DEEP LEARNING IN DIGITAL PATHOLOGY

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

QIANGQIANG GU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

ADVISOR: DR. STEVEN N HART

CO-ADVISOR: DR. CHAD L MYERS

JULY 2023

Acknowledgments

I would like to express my heartfelt gratitude to all the members of my doctoral defense committee for their valuable time, generous support, and expertise in their respective domains. I am especially grateful to Dr. Steven N. Hart for dedicating countless hours to mentor me over the past four years in my Ph.D. training. Additionally, I would like to extend my thanks to Dr. Yuk Y. Sham, Dr. Chad L. Myers, and Dr. Timothy K. Starr for their guidance and mentorship over the past four years, as well as for agreeing to serve on my doctoral defense committee. I would also like to acknowledge Dr. Bradley J. Erickson for his contributions to my doctoral preliminary examination committee and for providing valuable feedback that has helped me to be a better scientist.

I want to express my gratitude and appreciation to the University of Minnesota Bioinformatics and Computational Biology graduate program, Mayo Clinic Graduate School of Biomedical Sciences, Mayo Clinic Department of Quantitative Health Sciences, Mayo Clinic Department of Laboratory Medicine and Pathology, and Mayo Clinic Cloud Program for allowing me to conduct my research using their excellent resources. I would also like to thank the University of Minnesota Graduate School for granting me the 2022-2023 doctoral dissertation fellowship, which has been instrumental in supporting my doctoral thesis research. I am especially grateful to Heather Domke, Lisa Gabrielson, Bobby Jadav, Glenda Mueller, Madison Nelson, Carol Prestegard from the Mayo Clinic, and Miranda Nelson from the University of Minnesota for their unwavering support and assistance throughout the past four years.

I would like to express my gratitude to Dr. Ryan Gillard from Google, Dr. Thomas J. Flotte, Dr. Jun Jiang, Mrs. Trynda N. Kroneman, Dr. Jacob G. Levernier, Dr. Chady Meroueh, and Mr. Naresh Prodduturi from the Mayo Clinic for their kind assistance and unwavering support. Additionally, I want to extend my thanks to my wonderful colleagues at Roche, including Dr. Ping-Chang Lin, Dr. Yao Nie, Mr.

Nazim Shaikh, and Dr. Xingwei Wang for their encouragement and assistance throughout the previous year.

I would like to express my gratitude to all the faculty and staff members, as well as my families and friends, for their guidance, assistance, and motivation. I am also thankful to my two faithful companions, Leena (a Pembroke Welsh Corgi) and Cisco (a Ragdoll Cat), who have been by my side throughout these years. I am deeply grateful to my grandmother, Dr. Zuomei Jiao, my mother, Mrs. Yesong Kong, and my father, Mr. Zhidong Gu, for their unconditional love and support. I would also like to remember my late grandfather, Dr. Fanlin Kong, with sincere appreciation and fond memories. I would like to express my gratitude to Dr. Willam D. Freeman, Dr. Saif D. Salman, Dr. David D. Wang, Dr. Zengri Wang, Dr. Ankush Patel, Dr. Shiman Li, Mrs. Marina Uehara, Mrs. Adriana Taborda, Ms. Nancy Konter, and Ms. Siyu Tang for their encouragement and support during the challenging time in the past year. I am sincerely thankful to everyone who has assisted me over the last four years. Their generous help and valuable insights have made the completion of my doctoral thesis research an incredible journey.

Dedication

My doctoral dissertation work is dedicated to my parents, my grandmother, and my late grandfather. Their constant support and encouragement have motivated me to be honest and contribute towards making this world a better place.

My doctoral dissertation research is dedicated to Dr. Shiman Li, Ms. Siyu Tang, and Dr. Ankush Patel, my three closest friends, who have consistently supported and motivated me during my doctoral training program.

I dedicate my doctoral dissertation work to individuals who have experienced or are currently facing health challenges. Over the past twelve years, I have personally endured numerous health issues. I hope that my journey, despite these obstacles, will never alter my identity. Instead, I aspire to inspire you to overcome your own health struggles and achieve the life you aspire to. My wish is for my doctoral dissertation work and further research to make even a small contribution towards defeating cancer and enhancing the quality of life for cancer patients worldwide.

Abstract

Digital pathology (DP), enabled by the availability of digitized whole slide images (WSIs), opens up possibilities for incorporating deep learning (DL) models into the development of computer-aided diagnostic (CAD) tools for cancer diagnostics. Among the various approaches, image classification and segmentation are widely utilized to enhance cancer diagnostics. Image classification provides slide-level predicted labels, such as tumor or non-tumor, while segmentation generates masks with x- and y-coordinates of predicted tumor areas. The scope of this dissertation research spans across multiple aspects. It involves the application of existing image classification models to differentiate between malignant breast cancer and normal breast WSIs. Additionally, a novel anomaly detection model was developed to identify anomalous tissues in melanoma WSIs. Furthermore, the developed anomaly detection model was effectively utilized for tumor segmentations in colorectal cancer (CRC).

