_{\mathrm{major-axis}}$. To quantitate spatial spread based on crowding, we modeled the distance of each cell to 20 nearby cells. To capture uniform dispersion of cells or *spaced* feature, we evaluated the population density of cells within a 5×5 grid under *complete spatial randomness hypothesis* [67]. The bivariate features are modeled using the joint distribution of z-scores from two-component mixture of Gaussians (MoGs) of the two univariate features and similarly the trivariate feature (large-round-spaced) uses three-component MoG distribution. Further, higher-order features such as picket-fence and cribriform are identified from a group of simple univariate features. The picket-fence feature is characterized by the organization of i) crowded, ii) elliptical cells which are iii) oriented at 90° to the lumen. Finally, the cribriform feature is recognized from the i) polarization of cells around ii) multiple lumen which are iii) mostly circular.

Templates for feature-matching: For each of the 16 tissue features, we acquired a template stack from 5 to 10 image regions (gold-standard) strongly indicating the histological feature under consideration from an expert pathologist on our team.

Recap of building analytical models to obtain likelihood scores: We constructed probability distributions upon evaluating each feature present within its associated template. For example, to identify *large* feature within a duct, an analytical model is obtained from the statistical distributions of the size of cells within "large template" L_t. To identify *large* feature in a new input image, the model parameters of L_t is compared (using Kolmogorov-Smirnov (K–S) test for unimodal distributions and Kullback-Leibler divergence for MoGs) with the parameters of a new distribution generated from a training image (L_n) to predict the strength of the feature being present/absent. To account for the amorphous nature of these features, L_t is compared with distributions arising from each cell and looking at its neighborhood spanning $100\mu m$ to obtain L_{n-cell}. Finally, we transform the distance measurement into likelihood score by feeding the distances through an inverted S-function, thereby capturing the intuition that small distances reflect higher similarity to template model and hence a stronger presence of the feature. We follow the steps for all 16 intraductal features (ID).

To capture the complete intraductal information, for each duct with n cells we obtain a cell-level likelihood matrix (size $n \times 16$).

5.3.4 Encoding textural features (T)

On the original 3-channel ductal ROI (RGB image), we measure the following popular textural properties (T-Pop) including mean, variance, skewness, kurtosis, 5^{th} to 11^{th} central moments [98, 99], *local binary pattern (LBP)* [100], *gray-level co-occurrence matrix (GLCM)* [99,101] and *Gabor filters* [102] and also complex wavelet-derived properties (T-Wv) *auto-cor-real, auto-cor-mag, color-cor, cousin-mag-cor, cousin-real-cor, mag-means, parent-mag-cor, parent-real-cor, pixel-stats, pixel-lp-stats, variance-hpr* [103, 104]. T-Wv features comprise of coefficients from complex wavelet transform basis functions at adjacent spatial locations, orientations and scales. The organization of textural properties is shown in Ta-

ble 4. For representing invasive carcinoma, we extract the hematoxylin channel and process the entire image to measure textural properties.

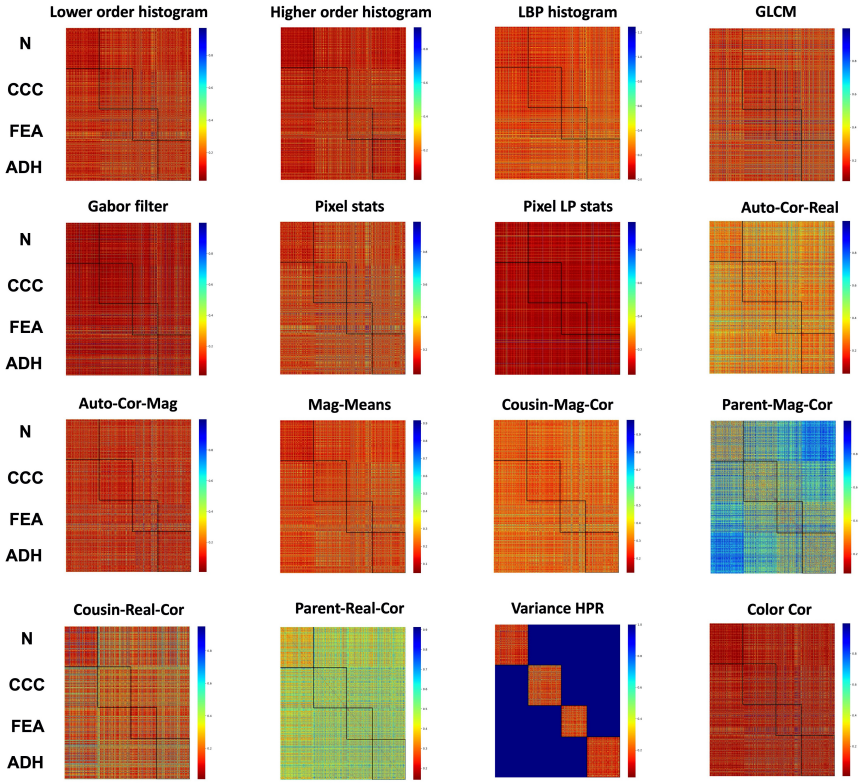

Figure 13: Visualization of the textural information differences across four diagnostic classes from HLRBB dataset. We can observe that some textural features (e.g., variance-HPR, parent-mag-cor, etc.) show less inter-class similarities and hence we analyze the performance of our ML framework on a subset of such features through manual selection

5.3.5 Computing information differences $f_k(x)$

The difference function $f_k(x)$ between input image x and a prototype image p_k (with $k = 1 \ldots p$) is computed by comparing the lumen/ductal morphology, tissue features and textural properties. We perform K-S test to compare LD-MAT and the intraductal tissue features. LD-MORPH features (area, roundness, and solidity of duct) and the textural

properties are compared by measuring the Euclidean distance between each property in the image x to p_k and is transformed by an inverted S-function such that $0 \leq f_k(x) \leq 1$. Lower distances indicate a stronger resemblance of the feature f.

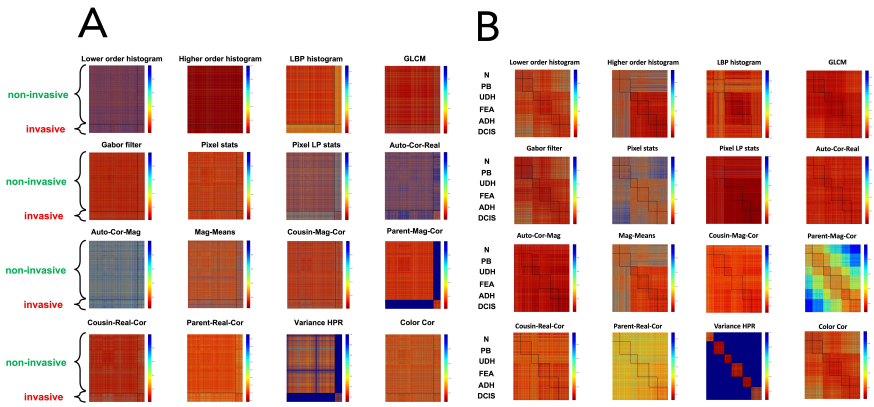

Figure 14: Visualization of the textural information differences across seven diagnostic classes from BRACS dataset. Panel A. shows the heatmap-based visualization for the invasive and non-invasive classes. Panel B. is a visualization of the textural feature differences among the non-invasive or benign breast lesion diagnoses. We can observe that some textural features (e.g., variance-HPR, parent-mag-cor, mag-means, etc.) show less inter-class similarities which is used in selecting the feature subset for assessing the ML framework's performance.

5.3.6 Implementation details

Extraction step of LD, ID, and T features is fully automated. The training examples are randomly shuffled after each iteration [83]. For training, the objective function $\mathcal{L}$ described in section 5.3.1 is minimized using gradient-descent and the function typically converges within 10^4 epochs (convergence tolerance $= 1e^{-3}$) to obtain $\beta_{optimal}$ and $\lambda_{optimal}$. These optimal parametric values are used in mapping function m (3) to obtain m_{test} and further generate prediction probability $\vec{p}$ through $tanh$ activation ($\in [-1, 1]$) and scale the values to fall between $0 - 1$, to obtain predicted label $l = 0$ if $p < 0.5$ and $l = 1$ if $p \geq 0.5$. A learning rate of $\alpha = 1e^{-4}$ is selected. The β and λ parameters are initialized with weights

drawn from *LeCun normal* with zero mean and standard deviation $= 1/ps$, where ps refers to the size of the prototype set [83]. For choosing regularization co-efficients $C_\beta = 2$ and $C_\lambda = 2$, we monitored the model parameters such as change in the objective function ($\Delta\mathcal{L}$) and error-rate (ϵ) with co-efficient values ranging from $10^{-2} - 10^2$. The entire pipeline was implemented in native Python 3.8 on a 64-bit single 3.4 GHz CPU. Computing $f_k(x)$ along with training the model with $\sim$ 1000 images for 10^4 epochs takes $\approx$ 4 hours. It takes $<$ 1 min. to obtain a diagnostic label for a new image.

Table 4: The sixteen textural features are made up of different number of textural properties which is listed above.

a. lower-order-histogram	5	b. higher-order-histogram	10	c. lbp-histogram	10
d. glcm	20	e. gabor	6	f. pixel-stats	6
g. pixel-lp-stats	24	h. auto-cor-real	736	i. auto-cor-mag	1764
j. mag-means	42	k. cousin-mag-cor	432	l. parent-mag-cor	288
m. cousin-real-cor	2304	n. parent-real-cor	864	o. variance-hpr	3
p. color-cor	9				

5.4 Experiments and results

5.4.1 Dataset and evaluation metrics

HLRBB dataset (20×): We evaluated our proposed method on the high- and low-risk benign breast biopsy dataset. It consists of 93 biopsy whole slide images (WSIs) scanned at 20× resolution using an Aperio ScanScope XT scanner (Leica Biosystems). To generate regions of interest (ROI) from WSIs, we applied the breast duct segmentation algorithm proposed in [80] which breaks the WSI into non-overlapping superpixels, assigns each superpixel a probability of belonging to a duct structure to create a probability map, extracts an estimate of a boundary for the ducts by applying a contour algorithm to the probability map, and finally uses the estimate of the boundary to generate a refined boundary for the duct.

Using this method on our dataset, we segmented a total of 1237 breast ducts from the WSIs, and further divided them into two sets: 1) 199 duct images, for which consensus annotation is collected from 3 expert pathologists and used for determining the ductal prototypes, 2) 1038 duct images, for which annotation is collected from 1 expert pathologist and used for training, validation, and testing of the proposed ML algorithm and baseline methods for comparison. Duct annotations belonged to either of the 4 diagnostic categories: Normal, CCC, FEA, or ADH. Details of the dataset proportions among classes are given in Table 5.

Table 5: Statistics of i) prototype set (HLRBB-PS-T30) for different diagnoses and feature configurations; ii) high- and low-risk benign breast lesion (HLRBB) dataset for five-fold cross validation.

	T-CRC	T-Wv	T-All	LD-MOR	LD-MAT	ID	LD-ID	T-LD-ID	Tr+Val+Te
Normal	5	5	5	5	4	5	5	5	275
CCC	5	5	5	5	5	4	6	5	275
FEA	5	5	5	2	2	2	5	5	213
ADH	5	5	5	5	5	5	5	5	275
Total	20	20	20	17	16	16	21	20	1038

To obtain lumen contours, the RGB duct image is first transformed to the HSV color space and then hierarchically clustered to find groupings of pixels representing the intraductal lumen [105]. Several morphological post-processing operations such as dilation and erosion were applied to obtain the final lumen mask. For nuclear segmentation, we used hematoxylin intensity thresholding followed by watershed and morphological operations as proposed in [38].

Selection of prototype set: To reflect the large intra-class heterogeneity present within the breast biopsies, we formulated the prototype selection as a four-step process. In step 1, we selected C ($= 199$) ducts with concordant diagnostic labels from three breast pathology experts. In step 2, we generated prototype similarity matrix S by comparing the feature set F (LD-MORPH, LD-MAT, ID, T, LD+ID, or LD+ID+T) using two-sample KS-test statistic and Euclidean distance where appropriate. In step 3, we performed t-distributed stochastic

neighbor embedding (t-SNE) on matrix S followed by k-means clustering and derived the optimal number of clusters using *elbow* method. Finally, in step 4, we obtained prototypical ducts closest to the cluster centers and were visually verified to display large heterogeneity. The ducts C and features F were selected based on the classification task (e.g. Normal-vs Rest, CCC-vs-Rest, etc.) and ML configuration (e.g. LD-MORPH, ID, LD+ID, etc.). For example, to select prototypes for evaluating Normal-vs-rest classifier using texture (T) features, we chose C ducts belonging to Normal class and F representing the 16 T features. To understand the effect of choice of prototypes on the classification performance, we tested our ML strategy on two prototype sets: HLRBB-PS-T30 and HLRBB-PS-T3 obtained by changing t-SNE *perplexity* hyperparameter from 30 to 3 respectively [84].

Train, validation, and test set: We implemented our models on 1038 breast ducts split in the ratio 3:1:1 for train (Tr), validation (Val), and test (Te) to run the baseline models. Data is separated at biopsy level to overcome the effect of correlated ducts. We combined Tr and Val to form the training set to run our ML framework. To generate confidence scores for evaluation, we considered five equally sized sets of Tr, Val, and Te splits selected randomly from the original HLRBB dataset.

BRACS dataset (40×): We additionally evaluated our method on the breast carcinoma subtyping (BRACS) dataset [106,107] consisting of 4539 ROIs extracted from 387 WSIs from a broad spectrum of breast biopsies including: Normal, Pathological Benign (equivalent to CCC in HLRBB dataset and denoted as P-Benign in 7), UDH, FEA, ADH, DCIS, and Invasive (5.2) scanned at 40× resolution using an Aperio AT2 scanner at 0.25 μm/pixel. BRACS dataset consists of 3163 training, 602 validation and 626 test images [1]. Nuclei segmentation was performed using HoVer-Net [108, 109]. Ducts were manually segmented. For the selection of prototype set, we operated on the validation dataset (C= 602) and followed a similar strategy as described in the previous section to obtain prototypical ducts (BRACS-PS-T30). To compute intraductal feature ID likelihood scores on the 40× BRACS data, we downsampled the measurements to be appropriate for our analytical models developed at 20×.

Evaluation metrics: We evaluated the performance of all models using recall (R) to focus on correct detection of positive class and weighted F-scores (wF) to indicate the correct

detection of both positive and negative classes. The class weights in wF are proportional to number of images present within each class. The final scores reported in 6 are calculated as the mean and standard deviation of the classification performance over five runs. For BRACS dataset (see 7), we trained the model three times with random parameter initializations as followed in [1].

5.4.2 Classification performances

We evaluated and compared all classification methods by implementing a one-vs-rest classifier for four models: *ADH*-, *FEA*-, *CCC*-, and *Normal-vs-rest* using 5-fold cross validation on prototype sets PS-T30 and PS-T3. The results from HLRBB-PS-T30 and HLRBB-PS-T3 are highlighted in 6. To understand the contributions of LD-MORPH, LD-MAT, ID, and T information separately and collectively (LD-ID and T-LD-ID) in diagnosing breast lesions, we modify the mapping function in 3 accordingly.

Table 6: Mean and standard deviation of recall (R) and weighted F-scores (wF) (in %) from diagnostic classification of benign breast lesions (20× images) in one-vs-rest binary classification tasks of the test dataset from 5-fold cross-validation using HLRBB-PS-T30 prototype set. Best results from baseline methods and our prototype-based methods (top two) are highlighted in bold.

Model	Metric	BG1	BL	BG2	T-Pop	T-Wv	T-All	LD-MORPH	LD-MAT	ID	LD-ID	T-LD-ID
ADH	**R**	25±4	19±1	**49±9**	50±7	65±8	69±5	71±8	51±9	70±4	**78±3**	**71±5**
	wF	69±2	70±1	**75±1**	64±4	67±3	66±2	56±2	61±1	68±4	**68±3**	**68±4**
FEA	**R**	19±7	13±6	**55±7**	52±5	63±5	66±5	48±9	51±5	73±9	**71±6**	**73±2**
	wF	73±2	75±2	**79±3**	67±2	69±2	70±3	71±3	62±5	69±2	**73±3**	**73±3**
CCC	**R**	35±4	28±4	**52±14**	66±3	66±3	70±3	71±4	67±7	78±2	**77±5**	**77±4**
	wF	71±2	74±1	**77±3**	59±4	66±4	67±3	58±3	62±3	75±2	**74±3**	**72±1**
Normal	**R**	56±6	44±7	**63±10**	83±2	80±6	**83±1**	71±5	64±6	76±7	80±8	**80±6**
	wF	74±1	79±2	**82±3**	69±5	76±2	**79±2**	60±5	58±2	76±3	79±2	**81±2**

Results on baseline methods: To benchmark our analytically modeled, prototype-driven approach, we compare and analyze our results with two recently proposed state-of-the-art methods on breast histopathology classification.

1. *Baseline with cell-graph GNNs (BG1-BG2)* [33]: We implemented cell-graph GNNs which was recently explored for tissue analysis [33–35, 37]. BG1 is obtained by generating a cell-graph topology by interlinking cells within a spatial neighborhood of $12.5\mu m$ for generating the graphs similar to the work presented in [33]. Each cell within the graph is embedded with morphological features such as *area*, *perimeter*, lengths of *major*- and *minor-axis*, *orientation*, *circularity*, *aspect-ratio*, and *solidity*.

2. *Baseline with likelihood scores (BL)* [80]:

To generate BL, we used mean of cell-level likelihood scores of the 16 ID features to represent each duct and implemented a logistic regression (LR) classifier with a 5-fold stratified cross-validation (hyperparameter optimized using GridSearchCV based on recall (R), and wF scores).

3. To combine the sophisticated classifier approach of BG1 with more meaningful features from BL, we additionally test the performance of GNN by replacing the cellular embeddings with diagnostically meaningful likelihood scores extracted in 5.3.3 to obtain BG2. As shown in 6, replacing the morphological features with our analytical models, there is a significant improvement in both recall and wF across all four classification scenarios.

Results on HLRBB dataset: 6 shows that our prototype-based method outperforms all the baseline methods by a significant margin (p <0.001) in recognizing the individual diagnostic categories (ADH, FEA, CCC, and normal) (greater R and wF scores). It should be noted that, for each one-vs-rest classification trial, the test set is class-imbalanced by a ratio of 1:3 (+:-ve class) (see 5.4.1). A higher wF and low R scores of baseline methods shown in 6 indicates that compared to our method, it is more successful in classifying ducts based on the majority class instead of correctly recognizing each diagnostic category. We observe that the classification performance improves with the successive inclusion of texture, ductal/lumen morphology, intraductal information, LD+ID, and finally T+LD+ID which is discussed in *ablation studies* (5.5.2).

Results on BRACS dataset: The classification performance on the seven diagnostic

categories from BRACS dataset is highlighted in 7. For comparison, we have included results from the seven-way classification performances of HACT-Net and aggregated statistics from three pathologists as reported in [1]. Interestingly, a variant of the textural feature (T-Wv) shows comparable performance to the baseline method (HACT-Net). For detecting diagnostically difficult benign cases, we exclude IC images. Comparatively, our method outperforms the baseline methods. Further, we also outperform domain experts on BRACS (see *Path* column in 7) by a significant margin.

Table 7: Mean and standard deviation of recall (R) and weighted F-scores (wF) (in %) from diagnostic classification of breast lesions from the BRACS dataset (40× images) using BRACS-PS-T30 prototype set. The classification performance of the pathologists (Path column) and seven-way HACT-Net results are reported in [1]. Best results from baseline methods and our prototype-based methods (top two) are highlighted in bold. The *invasive* classification task using our method is based on texture features only (see 14). *An additional result for detecting invasive cases using a binary-classification setting as stated in [1] is shown.