The contributions made by this doctoral dissertation research to the field of DP primarily stem from the development of the novel progressive context encoders anomaly detection (P-CEAD) model. This model successfully detects anomalies on melanoma WSIs and demonstrates extended applications for tumor segmentation on CRC WSIs. Furthermore, significant contributions arise from the utilization of existing image classifiers in differentiating malignant breast cancer from normal breast WSIs. The research findings shed light on the significance of hyperparameter configurations and dataset variations in the process of selecting model architectures. These findings highlight that non-specialized model architectures with optimized hyperparameter configurations, have the potential to surpass DP-specialized model architectures in achieving accurate classifications on binary breast cancer WSIs.

Table of Contents

List of Tables

List of Figures

Cases

CHAPTER 1: BRIDGING THE CLINICAL-COMPUTATIONAL TRANSPARENCY GAP IN DIGITAL PATHOLOGY

Abstract

Computational pathology (CP) combines clinical pathology with computational analysis, aiming to enhance diagnostic capabilities and improve clinical productivity. However, communication barriers between pathologists and developers often hinder the full realization of this potential.

This study aims to propose a standardized framework that improves mutual understanding of clinical objectives and computational methodologies. The goal is to enhance the development and application of computer-aided diagnostic (CAD) tools.

The paper suggests pivotal roles for pathologists and computer scientists in the CAD development process. It calls for increased understanding of computational terminologies, processes, and limitations among pathologists. Similarity, it argues that computer scientists should better comprehend the true use-cases of the developed algorithms to avoid clinically meaningless metrics.

CAD tools have been shown to improve pathology practice significantly. Some tools have even received the United States Food and Drug Administration (FDA) approval. However, improved understanding of machine learning models among pathologists is essential to prevent misuse and misinterpretation. There is also a need for a more accurate representation of the algorithms' performance compared to pathologists.

A comprehensive understanding of computational and clinical paradigms is crucial for overcoming the translational gap in CP. This mutual comprehension will improve patient care through more accurate and efficient disease diagnosis.

1.1 Introduction

The roots of modern pathology, which was initially established by Rudolf Virchow, can be traced back to the 19th century, wherein gold-standard principles of laboratory medicine were forged alongside primordial numerical apertures, compensating eyepieces, apochromatic objectives, and immersion lenses.[1,2] Virchow's creed to "think microscopically" extolled both the importance of cellular insight in disease pathology and of the devices enhancing understanding of tissue morphology.[1,2] For over 150 years, this has served the pathology community well. However, the field of pathology is rapidly evolving thanks to the increasing adoption of digital pathology (DP).

DP has been made possible by the advent and adoption of whole slide imaging (WSI) and computer aided diagnostic (CAD) tools – most notably advances in machine learning (ML).[3–8] The variety of tasks enabled by ML are varied – from segmentation of cellular structures to tumor detection and grading.[9–11] Computational scientists with expertise in quantitative analysis and programming, specifically machine learning engineers, are responsible for the development of CAD tools and ML models. Their role involves creating and implementing algorithms, pipelines, and software tools to support research and clinical applications in pathology diagnostics. They employ computational and statistical methods to analyze biological data in order to enhance the field. Such has introduced a conflation of terminologies: one from the world of medicine, the other from computer science. Communication between computational scientists and pathologists is of vital importance to developing useful, accurate, and beneficial algorithms.[12] A need, therefore, exists to harmonize the vernacular of computer science and pathology.

The qualitative versus quantitative outlook held from respective pathologist and computational scientist bodies marks a key difference existing at a fundamental level. Computer vision, regardless of application, is quantitative by nature, i.e., analyzed by a mathematical framework statistically modeled for a desired outcome. On the other hand, pathologists generally rely on their training, intuition, and experience. Yet, best practices are evolving to increasingly include quantitative variables in diagnostic applications. For example, International Ki67 in Breast Cancer Workgroup scoring recommendations include visual interpretation of at least 500 malignant invasive cells (and preferably at least 1000 cells) to achieve adequate precision, despite known interindividual variation that leads to weak analytical validity.[13–15] As counting cells is considered a monotonous but critical task, pathologist attitudes have shifted toward embracing ML solutions for otherwise subjective and time-consuming workflow tasks.[7,8,16]

For the innovators, early adopters, and early majority who have synergistically compelled an equally eager cadre of computational scientists to build clinically useful tools, the difficult task of persuasion may appear solved. However, this nascent relationship has struggled to advance due to a communication gap.