Model	Metric	Path	HACT-Net	BG1	BG2	T-Pop	T-Wv	T-All	LD-MORPH	LD-MAT	ID	LD-ID	T-LD-ID
Invasive	**R**	-	**82**	77±5	-	60±1	**82±1**	**71±1**	-	-	-	-	-
(I)	**wF**	93±2	**88**|96*	83±1	-	79±1	**86±1**	**85±1**	-	-	-	-	-
DCIS	**R**	-	65	61±3	**65±3**	71±2	51±2	64±2	**76±2**	61±4	71±5	**74±1**	66±2
(D)	**wF**	68±3	66±3	80±2	**77±2**	75±1	75±0	80±1	**79±1**	56±2	71±1	**81±1**	84±1
ADH	**R**	-	38	53±2	**55±6**	72±2	**71±3**	58±2	62±5	66±1	66±2	63±2	**75±1**
(A)	**wF**	46±20	40±3	75±2	**74±2**	61±1	**64±1**	65±1	65±2	62±2	71±1	74±1	**71±1**
FEA	**R**	-	**82**	71±2	82±1	**76±1**	67±2	66±1	57±1	77±8	62±1	72±1	**79±3**
(F)	**wF**	38±3	**74±1**	65±3	64±1	**75±1**	75±1	75±2	63±1	64±1	67±1	71±2	**76±1**
UDH	**R**	-	44	**61±4**	37±13	80±3	86±1	**75±2**	69±2	64±1	57±4	56±3	**73±3**
(U)	**wF**	39±10	44±2	**65±5**	71±2	59±1	57±1	**63±1**	48±2	65±0	68±1	68±1	**68±3**
P-Benign	**R**	-	48	**80±3**	42±8	**80±1**	74±1	69±3	54±2	75±3	61±4	57±2	**73±2**
(PB)	**wF**	52±2	48±3	**75±3**	69±7	**78±0**	75±1	74±1	56±2	51±1	56±2	57±2	**75±1**
Normal	**R**	-	62	**77±5**	73±4	65±3	64±1	62±1	**77±3**	29±6	36±6	37±1	**64±2**
(N)	**wF**	52±12	62±2	**78±1**	74±2	78±0	77±1	79±1	**76±1**	73±1	74±2	82±1	**83±1**

5.5 Discussion

5.5.1 Diagnostic explainability

For illustration, consider ADH-vs-rest classification scenario using LD and ID features on HLRBB dataset. Texture features are excluded here since they do not increase the classification performance (6). The prototype selection step results in five prototypical ducts P1-P5, and the ML framework assigns optimal β values: -1.2, 1.1, 1.5, -1.5, and 1.2 respectively (15).

Test duct D1 (true label: ADH) consists of monomorphically round and spaced nuclei (15) which partially resembles prototype P2 that positively guides in diagnosing ADH ($\beta = 1.1$). Further, D1 shows presence of multiple imperfectly circular lumen indicating a cribriform feature which is detected to resemble P3 and further contributes towards detecting ADH ($\beta = 1.5$). Interestingly, D1 shows similarity to P5 by paying attention to the ductal morphology (solidity) and indeed uniformly large-round (high solidity) structure describes some of the high-risk breast ducts. Overall, our ML achieves a high ADH prediction probability of $\sigma(m(x)) = 0.89$ (true positive) for D1.

Test duct D2 (true label: non-ADH) consists of simple picket-fence architecture with sparsely distributed spaced cells (15). The minor contributions of ductal similarity to prototypes P2 and P5 with respect to spaced and solidity is not sufficient in diagnosing ADH, thereby obtaining a low prediction probability $\sigma(m(x)) = 0.14$ (true negative) for D2.

Impact of lumen/ductal morphology (LD), and intraductal tissue features (ID):

We observe from Table 6, the successful inclusion of both LD and ID features shows an improvement in the classification performance. To further investigate this, we analyzed the prediction probability distributions from each of our ML configurations on one testing set in HLRBB dataset from the 5-fold cross-validation setting. Interestingly, the prediction probabilities from LD features are $\sim$ 60% for true positives (TPs) and $\sim$ 30% for true negatives (TNs). Combining this with ID features (LD+ID), the classification performance is vastly improved (Fig. 16) and is further shown to diagnose with higher confidence (TPs and TNs are pushed farther away from the mean value of 0.50).

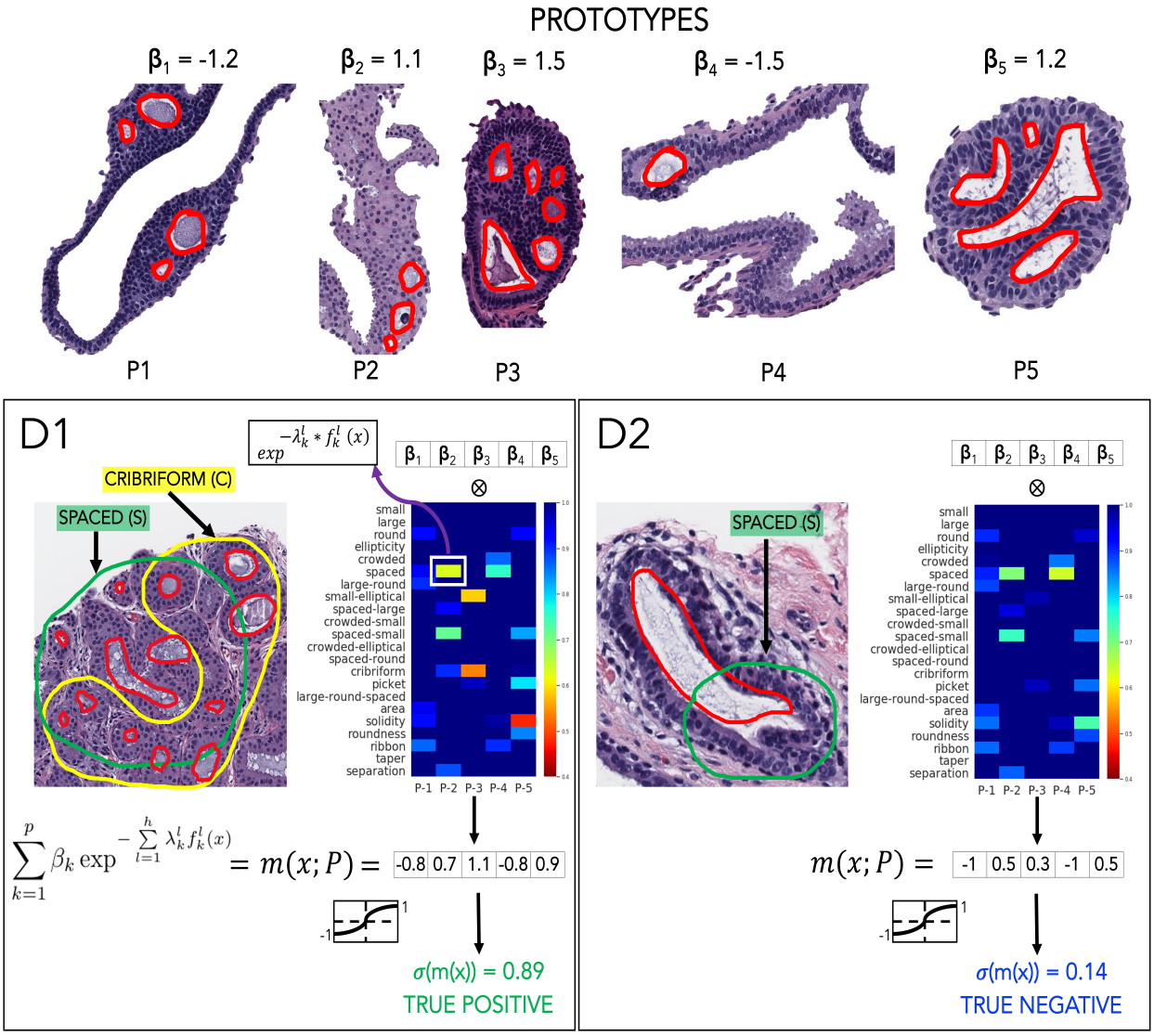

Figure 15: Diagnostic explanations (see 5.5.1 for more details): P1-P5 are prototypes selected for ADH-vs-rest classification. D1 and D2 are two test ducts. Each cell in the heatmap signifies the feature importance λ_k^l and feature difference $f_k^l(x)$ between prototype k and test duct x for l^{th} histological feature to obtain m (3) and generate prediction probability (5.3.6).

5.5.2 Model ablations

Impact of textural features: We investigated the inter-class heterogeneity and intra-class similarity of textural features across all classes from HLRBB and BRACS datasets to generate heatmaps (see Fig. 14).

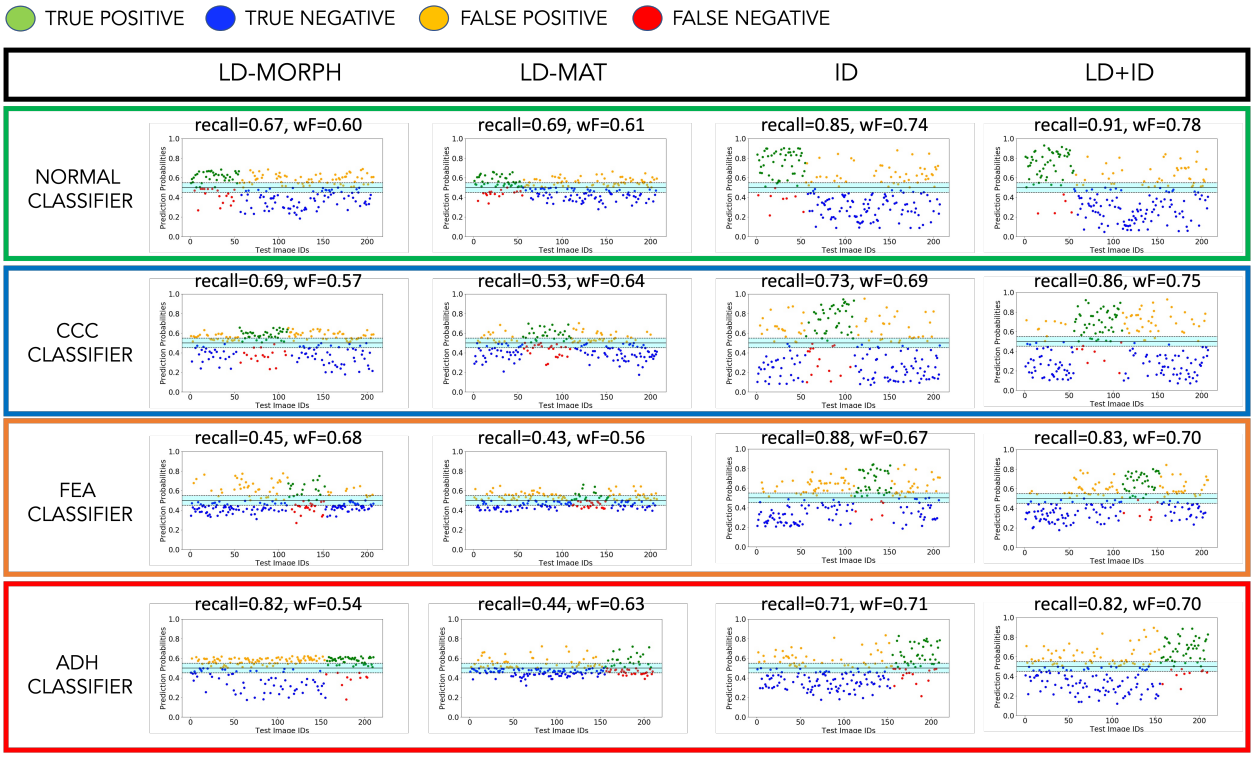

Figure 16: Distributions of prediction probabilities from each of our ML configurations on one testing set from the 5-fold cross-validation setting on the HLRBB dataset. Note the tight distribution along the decision boundary while using lumen/ductal morphology (LD-MORPH and LD-MAT). The inclusion of intraductal histological structural (ID) information improves the classification performance with higher confidence. Additionally, this decision boundary can be further deployed to explanation interface as the confidence level, where decisions within the boundary would have low confidence scores hence needs pathologist's intervention for this critical recommendation.

We performed additional experiments to analyze the impact of textural feature subset based on visualizing these heatmaps. Table 9 shows the mean of the recall and wF scores

obtained by performing the classification on a subset of textural features. For invasive, the textural features: *pixel LP stats*, *auto-cor-real*, *parent-mag-cor*, and *variance hpr* show less inter-class similarities and hence we used this feature subset to run the ML framework. However, this manual feature subset selection did not enhance the overall classification performance.

Table 8: Mean of the recall (R) and weighted F-scores (wF) (in %) from diagnostic classification of four classes (HLRBB dataset) and seven classes (BRACS dataset).

Dataset		I	D	A	F	U	B	N
HLRBB	**R**	-	-	64	65	-	62	77
	wF	-	-	65	68	-	66	75
BRACS	**R**	77	66	56	76	68	75	63
	wF	87	78	66	74	55	77	79

5.5.3 Conclusion

We presented a prototype-driven computational pathology pipeline based on analytical modeling of tissue features for the challenging task of diagnosing a broad spectrum of breast biopsies. We validated our approach on multiple classification scenarios across two datasets (HLRBB and BRACS) scanned at different resolutions and showed a significant improvement ($\approx 20\%$) over the state-of-the-art methods. A key highlight of our method is in its ability to provide pathologist friendly diagnostic explanations without largely compromising on the classification performance. We posit, the strategy outlined in this paper generalizes to tissue histologies from other organs as defined in the WHO Classification of Tumors book. Further, our approach can facilitate a communication platform between pathologists and computational scientists to interact and develop AI-driven algorithmic tools that can enhance patient care in a clinical setting.

Table 9: Mean and standard deviation of the recall (R) and weighted F-scores (wF) (in %) from diagnostic classification of four classes (HLRBB dataset) and seven classes (BRACS dataset) using manually selected subset of textural features (T-Subset).

Dataset		I	D	A	F	U	B	N
HLRBB	**T-Subset**	-	-	c,f,h,l,m,o	c,h,k,l,n,o	-	h,k,l,n,o	b,h,l,m,n,o
	R	-	-	64	65	-	62	77
	±	-	-	9	3	-	4	3
	wF	-	-	65	68	-	66	75
	±	-	-	2	3	-	3	1
BRACS	**T-Subset**	c,g,h,l,n,o	c,f,l,n,o	f,l,m,n,o	b,g,j,k,l,n	c,f,l,n,o	b,c,f,j,l,o	b,c,f,j,l,o
	R	77	66	56	76	68	75	63
	±	2	1	6	2	4	2	1
	wF	87	78	66	74	55	77	79
	±	1	1	1	1	1	1	1

6.0 Towards showcasing the potential for seamless integration of our computational pathology framework into clinical workflows

6.1 Chapter summary

A major bottleneck in developing and deploying AI-based anatomic pathology applications in clinical settings is the lack of access to high-performance computing required for training and testing computational pathology pipelines. On the contrary, easy access to low-compute devices ('edge devices') in a clinical setting permits deployment of already trained computational pathology pipelines. To integrate AI-based anatomic pathology applications in clinical settings, we propose training our computational pathology pipelines on remote customizable high-performance AI-enabled compute architectures provided by state-of-the-art data centers and applying the pipelines on edge devices for real-time clinical applications. For demonstration, our pipeline detects histological structures in breast tissue and classifies them into two diagnostic categories, high-risk and low-risk. The pipeline was trained on SambaNova SN108-R, an emerging customizable AI-based compute architecture provided by Argonne Leadership Computing Facility (ALCF). For performance comparison, we trained the same pipeline on CPU- and GPU-based architectures. We deployed the trained pipeline on an edge device to showcase the ability to generate diagnostic inferences in real-time.

6.2 Introduction

6.2.1 Background

Manual pathology practice is inefficient: Currently, manual review of glass microscope slides (GMS) is the standard for surgical pathology diagnosis. The manual workflow consists of pathologists first accessing and reviewing the Anatomic Pathology Laboratory Information (APLIS) to manually read the specimen information [110]. The slides are viewed sequentially

and one at a time to identify diagnostically salient regions such as calcifications, microorganisms, atypia, or cancer in a raster scan fashion. As a part of the diagnostic workflow, pathologists manually construct a report by transcription, typing or speech recognition [110]. Although, there exists few time saving strategies for report generation, however, it still requires final review and manual data entry. This manual pathology practice is inefficient, error-prone, and highly subjective [110, 111].

Additional issues affecting optimal patient care: In addition to inefficiencies within current analog routines of pathology, external developments are also concerning. The cancer cases are expected to rise with an increase in the aging population (increase from 1.7 million cases in 2012 to 2.3 million in 2025) [6]. As the number of cancer cases grow, the forecasted shortage of pathologists is alarming (declining from 5.7 to 3.7 per 100,000 people between 2010 and 2030) [7]. Today, major pathology practices are having problems with understaffing and increasing workloads. This is even more problematic in areas that are traditionally underserved (e.g., rural areas, community hospitals, etc.) [7]. The increase in the pathologists' workload coupled with the growth of the digital pathology market, has led to emerging black-box AI-based anatomic pathology applications.

Growth of digital pathology to improve anatomic pathology efficiency: DP enables the digitization of histological images, opening up many possible benefits. Having access to computing tools can provide pathologists with more quantifiable data relevant to risk assessment [13]. DP has proven to be successful for use in teleconferencing [14], allowing for simultaneous viewing of whole case files by multiple pathologists. This allows doctors the ability to virtually discuss a case with a pathologist who may be specialized in a field relative to the patient, such as lung pathology [13]. Additionally, DP makes the process of getting a second opinion much easier and faster. The digitization of case data also makes cross-site patient data synchronization much easier [13]. Further, the digitization also allows for the development of intelligent AI systems to aid in the diagnostic process. However, to realize the seamless integration of AI-based applications into the traditional diagnostic workflows, the system must achieve reasonable turnaround time and minimize the time-to-decision to achieve clinical adoption and possibly facilitate regulatory agency approvals.

Computational challenges to achieve required turnaround time in anatomic pathology

workflow: Turnaround time is defined as the time required by a system to execute an application which includes loading the data, running the application, and displaying the output on a screen [112]. In anatomic pathology, this turnaround time is critical for an AI system to achieve, especially when rapid and repetitive panning/zooming operations are performed during a search task. Further, the turnaround time might vary depending upon the application, image size, compute workload, etc.

However, the integration of AI-based applications in clinical workflows is not straightforward. A major bottleneck in developing and deploying AI-based anatomic pathology applications in clinical settings is the lack of access to high-performance computing required for training and testing computational pathology pipelines. On the contrary, easy access to low-compute devices ('edge devices') in a clinical setting permits deployment of already trained computational pathology pipelines. Additionally, to train these pipelines, we need to tackle computational barriers due to the massive data volume generated from high-resolution pathology WSIs. For example, a single core of breast biopsy tissue occupies between 1-20 gigabytes of storage space depending on the sample and resolution [112]. Although, the entire WSI is usually broken down into tiles, still the usage of a single GPU to train the ML models results in "out-of-memory" issues. Alternately, clustering multiple GPUs is not an optimal solution, since we need to deal with the problems caused by disaggregation of the computational workflows. Further, it is impossible to load the high-resolution images without sufficient tiling or downsampling which could lead to loss of contextual information and deliver less accurate results.

Our approach: To integrate AI-based anatomic pathology applications in clinical settings, we propose training our computational pathology pipelines on remote customizable high-performance AI-enabled compute architectures provided by state-of-the-art data centers and applying the pipelines on edge devices for real-time clinical applications. For demonstration, our pipeline detects histological structures in breast tissue and classifies them into two diagnostic categories, high-risk and low-risk. The pipeline was trained on SambaNova SN108-R [113, 114], an emerging customizable AI-based compute architecture provided by Argonne Leadership Computing Facility (ALCF). For performance comparison, we trained the same pipeline on CPU- and GPU-based architectures. We deployed the trained pipeline

on an edge device to showcase the ability to generate diagnostic inferences in real-time.

6.2.2 Related Work

Turnaround time in digital pathology: An innovative framework called *pathologists' computer-assisted diagnosis* (pCAD) was projected as a result of the emerging trends in DP which motivated the application of AI towards visual assessment of the tissue slides [20]. This framework was used to carry out an experiment to evaluate time spent by the pathologists in diagnosing breast core biopsies on each field of view under the microscope. The average time required to manually review a biopsy was 221.6 seconds and the average simulated time to review using the pCAD framework was 98 seconds, amounting to a 56% reduction in time-to-decision. The experimental setup described in [115] is significant to support the hypothesis that computational pathology workflows can improve anatomic pathology efficiency and minimize time-to-decision. This hypothesis is further supported by the work in [116], where the author demonstrated a 10% shorter time-to-decision from the digital pathology workflow when compared to traditional pathology for diagnosing 400 biopsy cases.

Application of AI systems: The authors in [113] evaluated the relevance of using customizable AI-based computing architecture (SambaNova system) on a diverse range of applications such as, *BERT* (Bidirectional encoder representations from transformers) for natural language processing, *CosmicTagger* to detect neutrino interactions from high-resolution images obtained from the neutrino detectors, *Gravwaves* to observe astrophysical phenomena using gravitational waves, etc. Brace et al. explored the suitability of Cerebras for running molecular dynamics simulations for guiding the conformational search and demonstrated significant performance gains over traditional CPU/ GPU hardware. The authors in [117] demonstrated an automated workflow for rapid (re)training of deep-learning networks on AI accelerators (Cerebras and Sambanova) and facilitating model deployment in real-time at the edge. This workflow showed the feasibility of using high-powered remote AI accelerators (e.g., Sambanova) to promote fast training/ retraining for low-cost data processing on edge devices. However, no substantial studies have been performed to demonstrate a proof-of-concept blueprint to bring such technologies to anatomic pathology workflows.