When performed manually by a human reader, "quantitative" tasks (like counting) are often only "semi-quantitative". Subjective biases lead to suboptimal reproducibility, e.g., from variations in microscope configuration and field of view, non-uniform standardization, inconsistent application of scoring rules, and limited spatial coverage from visual evaluation.[14,16] Quantitative tasks are essential to diagnostic applications but require a great degree of accuracy and are often pedantic, fatiguing, and otherwise impossible to scale. Computational analysis of an image in its entirety offers to circumvent the subjectivity, inconsistency, and other limitations inherent to human investigation.[15,16]

In this article, we expand the lines of communication between computational scientists and pathologists by delving deeper into concepts that should be familiar to both parties – with examples for readers on either end of the spectrum. By meeting on grounds paved by shared language, clinicians and computational scientists may build toward synchronicity. Though barriers to implementing CADs within the clinical pathology workflow are extensively covered in literary review with many proposals for their mitigation,[10,12,16–23] none have yet proposed a framework for bridging the divided understanding between engineers and practitioners - two of the most essential elements in CAD development. With a shared vocabulary, "think microscopically" may need to be revised to "think quantitatively".

1.2 Previous Work

Emphasis has been placed on closer collaboration amongst primary drivers of the pathology CAD sphere. Asif et al. identify four primary stakeholder cohorts from clinical, academic, industrial, and patient/pubic sectors projected to spearhead implementation of guidelines and standards for the development and deployment of CP models into clinical practice.[24] Kim et al. have highlighted the importance of collaboration between clinicians and informaticians in the field of oncology, leading to technical advancements that improve clinical care.[25] With the increasing reliance on technology for cancer staging, risk stratification, and treatment, clinical informatics, including the role clinical informaticians play within medicine and the clinical laboratories, are essential in bridging the gap between physicians and technicians.

The multidisciplinary nature of clinical informatics provides a contextual framework that aids in progress in both clinical and computational forums. Programming expertise is not a requirement for clinical informaticians; however, an understanding of professional terminology, procedures, and their applications in both clinical and computational sectors is crucial for effective teamwork in the development of clinical

Artificial Intelligence (AI) and the translation of clinical AI tools. Clinical informaticians with an understanding of real-world clinical problems from real-world clinical experience have the acumen to triage problems plaguing patient care and address unmet clinical needs through informatics-based approaches.[25]

The field of digital pathology is at the forefront of the scramble for medical AI, and the exponential surge in interest is evidenced by the increasing number of publications detailing machine learning advances for diagnostic applications.[26] Steiner et al. describe recent advances in AI applications for digital pathology image analysis and factors contributing to the clinical "translation gap" for pathology CAD tools.[27]

Confusion, uncertainty, and a lack of transparency are prominent factors contributing to stymied AI development and adoption for pathology. The lack of verifiable mechanisms for interpreting machine learning rational has drawn pathologist scrutiny of the unexplainable nature of computationally deduced predictions.[22] Ethical quandaries and legal concerns stemming from the “black box” of algorithmic understandability invoke apprehension from pathologists and the litany of non-AI versed healthcare practitioners and stakeholders within the multidisciplinary milieu of laboratory medicine.[28] The nascency of AI penetration for clinical laboratory diagnostics in comparison to similar diagnostic disciplines, e.g., radiology, leaves many pathologists with little CAD experience and little understanding of appropriate use, strengths and weakness, and interpretation of machine-output predictions.[29] The impact of such inexperience has led to insufficient contemplation of the essential factors for appropriately developing AI for fundamental diagnostic use-cases and results in clinically non-deployable solutions.[29–32] Although suggestions for greater CAD awareness and pathologist interaction with AI tools is suggested, little is typically mentioned in lieu of charting a process of execution extending beyond mere identification of barriers curtailing the implementation of AI-based CADs in the clinical workflow.[29]

The literature to date provides extensive coverage of CADs for pathology from both computer engineer and pathologist perspectives, for which problems are identified and solutions proposed (**Table 1. 1**).[12,20–22,33] Fostering communication between pathologists and computational biologists is integral to CAD development though often relegated to an afterthought within a sea of discourse in each isolated sector. We aim to extrapolate upon the suggestion of shared communication through offering a structured and actionable solution. By facilitating the construction of a bridge between pathologist and computational biologist thought and communication, we hope to overcome the largest obstacle preventing clinical CAD translation in digital pathology.

Table 1. 1 *Challenges facing clinical deployment of CAD solutions for CP.*[11,15,17–25]

Challenge	Limiting Factors	Proposals and Implementations
Standardization	• Variations in data source, file format, and AI modeling methodology predispose to variations in CAD analytics, e.g., classification, with accompanying concerns of output validity. • Community adoption of conversion tools generating the Digital Imaging and Communications in Medicine (DICOM)-compliant files is limited. • Lack of high-performance software libraries capable of supporting intensive data processing formats for advanced machine learning applications, e.g., training of a convolutional neural network (CNN), requiring large volumes of WSI data. • Lack of unified annotation protocol and standardization for metadata storage and imaging data.	• Adoption of an open-source file format based on known, established standards, e.g., DICOM standard, supporting WSI storage and exchange in the Picture Archiving and Communication System (PACS), that is supported by both open-source and commercial software, e.g., Tagged Image File Format (TIFF), and capable of supporting multi-channel, multiplexed, mass spectrometry, extended depth of focus (“Z-stacked”), and structured annotation data, e.g., Open Microscopy Environment (OME)-TIFF. • Development of a regularly updated interoperable software that is supported across platforms and used for many different use-cases and applications.
Availability	• Lack of pathologist annotated WSI dataset is compounded by limited specialist availability, limited	• User-friendly, easily accessible, e.g., internet-cloud based, open-source annotation software with