6.3 Methods

We built a workflow to train computational pathology pipelines on remote customizable AI-based compute architecture (SambaNova SN108-R) provided by Argonne Leadership Computing Facility. In this section, we first provide an overview of the SambaNova architecture. Second, we describe the experimental setup of our computational pathology training pipeline. Third, we evaluate the limitations of running the training pipeline on traditional CPU- and GPU-based architectures. Finally, we demonstrate the deployment of trained pipeline on an edge device to showcase the ability to generate diagnostic inferences in real-time.

6.3.1 SambaNova reconfigurable dataflow architecture

The SambaNova Reconfigurable Dataflow Architecture (RDA) is a high-performance computing architecture catered to the next generation of AI applications. The reconfigurable dataflow architecture enables high-throughput processing of the high-resolution tissue biopsy images. Further, the SambaFlow software framework allows seamless operation by providing an automated support for convolution overlap handling, intermediate tensors, and tiled images. Additionally, it is capable of processing images of very high-resolution ($\approx 4K - 50K$) on a single DataScale system which favors its usage for deploying AI-based anatomic pathology applications. For this study, we used the SambaNova DataScale system, SN108-R deployed at the Argonne Leadership Computing Facility (ALCF). Each system consists of a host module and 8 Reconfigurable Dataflow Units (RDUs) which are interconnected via the RDU-Connect, while the systems are connected using an InfiniBand-based interconnect. This allows for a seamless model and data parallelism across the RDUs in the SN108-R SambaNova system required for integrating AI-based applications in a clinical setting.

6.3.2 Experimental setup

A visual illustration of the the experimental setup is shown in Fig. 17. As shown in Fig. 17A, we implement a computational pathology framework designed for rapid training of two deep-learning (DL) networks on the SambaNova system to detect the diagnostically important histological structure, duct(s) in the breast tissues and classify them into two diagnostic categories, high-risk and low-risk. In addition, we demonstrate the clinical anatomic pathology application pipeline to deploy the trained models on an edge device and showcase the ability to generate diagnostic inferences in real-time (see Fig. 17B). The methods outlined here can be easily adapted to different datasets or different DL networks. We quantitate the performance based on end-to-end training time, model throughput, and latency. We further compare the end-to-end training time of the proposed workflow of the AI system (SambaNova) with the following CPU- and GPU-based architectures: Intel Xeon Gold 6234 CPU, TitanX GPU, NVIDIA V100 GPU, and NVIDIA A100 Tensor core GPU. To implement, we developed a workflow orchestrator on the open-source PyTorch framework to enable data loading, model configuration, and deployment on the machines. Further, to enable data parallelism, we packaged our workflow using DataDistributedParallel (DDP) training module from the PyTorch framework.

6.3.3 Training computational pathology pipelines

To demonstrate the proposed workflow of training our computational pathology pipelines, we implemented two DNNs, U-Net [118] and ResNet18 [119]. For training, we used the breast carcinoma subtyping (BRACS) dataset [107] scanned at 40X resolution. A detailed description of the dataset is provided in Section 4.4.1. For the sake of classification, the diagnostic labels *Normal*, *Pathological benign*, and *UDH* were regrouped as low-risk and *FEA*, *ADH*, and *DCIS* were regrouped as high-risk. The details of model implementation is described below.

1. U-Net: For duct segmentation, we deployed a fully convolutional neural network-based U-Net architecture [118]. U-Net has shown promising results for semantic segmentation in medical images. The convolution and pooling operations in U-Net allow the model to

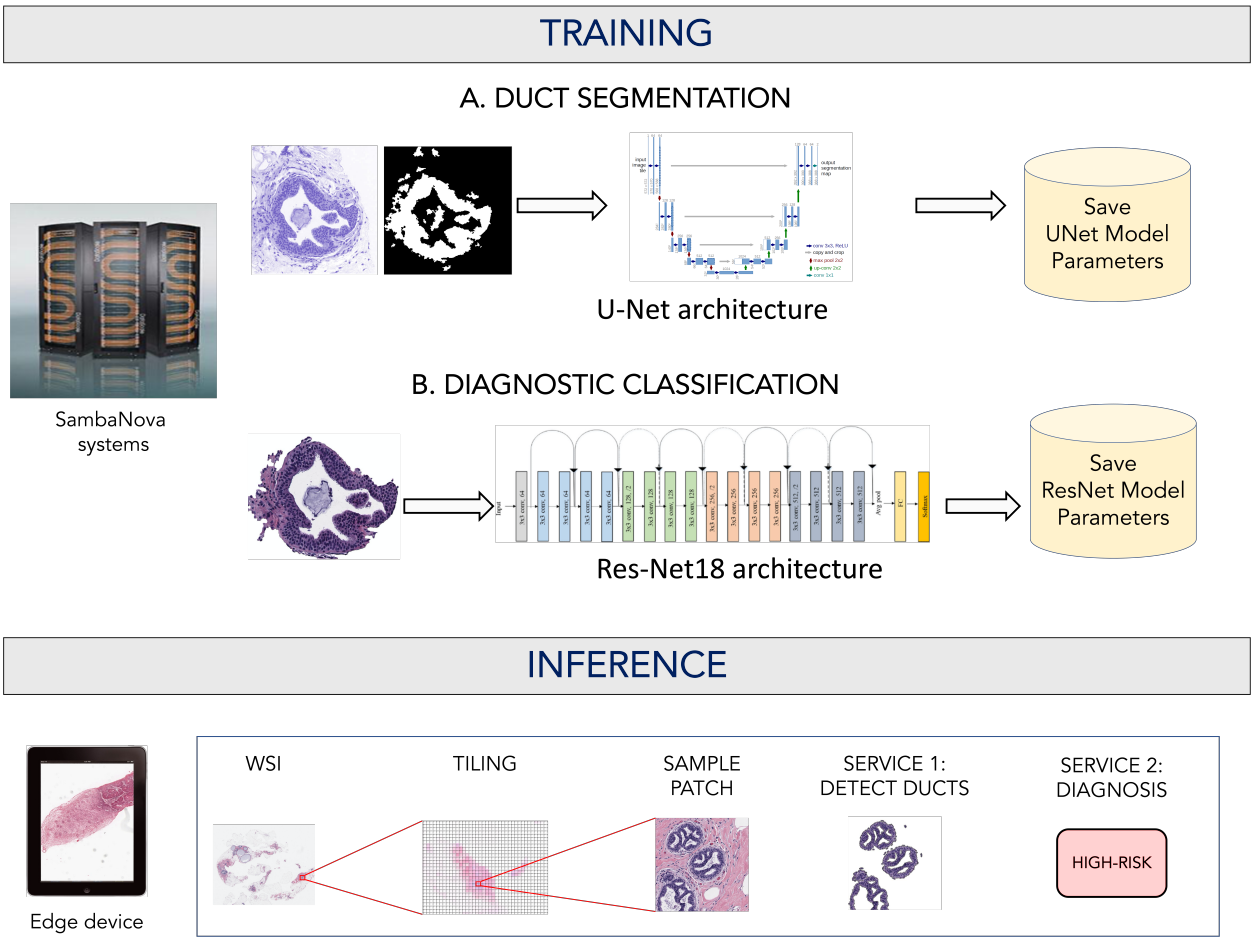

Figure 17: (A) Computational pathology training pipeline on a remote customizable high-performance AI-enabled compute architecture (SambaNova) to detect duct(s) in breast tissue and classify them into two diagnostic categories, high-risk and low-risk using U-Net and ResNet-18 DL networks respectively. (B) Clinical anatomic pathology application pipeline to deploy the trained model from (A) on an edge device and demonstrate real-time diagnostic inferences to detect ducts in a breast tissue and predict its diagnostic label.

learn important structural relationships that will aid in the segmentation. In the case of duct segmentation, we want the model to learn the spatial relationships between the duct and stroma.

Data preprocessing: To organize the dataset for duct segmentation, we first normalized the ROI images from BRACS dataset to $1K \times 1K$ and extracted the hematoxylin channel by performing color deconvolution. Next, we obtained ground-truth duct masks upon applying the duct segmentation algorithm described in section 3.3.1. To summarize, the algorithm breaks the WSI into non-overlapping superpixels, assigns each superpixel a probability of belonging to a duct structure to create a probability map, and finally uses the estimate of the boundary to generate a refined boundary for the duct. Since the original method [80] was trained on 20X images, the images from BRACS dataset were downsampled from 40X to 20X prior to applying this method. The 500×500 downsampled images along with their ground-truth masks were used to train the U-Net architecture on the Sambanova system. However, to overcome memory issues on CPU and single node GPU machines, the images and binary masks were resized to (112,112) prior to training.

Model implementation: To train the U-Net architecture, we used the Dice loss function and monitored the Intersection over Union (IoU) scores with a decision threshold of 0.5. The U-Net model has 23 convolutional layers generating ≈ 36.9 million trainable parameters as indicated in Table 10. We used an Adam optimizer with a learning rate of 1×10^{-4} and trained the network for 200 epochs.

2. ResNet - We deployed a ResNet-18 CNN model to diagnose the breast biopsy images as low-risk or high-risk. The ResNet architecture introduced skip connections to handle information loss during training. Several studies have explored the performance of ResNet on histopathology images [119]. The training and validation dataset consisted of 2358 and 474 ductal ROIs respectively of size $1K \times 1K$. To overcome memory issues on CPU and single node GPU machines, the RGB images were resized to (256,256).

Model implementation: We used transfer learning with a pretrained ResNet-18 model from the PyTorch vision models. The ResNet-18 model has 18 convolutional layers generating ≈ 11.2 million trainable parameters as indicated in Table 10. Adam optimizer with a weight decay of 5×10^{-4}, learning rate of 1×10^{-4} was used to optimize the categorical cross-entropy loss function for 200 epochs with a batch size of 32.

To show the feasibility of training high-resolution images on the AI-based system on larger batch sizes, we performed additional experiments on the SambaNova system by scaling the

batch-size from 32 to 64 on the training images without further downsampling or tiling.

6.3.4 Deploying trained pipeline on an edge device

We deployed the trained pipeline on an edge device to showcase the ability to generate diagnostic inferences in real-time. For demonstration purpose, we used the low-computing Intel(R) Core(TM) i7-4770 CPU @ 3.40 GHz machine as the edge device. To simulate a typical anatomic pathology workflow, we designed our inference pipeline to load a WSI, segment it into multiple tiles, load the trained pipelines from U-Net and ResNet-18 models and obtain segmented duct(s) and diagnostic label for each tile (see Fig. 17B).