	dedicated time for annotating, and limited data accessibility from data privacy and proprietary sourcing. ● Clinical deployment of algorithms requires well-annotated datasets to assure strong validation and interpretability, however pathologists without programming experience or familiarity with command line interface for software operations may be deterred by WSI segmentation software that is complex to deploy and use.	easy-to-use graphical user interface enabling pathologists without programming experience to segment Whole Slide Images (WSIs) for algorithm training. ● Easily searchable open-source WSI databases housing quality-controlled data and AI algorithms, e.g., the Cooperative Human Tissue Network, BIGPICTURE proposition (Innovative Medicines Initiative Call 18, Europe). ● Accurate assessment of workload distribution for annotations incorporating consideration of various levels, e.g., cases, regions, cells, reports, and details, e.g., level of exhaustiveness, in conjunction with pathologist experience, clinical time constraints, and daily work commitments. ● Further development of existing techniques for mitigating (though not yet substitutable for) data shortages and time-consuming annotation processes, e.g., transfer learning and data augmentation.
Cost	● Labeled WSI data is expensive to produce. ● WSI data storage costs compound upon production expenses. ● Graphical Processing Unit (GPU) clusters are required for training and deploying deep learning (DL) CADs in practice, however, are financially limiting barriers for many pathology laboratories.	● Unsupervised learning techniques for training algorithms, which do not require labeled data (however lack explainability). ● Greater accessibility of available training data through use of open-source data unrestricted by data privacy and proprietary limitations.
Quality	● Reliable quantitative analysis is dependent upon WSI quality that is subject to variations in pre-analytical preparation, e.g., tissue sectioning, staining, and scanning. ● High-resolution image reduction techniques used in DL development	● Use of an automated image quality analysis pipeline, e.g., HistoQC, PathProfiler, for identifying image artifact and other slide scanning errors warranting rescanning prior to further processing and

	may compromise training quality, e.g., higher-level structural information such as tumor shape or extent may only be capturable through analysis of larger WSI regions.	computational analysis. ● Regular quality review during the annotation process for error identification and improvement in annotation quality. ● Automated quality control for annotations with metrics such as exhaustiveness, diversity, and concordance for further analysis of regions that may be prioritized based on the current cell count or area annotated. ● Normalization techniques such as scale normalization may mitigate variations in pixel sizes and differences in WSI scanning devices. ● Additional data normalization techniques include those for stain normalization, flexible thresholding for variations in data luminance, focus spatial correlation and multi-level magnification for WSI patches.
Security	● Cloud-based storage of WSI data is susceptible to network delays and unreliable compliance with the Health Insurance Portability and Accountability Act (HIPAA) protocols for secure safeguarding of uploaded patient data from unauthorized personnel. ● Adversarial attacks capable of misleading a robust ML network may result from targeted manipulation of only a small number of pixels within an image. ● Though adversarial attacks on CP CADs are only hypothesized, demonstration of vulnerability to targeted attacks has lent to concerns of algorithmic contamination from presence of small artifacts or minimal noise.	● Construction of local AI resources for WSI upload security.
Transparen	● Clinical, legal, and regulatory	● “Rule extraction” techniques

cy	obstacles arise from the "black box" nature of AI decision-making processes, emphasized in DL-based systems for which perceptions regarding interpretability are marred by uncertainty. Such has led to regulatory restrictions, e.g., the General Data Protection Regulation of Europe, and liability concerns from pathologists contemplating the medical implications of inexplicable AI deductions.	enabling easier algorithmic interpretability through providing information revelatory of previously hidden algorithmic segmentation processes, e.g., sectioning algorithmic deduction into a stepwise process in which histopathologic features of algorithmic focus are displayed to facilitate human understanding of algorithmic "thought".

1.3 Computer Vision Application in Digital Pathology

1.3.1 Image Classification

1.3.1.1 Definition

Computational image classification is guided by specific rules[34] for tissue categorization that aid predictive labeling of specific pixel-groups within an image, e.g., benign vs. cancerous tissue.

Of critical importance is the definition of the label. A label is any metadata that one wants the model to associate the input data with. The label could be an attribute about a patient, a slide, a region, or individual pixels that one wants to derive from input data. The type and quality of the label can have a profound impact on the type and meaning of model outputs (**Figure 1. 1** and discussed below).

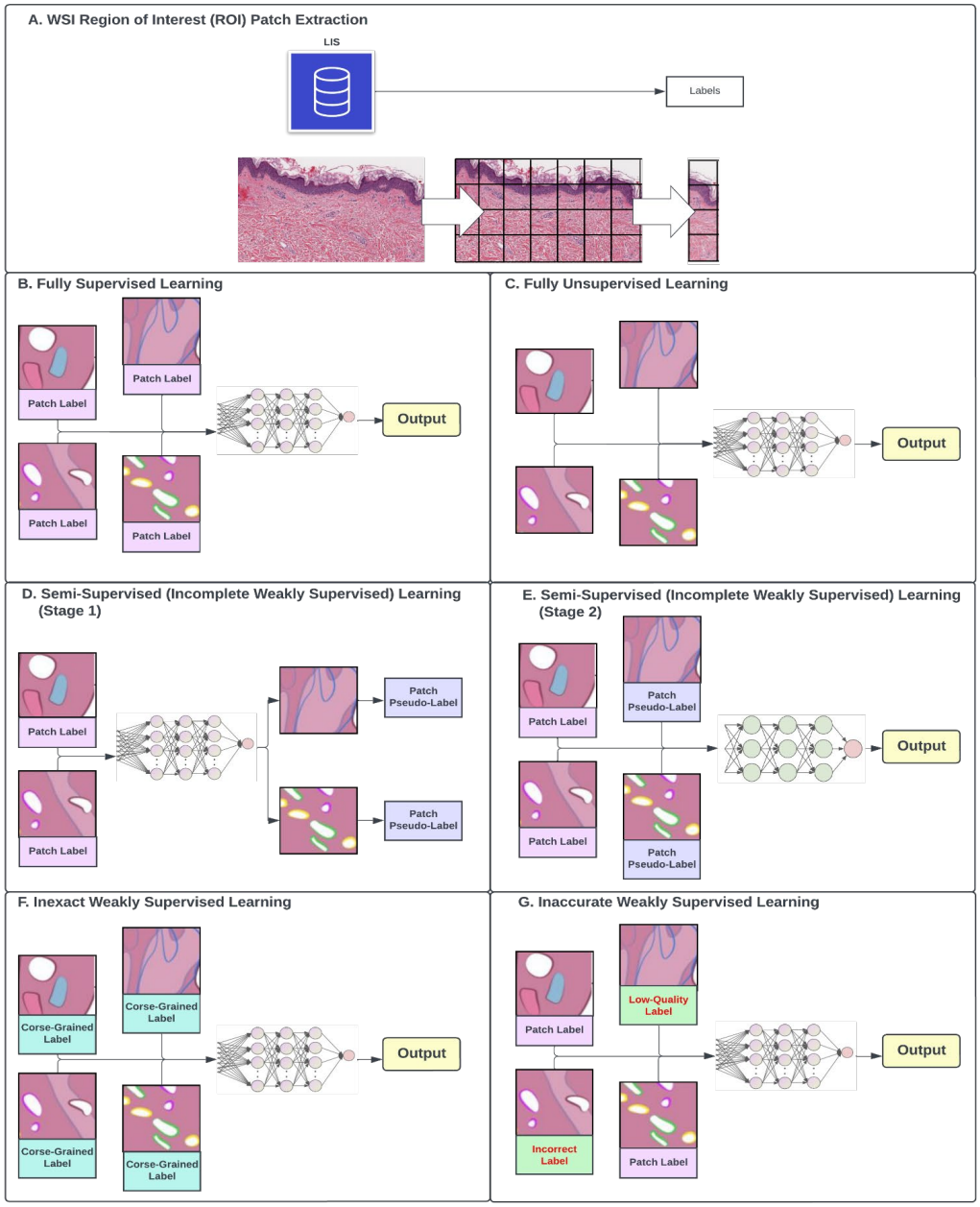


Figure 1. 1 *Model training methods. A). WSI region of interest (ROI) patch extraction; B). Fully supervised learning; C). Fully unsupervised learning; D). Stage one of semi-supervised (incomplete weakly supervised) learning; E). Stage two of semi-supervised (incomplete weakly supervised) learning; F). Inexact weakly supervised learning; G). Inaccurate weakly supervised learning.*

1.3.1.2 Clinical Use Cases of Image Classification

Clinical approach to detection of ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC) of the breast may entail pathologist analysis of a lesion for the presence of either entity. With this procedure in mind, a computational biologist may derive a binary classification system for the detection of features indicative of either DCIS or IDC. As clinical reality indicates incidence rates of DCIS coexisting with IDC in up to 76.9% of all DCIS cases examined in the literature,[35,36] the binary framework of the algorithm will be fundamentally flawed in its ability to detect dual presentations that may warrant unique treatment schemes.

Another common failure point is the use of homogeneous data. For example, assume the breast cancer patient has thirty slides in a study, but only one diagnosis. ML algorithms could simply learn to predict if a slide belongs to the patient, rather than predicting the presence or absence of relevant features. The model's predictive output has limited clinical merit due to a lack of generalizability, or applicability to a wide patient population, while demonstrating misleading accuracy and precision. Therefore, pathologists must communicate the nature of provisioned data with computational scientists, emphasizing data diversity and its relevance for an intended clinical problem throughout model training and testing. This communication is crucial to ensuring that the image classifier provides clinically useful insights for a broad patient population.