6.4 Results and discussion

6.4.1 Computational performance of CPU- and GPU-based architectures

Table 10: Computational performance

Model	UNet	ResNet
Xeon CPU	1100.18	270.93
TitanX GPU	151.22	55.62
V100 GPU	276.87	97.37
A100 ThetaGPU	583.97	82.39

(a) Training time on different hardware devices (in minutes) to run UNet and ResNet18 networks for 200 epochs.

Model	UNet	ResNet
Input	(112,112)	(256,256,3)
Loss fcn	Dice	BCE
# parameters	36.9 M	11.2 M

(b) Model description

For performance comparison, we trained the computational pathology pipeline described in the previous section on CPU- and GPU-based architectures. We recorded the training time, training loss and validation accuracies generated per epoch. Fig. 18 illustrates the

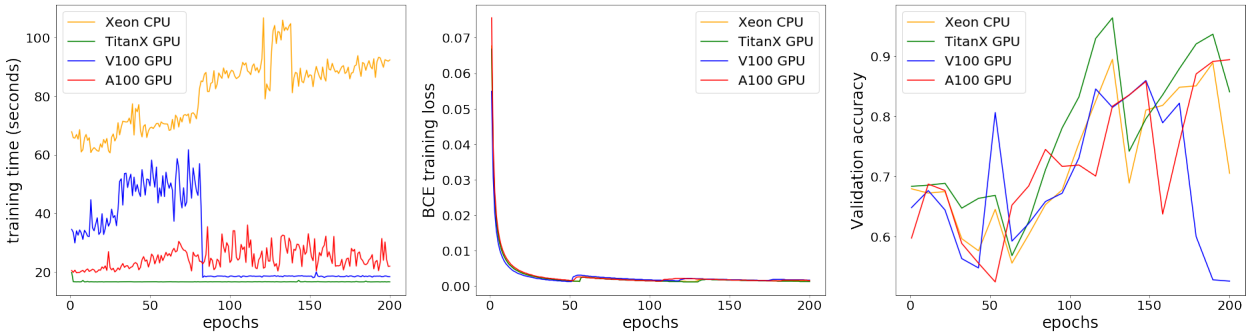

Figure 18: Comparison of (i) training time per epoch, (ii) BCE training loss, and (iii) validation accuracy generated by implementing a Res-Net18 architecture for diagnostic classification on different hardware devices.

computational performance measurement for 200 epochs. The end-to-end training time from the traditional architectures is reported in Table 10a. For all the experiments, we resized the images to (112,112) for U-Net and to (256,256) for Res-Net to avoid "out-of-memory" issues that were initially encountered.

We analyze the computational performance of all the four traditional hardware architectures to train the U-Net network. TitanX GPU provides an approximate speed up of 7x, 1.8x, and 3.8x compared to Xeon CPU, V100 GPU (deployed on Summit at Oak Ridge National Laboratory), and NVIDIA A100 GPU. Additionally, to train the Res-Net, TitanX GPU exceeds the performance speed when compared to CPU, V100, and A100 GPUs by 4.9x, 1.7x, and 1.5x respectively. However, neither of the devices could support the original image size of 500×500 and larger batch sizes which is a major bottleneck in deploying AI-based applications on the high-resolution pathology images.

6.4.2 Computational performance of customizable AI-based compute architecture

To measure and compare the computational performance on the customizable high-performance AI-enabled compute architecture, SambaNova SN108-R, we capture several

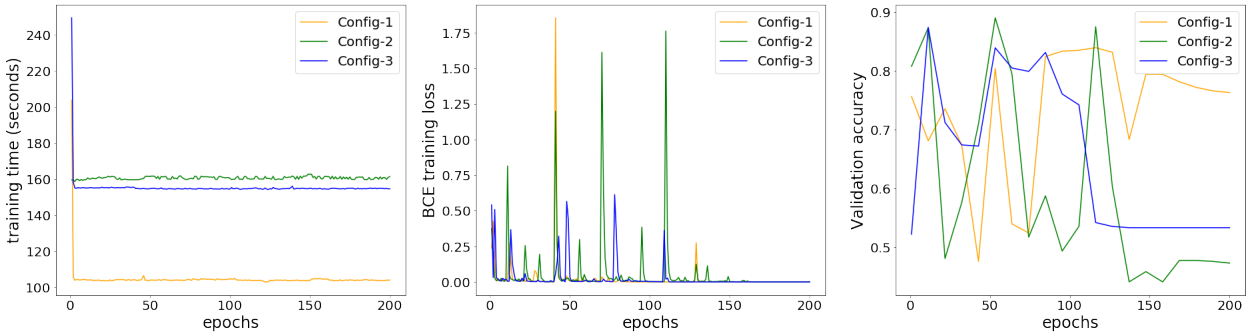

Figure 19: Comparison of (i) training time per epoch, (ii) BCE training loss, and (iii) validation accuracy generated by implementing a Res-Net18 DL network on the SambaNova system for three training configurations. (1) Config-1: image size of (256,256) and batch-size of 32. (2) Config-2: image size of (500,500) and batch-size of 32, and (3) Config-3: image size of (500,500) and batch-size of 64.

performance metrics during training and evaluate the feasibility of using high-performance AI systems on high-resolution images and larger batch sizes by performing the following three experiments:

1. Training on downsampled images (256,256) with batch-size 32: The performance of training the UNet architecture on (256,256)-sized images resulted in an end-to-end throughput of 65.90 samples/second and average latency of 0.48 seconds. Further, training the RescaleNet18 (a variant of ResNet suitable for the SambaNova software framework) on the downsampled image with batch-size of 32 resulted in an end-to-end throughput of 241.15 samples/second and an average latency of 0.13 seconds.

2. Training on original images (500,500) with batch-size 32: The performance of training the UNet architecture on (500,500)-sized images resulted in an end-to-end throughput of 16.68 samples/second and average latency of 1.92 seconds. Further, training the RescaleNet18 on the original image without resizing with batch-size of 32 resulted in an end-to-end throughput of 41.45 samples/second and an average latency of 0.77 seconds.

3. Training on original images (500,500) with batch-size 64: Scaling the batch-size to 64 to run RescaleNet resulted in an end-to-end throughput of 42.04 samples/second and an

average latency of 1.52 seconds.

The training time, training loss and validation accuracies generated per epoch from running the experiments on the SambaNova system is illustrated in Fig. 19. We observe that training the pipeline on downsampled images is faster even on the SambaNova system, however it was possible to run the full-sized images on the high-performance computing architecture. Interestingly, increasing the batch-size by 2x did not significantly reduce the end-to-end throughput (41.45 samples/ second vs 42.04 samples/second). However, the end-to-end training time on the traditional CPU- and GPU-based architectures was lower than the SambaNova system. Upon further investigation, we could justify the suboptimal performances based on the fact that more hands-on-work needs to be done by the SambaNova engineers in order to optimize the compilation to take full advantage of the AI-hardware accelerators. Currently, the system is under active development stage and not all layers of the SambaNova software stack (SambaFlow) is fully optimized. It should be noted that, the purpose of this study was to demonstrate the feasibility of using powerful AI-enabled systems for deploying on the edge devices. The proposed computational pathology pipeline is highly flexible and can be readily adapted on future versions of the SambaNova system.

6.4.3 Deployment of trained pipeline on the edge device

We successfully implemented the workflow described in section 6.3.4 for detecting duct(s) and predicting the diagnostic category on the edge device. We were able to achieve an average diagnostic inference time of 38.28 milliseconds and thereby showcasing the ability to deploy real-time clinical applications.

6.5 Conclusions and future work

In this chapter, we discussed some issues to integrate AI-based anatomic pathology applications into clinical workflows and presented an approach to address them. In particular, we trained our computational pathology pipelines on remote customizable high-performance AI-

enabled compute architectures provided by state-of-the-art data centers such as, ALCF and applied the trained models on edge devices for real-time clinical applications. For demonstration, we tested our framework on two anatomic pathology applications of detecting duct(s) in the breast tissue and classifying them into two diagnostic categories of high- or low-risk. We successfully showed the feasibility of using the emerging customizable AI-based compute architecture, SambaNova SN108-R for training and deploying the trained models in real-time on low-compute edge devices which is easily accessible in a clinical setting.

Future work: This work can be developed further in several directions. First, more hands-on work needs to be done by the Sambanova engineers in order to optimize the compilation to take advantage of the AI-hardware accelerators. The system is currently under active development stage and all the layers in the software stack are not fully optimized. Second, the current workflow demonstrates the feasibility of using powerful AI systems, however, a systematical benchmarking study needs to be conducted to compare the computational performance. Third, similar to the study conducted on the pCAD framework [20], a visual psychophysics experiment can be conducted to provide qualitative and quantitative assessment of integrating AI-based solutions into the clinical workflow to aid the pathologists in decision-making. Fourth, the existing computational pathology training pipeline can be modified to enable re-training or on-the-fly training as and when new pathologist annotated training data becomes available.

7.0 Conclusions

In this thesis, we built an interpretable computational pathology framework based on analytical modeling of already established anatomic pathology knowledge. Additionally, we demonstrated the feasibility of integrating AI-based applications into the clinical workflows by training on the emerging remote high-performance AI-based compute architecture and deploying the trained model on low-compute edge devices which are easily accessible within a clinical setting. The major advantage of invoking analytical models of the diagnostically meaningful hand-crafted features for prototype-driven ML classification is that they are flexible to adapt as the clinical practice changes. Also, our proposed computational pathology framework can easily incorporate new morphology tissue features. Further, the methods developed in this thesis generalizes to tissue histologies from other organs since a majority (80-90%) of cancer cases originate from epithelial tissue malignancies. The key contribution of our framework is that we have built a communication platform for pathologists and computational scientists to interact and develop AI-based applications and enhance patient care in a clinical setting.

In **Chapter 2**, we developed analytical models for a dictionary of tissue features that aid in the differential diagnosis of atypical breast biopsies. Following guidelines in the WHO classification of tumors of the breast and in consultation with our team of breast pathology sub-specialists, we assembled a visual dictionary of 16 tissue features that the pathologists frequently use in making complex diagnostic decisions for atypical breast biopsies. This strategy has the potential to extend to other organ systems and act as a surrogate in the case review and quality assurance discussions for reducing discordance between pathologists. Our approach of analytically modeling the tissue features that traditionally define the standard on tumor classification/ nomenclature for pathologists worldwide is the first of its kind.

In **Chapter 3**, we reframed the computational diagnosis of breast biopsies as a problem of prototype recognition following a hypothesis that pathologists mentally relate current tissue features to previously encountered features during their routine diagnostic work. We developed an unsupervised method to obtain the relative importance of the tissue features in

a set of pre-selected prototypical ductal ROIs in classifying the breast biopsies. This model provided clinically relevant explanations to its recommendations which is easily interpretable.