1.3.1.3 Considerations in Model Construction for Image Classification

We propose a framework to help guide pathologist understanding of computational procedures for image classifier development while allowing computational scientists ease of reciprocity in communicating their concerns and requests to pathologists at various stages to ensure algorithms meet clinical objectives.

Our framework is based on essential questions and considerations that arise during computer scientist development of image classification algorithms. These include the level of supervised learning to use during training, the size and diversity of training data, "balanced" vs. "imbalanced" classification, designing a validation experiment for an image classifier, and evaluating model performance. By addressing these questions and considerations, pathologists and computational scientists can work together more effectively in developing clinically relevant image classifiers.

1.3.1.3.1 Level of Supervision

There are two primary levels of supervision when building models: supervised and unsupervised. Supervised learning means that labels are provided, whereas unsupervised simply uses patterns in the data to form natural groupings. Supervised learning can be further broken-down into specialized types including weakly-supervised, semi-supervised, and self-supervised (**Figure 1. 1**).

Fully Supervised & Fully Unsupervised Learning

Labels are requisite for all images used for fully supervised model training. “Slide-level” (or image-level) labels present an overview diagnosis of an entire WSI. “Pixel-level” labels are derived from extensive annotations at the pixel-level, whereby multiple tissue types and/or pathophysiology may be identified on a single WSI. Annotations characterized as “strong” in the literature typically refer to those at the pixel-level.[37–39] Image patches, grids with certain width and height that overlay a WSI, are computationally extracted from a single WSI to create multiple training images (**Figure 1. 1 A**).

An example of fully supervised learning is an algorithm that is fed a series of patch images with corresponding patch-level labels, e.g., tumor vs. benign (**Figure 1. 1 B**).[40–46] Fully supervised learning

requires that training images and their labels must be paired using the same scale, e.g., patch-level labels must correspond with image patches and slide-level labels must correspond with entire WSIs. Conversely, fully unsupervised learning involves model training with images absent of labels (**Figure 1. 1 C**).[47–50] Data is not assigned a class in unsupervised training but is instead "classified" using another measurement of similarity (e.g., principal component analysis, t-stochastic neighbor embedding, uniform manifold approximation and projection, etc.).

Weakly Supervised Learning

Weakly supervised learning is an overarching term characterizing a subset of supervised learning techniques distinguished by noisy or vague label associations for training image data. Three subtypes of weakly supervised learning[51,52] include incomplete supervision (whereby partially labeled images are provided) (**Figure 1. 1 D - 1. 1 E**), inexact supervision (whereby coarse-grained labels, i.e., those which broadly apply to an entire image, set of images, or image portion in lieu of more nuanced or detailed annotations, are provided) (**Figure 1. 1 F**), or inaccurate supervision (whereby "noisy" processes contributing to low-quality- or incorrect labels, e.g., from the machine learning framework or human behavior, are provided) (**Figure 1. 1 G**).[37,53,54] Inconsistency lingers throughout literature definitions of weak supervision. For example, weak supervision may solely be characterized as "inexact supervision", whereby image pairs and their corresponding coarse-grained, e.g., "slide-level", labels are used for training.[37,53,54]

As detailed annotations for entire large datasets are often tedious, time-consuming, and expensive, they are thereby infeasible to acquire for many histology tasks.[55] Incomplete learning may be utilized to mitigate for cost or labor shortages in such circumstances.

Intra- and inter-observer variability from subjective pathologist examinations poses challenges for the creation of ground truth annotations which may prompt specialists, i.e., "expert", opinion for resolution. Ideally, consensus scoring from enough experts is of importance for ensuring accurate model reproducibility.[56] Additionally, inclusion of discordant cases in model training is of importance for assuring unbiased models that are robust and capable of flagging the most challenging cases. The absence of expert opinion in instances of non-concordance increases the risk of inaccurate labeling and therefore inaccurate training, i.e., inaccurate supervision. High inter-observer variability is observed in pathologist assessment of Gleason grade in prostate cancer WSIs (used to calculate Gleason score, the strongest prognostic predictor of prostate cancer).[57] Gleason grading algorithms are typically complex and require extensive region-level manual annotations by experts, who themselves often fall to interobserver disagreement for particularly challenging histopathological presentations of prostate cancer that require Gleason grading, e.g., poorly formed glands.[10] To mitigate burdensome and ostensibly unnecessary requirements for detailed pixel-level annotations, slide-level annotations may instead be utilized in inexact training schemes for algorithms purposed to learn differentiating features of Gleason patterns that are then used for predicting corresponding grade groups.[57]