In **Chapter 4**, we significantly enhanced the entire computational pathology framework for differential diagnoses of a broad spectrum of breast biopsies. We validated our approach on multiple classification scenarios across two datasets scanned at different resolutions and showed a significant improvement over the state-of-the-art methods. The key highlight of this chapter was in showing the ability of our ML method to provide pathologist friendly explanations without largely compromising on the classification performance.

In **Chapter 5**, we demonstrated the feasibility of integrating AI-based anatomic pathology applications in clinical settings by training our computational pathology pipelines on remote customizable high-performance AI-based compute architectures and deployed the trained models in real-time on low-compute edge device.

As the next target for development, there are some addressable limitations which include: a. DCIS dominant patterns such as *solid*, *comedo necrosis*, and *micropapillary* have not been included. b. For invasive carcinoma, cell specific features such as the mitotic count can be beneficial. We anticipate that the inclusion of analytical models of missing features will further enhance the classification performance. c. imperfect duct, lumen, and nuclei segmentation can negatively impact the analytical models. d. Need for rigorous feature perturbations analysis on the diagnostic classification performance, such as increasing size of prototype dataset.

The digital and computational pathology is revolutionizing the anatomic pathology workflows as a result of which several AI-based applications are being developed. However, for clinical adoption and FDA approvals, the AI system must be transparent and trustworthy. Keeping this in mind, in this thesis we collaborated with a pathology expert to help us understand what tissue features are they paying attention to while diagnosing breast lesions and then we built analytical models to capture them. For easy visual interpretability, we used these analytical models to drive a prototype-driven ML framework. Finally, we also demonstrated a proof-of-concept study to integrate our computational pathology framework into clinical workflows. The principal advantage of our computational pathology framework is that it is flexible to adapt as the clinical practice changes and we can incorporate addi-

tional tissue features to make the system more robust. As an additional future direction, we can conduct visual psychophysics experiments of our framework with the pathologists to evaluate if our AI system can improve the workflow efficiency.

Bibliography

[1] Pushpak Pati, Guillaume Jaume, Antonio Foncubierta, Florinda Feroce, Anna Maria Anniciello, Giosuè Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, et al. Hierarchical graph representations in digital pathology. *arXiv preprint arXiv:2102.11057*, 2021.

[2] American Cancer Society. Breast cancer facts & figures 2019–2020. *Am. Cancer Soc*, pages 1–44, 2019.

[3] Andrew J Evans, Thomas W Bauer, Marilyn M Bui, Toby C Cornish, Helena Duncan, Eric F Glassy, Jason Hipp, Robert S McGee, Doug Murphy, Charles Myers, et al. Us food and drug administration approval of whole slide imaging for primary diagnosis: a key milestone is reached and new questions are raised. *Archives of pathology & laboratory medicine*, 142(11):1383–1387, 2018.

[4] Clia laboratory guidance during covid-19 public health emergency.

[5] Anil V. Parwani. Next generation diagnostic pathology: use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. *Diagnostic Pathology*, 14(1):138, Dec 2019.

[6] Digital pathology market forecast to 2022 by markets & markets, 2017.

[7] Stanley J Robboy, Sally Weintraub, Andrew E Horvath, Bradden W Jensen, C Bruce Alexander, Edward P Fody, James M Crawford, Jimmy R Clark, Julie Cantor-Weinberg, Megha G Joshi, et al. Pathologist workforce in the united states: I. development of a predictive model to examine factors influencing supply. *Archives of Pathology and Laboratory Medicine*, 137(12):1723–1732, 2013.

[8] Sandy Mullay, Theodore R Newell, and Jeffery Prichard. Evolving workflow drives anatomic pathology design. *Evolving Workflow Drives Anatomic Pathology Design : December 2017 - MedicalLab Management Magazine*, Dec 2017.

[9] Jeffrey Chun Tatt Lim, Joe Poh Sheng Yeong, Chun Jye Lim, Clara Chong Hui Ong, Siew Cheng Wong, Valerie Suk Peng Chew, Syed Salahuddin Ahmed, Puay Hoon Tan, and Jabed Iqbal. An automated staining protocol for seven-colour immunofluorescence of human tissue sections for diagnostic and prognostic use. *Pathology*, 50(3):333–341, Apr 2018.

[10] Alexi Baidoshvili, Anca Bucur, Jasper van Leeuwen, Jeroen van der Laak, Philip Kluin, and Paul J van Diest. Evaluating the benefits of digital pathology implementation: time savings in laboratory logistics. *Histopathology*, 73(5):784–794, 2018.

[11] Kabeer K. Shah, Julia S. Lehman, Lawrence E. Gibson, Christine M. Lohse, Nneka I. Comfere, and Carilyn N. Wieland. Validation of diagnostic accuracy with whole-slide imaging compared with glass slide review in dermatopathology. *Journal of the American Academy of Dermatology*, 75(6):1229–1237, Dec 2016.

[12] Enforcement policy for remote digital pathology devices during the coronavirus disease 2019 (covid-19) public health emergency, 2020.

[13] Kevin S. McDorman, Curtis Chan, Jennifer Rojko, Christina M. Satterwhite, and James P. Morrison. Chapter 7 - special techniques in toxicologic pathology. In Wanda M. Haschek, Colin G. Rousseaux, and Matthew A. Wallig, editors, *Haschek and Rousseaux's Handbook of Toxicologic Pathology (Third Edition)*, pages 175 – 214. Academic Press, Boston, third edition edition, 2013.

[14] Liron Pantanowitz, Paul N. Valenstein, Andrew J. Evans, Keith J. Kaplan, John D. Pfeifer, David C. Wilbur, Laura C. Collins, and Terence J. Colgan. Review of the current state of whole slide imaging in pathology. *Journal of pathology informatics*, 2:36–36, 2011.

[15] Liron Pantanowitz, Ashish Sharma, Alexis B. Carter, Tahsin Kurc, Alan Sussman, and Joel Saltz. Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *Journal of pathology informatics*, 9:40–40, Nov 2018.

[16] David N. Louis, Michael Feldman, Alexis B. Carter, Anand S. Dighe, John D. Pfeifer, Lynn Bry, Jonas S. Almeida, Joel Saltz, Jonathan Braun, John E. Tomaszewski, John R. Gilbertson, John H. Sinard, Georg K. Gerber, Stephen J. Galli, Jeffrey A. Golden, and Michael J. Becich. Computational pathology: A path ahead. *Archives of pathology & laboratory medicine*, 140(1):41–50, Jan 2016.

[17] Tony J Collins. Imagej for microscopy. *Biotechniques*, 43(S1):S25–S30, 2007.

[18] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

[19] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.

[20] Navid Farahani, Zheng Liu, Dylan Jutt, and Jeffrey L Fine. Pathologists' computer-assisted diagnosis: a mock-up of a prototype information system to facilitate automation of pathology sign-out. *Archives of pathology & laboratory medicine*, 141(10):1413–1420, 2017.

[21] Baris Gecer, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern recognition*, 84:345–356, 2018.

[22] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. From patch-level to roi-level deep feature representations for breast histopathology classification. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 109560H. International Society for Optics and Photonics, 2019.

[23] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G Elmore, and Linda Shapiro. Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 893–901. Springer, 2018.

[24] Yan Hao, Shichang Qiao, Li Zhang, Ting Xu, Yanping Bai, Hongping Hu, Wendong Zhang, and Guojun Zhang. Breast cancer histopathological images recognition based on low dimensional three-channel features. *Frontiers in Oncology*, 11:2018, 2021.

[25] A. D. Belsare, M. M. Mushrif, M. A. Pangarkar, and N. Meshram. Classification of breast cancer histopathology images using texture feature analysis. In *TENCON 2015 - 2015 IEEE Region 10 Conference*, pages 1–5, 2015.

[26] Taha J. Alhindi, Shivam Kalra, Ka Hin Ng, Anika Afrin, and Hamid R. Tizhoosh. Comparing lbp, hog and deep features for classification of histopathology images. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2018.

[27] Ahmad Chaddad, Christian Desrosiers, Lama Hassan, and Matthew Toews. Multi-spectral texture analysis of histopathological abnormalities in colorectal tissues. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2628–2632, 2016.

[28] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

[29] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6), 2017.

[30] Iciar 2018 grand challenge on breast cancer histology images.

[31] Kamyar Nazeri, Azad Aminpour, and Mehran Ebrahimi. Two-stage convolutional neural network for breast cancer histology image classification. In *International Conference Image Analysis and Recognition*, pages 717–726. Springer, 2018.

[32] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *arXiv preprint arXiv:2004.09666*, 2020.

[33] Pushpak Pati, Guillaume Jaume, Lauren Alisha Fernandes, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Daniel Riccio, Maurizio Di Bonito, et al. Hact-net: A hierarchical cell-to-tissue graph neural network for histopathological image classification. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pages 208–219. Springer, 2020.

[34] Guillaume Jaume, Pushpak Pati, Behzad Bozorgtabar, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Tilman Rau, Jean-Philippe Thiran, Maria Gabrani, and Orcun Goksel. Quantifying explainers of graph neural networks in computational pathology. *arXiv preprint arXiv:2011.12646*, 2020.

[35] Guillaume Jaume, Pushpak Pati, Antonio Foncubierta-Rodriguez, Florinda Feroce, Giosue Scognamiglio, Anna Maria Anniciello, Jean-Philippe Thiran, Orcun Goksel, and Maria Gabrani. Towards explainable graph representations in digital pathology. *arXiv preprint arXiv:2007.00311*, 2020.

[36] Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical image analysis*, 63:101696, 2020.

[37] Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[38] Akif Burak Tosun et al. Histological detection of high-risk benign breast lesions from whole slide images. In *International Conference on MICCAI*, pages 144–152, 2017.

[39] Ezgi Mercan et al. Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. *JAMA Network Open*, 2(8):e198777–e198777, 2019.

[40] Gabriel García, Adrián Colomer, and Valery Naranjo. First-stage prostate cancer identification on histopathological images: Hand-driven versus automatic learning. *Entropy*, 21(4):356, 2019.

[41] Gabriel García, Adrián Colomer, and Valery Naranjo. Cvblab.

[42] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[43] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Deeppath.

[44] Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific reports*, 9(1):1–8, 2019.

[45] Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. Deepslide.

[46] Jonine D Figueroa et al. Standardized measures of lobular involution and subsequent breast cancer risk among women with benign breast disease: a nested case–control study. *Breast Cancer Research and Treatment*, 159(1):163–172, 2016.

[47] Richard J Santen. Benign breast disease in women. In *Endotext [Internet]*. MDText. com, Inc., 2018.