Semi-Supervised Learning

Semi-supervised learning (or incomplete supervision) is a subset of weakly-supervised learning[52] that may be distinguished from other weakly-supervised subtypes from its focus on propagating that which is "already established", i.e., using partially labeled and unlabeled image portions for learning, as opposed to a multi-step process that trains an image classification model using both annotated and unlabeled data from the same dataset. The training process begins with a supervised approach using the annotated proportion of dataset images. After supervised training, the "base" model is used to classify the remaining unlabeled images in the dataset. The most confident predictions are selected as "pseudo-labels," and

pseudo-labelled WSIs are used in conjunction with annotated WSIs to further refine model training. Semi-supervised approaches rely on shared classification between annotated and unlabeled training data.[58–60] Semi-supervised learning is ideal when limited annotated data is available, such as in rare disease cases,[61] instances requiring unique datasets or completely annotated large datasets for clinical problem modeling, and time constraints limiting pathologist domain expertise.[62]

Self-Supervised Learning

Self-supervised learning is characterized by machine generation of labels for unknown (or hidden) input regions from predictions derived from known input regions.[63–65] Self-supervised learning does not require domain expertise for labels.[63–66] In contrast to semi-supervised learning, annotations derived for self-supervised learning arrive from model-prediction rather than human deliberation.[66] Therefore, self-supervised learning is ideal when pathologist annotations are unattainable.

Self-supervised learning begins with an unsupervised approach to “pseudo-classify” unlabeled WSIs. The model is then trained in a supervised fashion using the WSIs and their corresponding auto-generated pseudo-labels from the dataset. Without the pathologist, the computer scientist will know that these labels have some meaning to their associated pattern but will not have a human interpretable label that describes what the class contains.[67–69]

1.3.1.3.2 Size and Diversity of Training Dataset

The size and diversity of a training dataset are crucial factors that impact algorithm performance. A small or homogeneous training dataset may lead to subpar model performance, i.e., "underfitting," as the model may not detect all characteristics featured in the training data. On the other hand, insufficient training data may also result in model "overfitting," demonstrated by model performance that meets or exceeds

expectations when classifying images from the training dataset but has poor generalizability to new data used for testing. Both examples emphasize the importance of a large and diverse training dataset for achieving optimal model performance and generalizability.

A student-T-test[70] can be adopted by computational pathologists to determine the optimal size of a training dataset. Established on the assumption (null hypothesis) that no significant difference exists in model performance following training with current vs. proposed dataset sample sizes, the student T-test can be used to determine whether WSI addition will lead to significant improvements in model performance. (**Table 1. 2**).

Table 1. 2 *Using a student T-test to determine optimal training dataset size.*

H_0 **(null hypothesis):**	**Outcomes and corresponding implications:**
Identical image classification performance following model training with current and proposed training sample sizes.	If P-value is ≤ 0.05 and the model trained with proposed sample size yields superior performance to the current sample size, then the proposed training size is suggested for use.
	If P-value is > 0.05 and the model trained with the current sample size did not achieve target performance, then more samples are suggested for incorporation within the training set.

While adding WSIs to increase the size of a training dataset can address some overfitting problems, model overfitting may persist due to training data homogeneity. A more diverse dataset containing WSIs from a range of patient populations is ideal. A heterogeneous dataset can help to ensure that the image classifier can generalize to a greater diversity of problem domain data, improving its clinical utility and generalizability.

1.3.1.3.3 Imbalanced Datasets

Balanced datasets are ideal for image classifier training as they represent a specific problem domain with an equal number of images from each class. However, imbalanced datasets with unequal sample classes are more common in practice. Imbalanced datasets, for example 80% tumor versus 20% normal, used for model development may result in model overfitting, leading to good predictive output for tumor samples but poor predictive output for normal samples. The model could simply decide to predict everything tumor, and it would be accurate 80% of the time. This is a dangerous outcome – especially if 80 / 20% reflects the actual balance seen in clinical workflows as it might go unnoticed for a while. Adding additional WSIs containing the minority class is a simple approach to balancing datasets, but it is often hindered by the unavailability of data from the non-dominant class(es) given the time and expense required for collecting additional data. To mitigate imbalanced classification in instances where supplementing additional WSIs is infeasible, two alternative solutions are common: 1) image augmentation of training data and 2) ensemble learning.

Image Augmentation

Image augmentation is purposed to promote feature "invariance" within training data to reduce the likelihood of model fixation on WSI features such as color or artifact that may differ in data samples created with differing procurement processes, e.g., type of WSI scanner used. Image augmentation may be used to increase the effective sample size in order to achieve data balance, however, is not an equal substitute for the addition of independent samples. Image augmentation methods include cropping, shifting, color augmentation, kernel filtration (e.g., sharpness, blurring), rotation, and flipping (**Figure 1. 2**). Image cropping involves the removal of specific regions of a WSI, resulting in a sub-sectioned WSI area with smaller pixel dimensions and file size. Image shifting moves each pixel of a WSI to a different position. Image rotation involves rotating a WSI in varying degrees of clockwise or counterclockwise

direction, while image flipping mirrors a WSI across a horizontal axis. These methods can be used individually or in combination to increase the diversity of samples in imbalanced datasets, improving the performance and generalizability of image classifiers.