[48] Sara W Dyrstad et al. Breast cancer risk associated with benign breast disease: systematic review and meta-analysis. *Breast Cancer Research and Treatment*, 149(3):569–575, 2015.

[49] Joann G Elmore et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*, 313(11):1122–1132, 2015.

[50] E. Mercan et al. Assessment of Machine Learning of Breast Pathology Structures for Automated Differentiation of Breast Cancer and High-Risk Proliferative Lesions. *JAMA Netw Open*, 2(8):e198777, Aug 2019.

[51] B. E. Bejnordi et al. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *J Med Imaging (Bellingham)*, 4(4):044504, Oct 2017.

[52] H. Li et al. Quantitative nuclear histomorphometric features are predictive of Oncotype DX risk categories in ductal carcinoma in situ: preliminary findings. *Breast Cancer Res.*, 21(1):114, 2019.

[53] F. Dong et al. Computational pathology to discriminate benign from malignant intraductal proliferations of the breast. *PLoS ONE*, 9(12):e114885, 2014.

[54] Sunil R Lakhani. *WHO Classification of Tumours of the Breast*. International Agency for Research on Cancer, 2012.

[55] J. Schindelin et al. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012.

[56] R. Achanta et al. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans Pattern Anal Mach Intell*, 34(11):2274–2282, Nov 2012.

[57] T. F. Chan et al. Active contours without edges. *IEEE Trans Image Process*, 10(2):266–277, 2001.

[58] Kenneth J Pienta et al. Correlation of nuclear morphometry with progression of breast cancer. *Cancer*, 68(9):2012–2016, 1991.

[59] Yan Cui et al. Nuclear morphometric features in benign breast tissue and risk of subsequent breast cancer. *Breast Cancer Research and Treatment*, 104(1):103–107, 2007.

[60] Anamika Kashyap et al. Study of nuclear morphometry on cytology specimens of benign and malignant breast lesions: A study of 122 cases. *Journal of Cytology*, 34(1):10, 2017.

[61] Aparna Narasimha et al. Significance of nuclear morphometry in benign and malignant breast aspirates. *International Journal of Applied and Basic Medical Research*, 3(1):22, 2013.

[62] Ellen CM Mommers et al. Prognostic value of morphometry in patients with normal breast tissue or usual ductal hyperplasia of the breast. *International Journal of Cancer*, 95(5):282–285, 2001.

[63] Yoshiko Yamashita et al. Does flat epithelial atypia have rounder nuclei than columnar cell change/hyperplasia? a morphometric approach to columnar cell lesions of the breast. *Virchows Archiv*, 468(6):663–673, 2016.

[64] Angela Flavia Logullo et al. Columnar cell lesions of the breast: a practical review for the pathologist. *Surgical and Experimental Pathology*, 2(1):1–8, 2019.

[65] Sarah E Pinder et al. Non-operative breast pathology: columnar cell lesions. *Journal of Clinical Pathology*, 60(12):1307–1312, 2007.

[66] Kimberly H Allison et al. Histological features associated with diagnostic agreement in atypical ductal hyperplasia of the breast: illustrative cases from the b-path study. *Histopathology*, 69(6):1028–1046, 2016.

[67] Sergio Rey et al. pysal/pointpats: pointpats 2.1.0, July 2019.

[68] Naiyun Zhou et al. Large scale digital prostate pathology image analysis combining feature extraction and deep neural network. *arXiv:1705.02678*, 2017.

[69] Nitesh V Chawla et al. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[70] Kim A Ely et al. Core biopsy of the breast with atypical ductal hyperplasia: a probabilistic approach to reporting. *The American Journal of Surgical Pathology*, 25(8):1017–1021, 2001.

[71] Lin-Ying Chen et al. Diagnostic upgrade of atypical ductal hyperplasia of the breast based on evaluation of histopathological features and calcification on core needle biopsy. *Histopathology*, 75(3):320–328, 2019.

[72] Yann LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20:5, 2015.

[73] Alex Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[74] Pierre Sermanet et al. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR, CBLS.* 2014.

[75] Melvin Silverstein. Where's the outrage? *Journal of the American College of Surgeons*, 208(1):78–79, 2009.

[76] Stuart J Schnitt and James L Connolly. Processing and evaluation of breast excision specimens: a clinically oriented approach. *American journal of clinical pathology*, 98(1):125–137, 1992.

[77] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, 2009.

[78] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in neural information processing systems*, pages 8930–8941, 2019.

[79] Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 32–40, 2019.

[80] Akash Parvatikar, Om Choudhary, Arvind Ramanathan, Olga Navolotskaia, Gloria Carter, Akif Burak Tosun, Jeffrey L Fine, and S Chakra Chennubhotla. Modeling histological patterns for differential diagnosis of atypical breast lesions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 550–560. Springer, 2020.

[81] Sachin Mehta, Ximing Lu, Donald Weaver, Joann G Elmore, Hannaneh Hajishirzi, and Linda Shapiro. Hatnet: An end-to-end holistic attention network for diagnosis of breast biopsy images. *arXiv preprint arXiv:2007.13007*, 2020.

[82] Beibin Li, Ezgi Mercan, Sachin Mehta, Stevan Knezevich, Corey W Arnold, Donald L Weaver, Joann G Elmore, and Linda G Shapiro. Classifying breast histopathology images with a ductal instance-oriented pipeline.

[83] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[85] Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, pages 2403–2424, 2011.

[86] Sarah B Hugar, Rohit Bhargava, David J Dabbs, Katie M Davis, Margarita Zuley, and Beth Z Clark. Isolated flat epithelial atypia on core biopsy specimens is associated with a low risk of upgrade at excision. *American journal of clinical pathology*, 151(5):511–515, 2019.

[87] Ronen Basri. Recognition by prototypes. *International Journal of Computer Vision*, 19(2):147–167, 1996.

[88] Mingbo Ma, Ming Shao, Xu Zhao, and Yun Fu. Prototype based feature learning for face image set classification. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[89] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017.

[90] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[91] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.

[92] John Wilder, Morteza Rezanejad, Kaleem Siddiqi, Allan Jepson, Sven Dickinson, and Dirk Walther. Local contour symmetry facilitates the neural representation of scene categories in the ppa. 01 2019.

[93] John Wilder, Morteza Rezanejad, Sven Dickinson, Kaleem Siddiqi, Allan Jepson, and Dirk Walther. The perceptual advantage of symmetry for scene perception. *Journal of Vision*, 17:1091, 08 2017.

[94] John Wilder, Morteza Rezanejad, Sven Dickinson, Kaleem Siddiqi, Allan Jepson, and Dirk B. Walther. Local contour symmetry facilitates scene categorization. *Cognition*, 182:307–317, 2019.

[95] Vladislav Ayzenberg, Yunxiao Chen, Sami R. Yousif, and Stella F. Lourenco. Skeletal representations of shape in human vision: Evidence for a pruned medial axis model. *Journal of Vision*, 19(6):6, June 2019.

[96] Morteza Rezanejad, Gabriel Downs, John Wilder, Dirk B. Walther, Allan Jepson, Sven Dickinson, and Kaleem Siddiqi. Scene categorization from contours: Medial axis based salience measures, 2018.

[97] Brian Falkenstein, Adriana Kovashka, Seong Jae Hwang, and S. Chakra Chennubhotla. Classifying nuclei shape heterogeneity in breast tumors with skeletons. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 310–323, Cham, 2020. Springer International Publishing.

[98] Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981.

[99] Jürgen Beyerer, Fernando Puente León, and Christian Frese. *Machine vision: Automated visual inspection: Theory, practice and applications*. Springer, 2015.

[100] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. *Computer vision using local binary patterns*, volume 40. Springer Science & Business Media, 2011.

[101] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[102] Dennis Dunn and William E Higgins. Optimal gabor filters for texture segmentation. *IEEE Transactions on image processing*, 4(7):947–964, 1995.

[103] Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, Oct 2000.

[104] texturesynth: A software for analyzing and synthesizing digital image of visual texture. https://github.com/labforcomputationalvision/texturesynth.

[105] Mark D. Zarella, David E. Breen, Andrei Plagov, and Fernando U. Garcia. An optimized color transformation for the analysis of digital images of hematoxylin & eosin stained slides. *Journal of pathology informatics*, 6:33–33, Jun 2015. 26167377[pmid].

[106] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *arXiv preprint arXiv:2111.04740*, 2021.

[107] Bracs: Breast carcinoma subtyping.

[108] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[109] Guillaume Jaume, Pushpak Pati, Valentin Anklin, Antonio Foncubierta, and Maria Gabrani. Histocartography: A toolkit for graph analytics in digital pathology. In *MICCAI Workshop on Computational Pathology*, pages 117–128. PMLR, 2021.

[110] Jeffrey L Fine. 21st century workflow: a proposal. *Journal of pathology informatics*, 5, 2014.

[111] Shaimaa Al-Janabi, André Huisman, and Paul J Van Diest. Digital pathology: current status and future perspectives. *Histopathology*, 61(1):1–9, 2012.

[112] Food, Drug Administration, et al. Technical performance assessment of digital pathology whole slide imaging devices. guidance for industry and food and drug administration staff. 2016.

[113] Murali Emani, Venkatram Vishwanath, Corey Adams, Michael E Papka, Rick Stevens, Laura Florescu, Sumti Jairath, William Liu, Tejas Nama, and Arvind Sujeeth. Accelerating scientific applications with sambanova reconfigurable dataflow architecture. *Computing in Science & Engineering*, 23(02):114–119, 2021.

[114] Sambanova systems.

[115] Terrell E Jones, Luong Nguyen, Akif Burak Tosun, S. Chakra Chennubhotla, Miroslawa W Jones, and Jeffrey L. Fine. Computational pathology versus manual microscopy: Comparison based on workflow simulations of breast core biopsies. *106th Annual Meeting of United States and Canadian Academy of Pathology, San Antonio, Texas, March 2017*, 2017.

[116] Aleksandar Vodovnik. Diagnostic time in digital pathology: A comparative study on 400 cases. *Journal of Pathology Informatics*, 7, 2016.

[117] Zhengchun Liu, Ahsan Ali, Peter Kenesei, Antonino Miceli, Hemant Sharma, Nicholas Schwarz, Dennis Trujillo, Hyunseung Yoo, Ryan Coffee, Naoufal Layad, et al. Bridging data center ai systems with edge computing for actionable information retrieval. In *2021 3rd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP)*, pages 15–23. IEEE, 2021.

[118] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[119] Muhammed Talo. Convolutional neural networks for multi-class histopathology image classification. *arXiv preprint arXiv:1903.10035*, 2019.