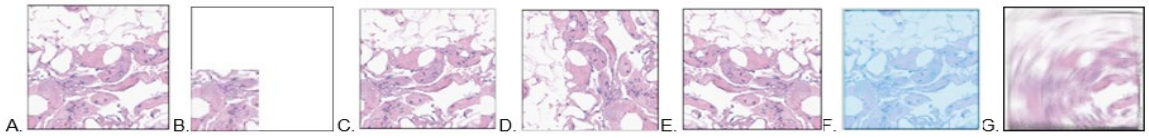


Figure 1. 2 *Image augmentation techniques. A). Original image; B). Cropped image; C). Image shifting (left-sided); D). Image rotation (270 degrees); E). Image flipping (mirror-reversal); F). Image color augmentation; G). Image blurring.*

Ensemble Learning

Ensemble learning[71] improves predictive classifier performance at the model-level through combining predictions from multiple "weaker" models to create a stronger overall model.[72] Benefits of the technique are reaped from the creation of weaker models that are substantially dissimilar to one another, thereby imparting significant training variability to create a stronger overall model structure. This may be contrasted with less effective techniques involving repetitive retraining of a single model with only minor changes afforded through each turn. Bagging,[73] stacking,[74] and boosting[75] are the three primary methods of ensemble learning.

Bagging is a model-level ensemble learning approach that uses smaller portions of a composite WSI dataset to train a number of "weak" classification models independently, in parallel. These weak models are trained with smaller amounts of data rather than the entire dataset, introducing diversity among weak learners. The predictions from these weak models are aggregated for a final output, potentially improving the overall accuracy and robustness of the image classifier (**Figure 1. 3 A**). Bagging methods may be useful for situations limited by development time and computational processing availability.[76,77]

Stacking is another model-level ensemble learning approach that employs different sets of input images from a clinical problem dataset to train multiple image classification models. The input data used for model training and the predictive outputs of all intermediate models are then used to train a final classification model with potentially improved accuracy and robustness (**Figure 1. 3 B**). However, stacking is a more complex approach that may require greater time and resources compared to other ensemble learning methods.[78,79]

Boosting is another model-level ensemble learning approach that uses weak models trained sequentially rather than independently. Each new model is trained to correctly classify the misclassified predictions of its predecessor, thereby increasing in strength, i.e., predictive capacity, as boosting progresses. The final classification model rendered is the product of the weighted sum of all sequentially trained weak learners, resulting in a more accurate and robust image classifier (**Figure 1. 3 C**). Of the three ensemble learning methods, boosting is the most ideal for optimizing classifier predictive value due to its sequential approach enabling the progressive development of its weaker models.[80,81]

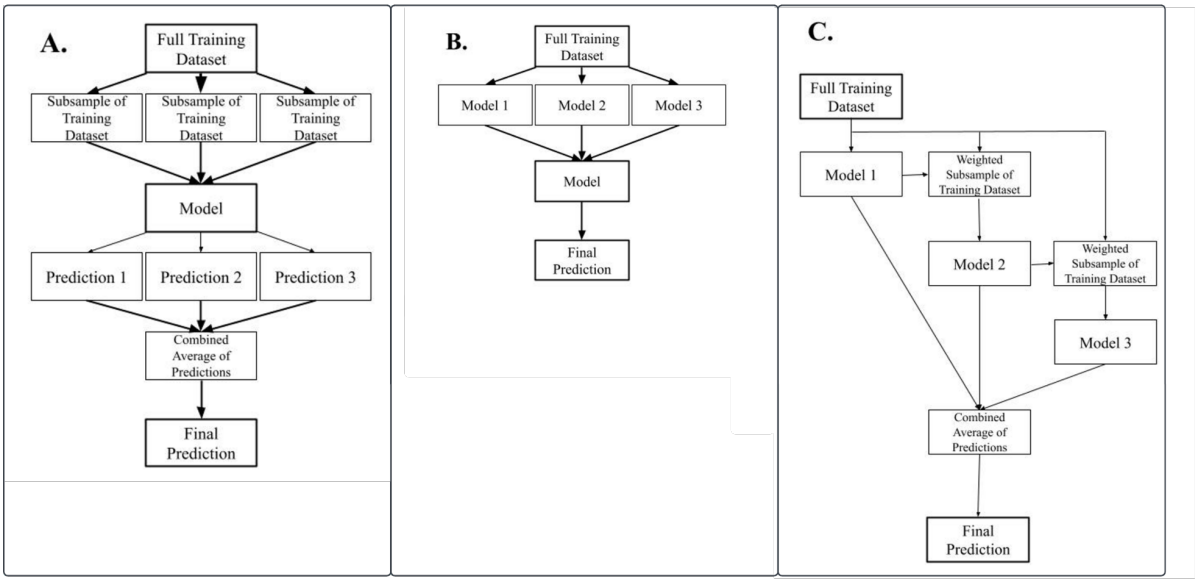


Figure 1. 3 *Ensemble learning. A). Bagging; B). Stacking; C). Boosting.*

1.3.1.3.4 Validation Experimental Design

The validation process is a crucial step for ensuring the credibility of the predictions generated by ML algorithms. While model training focuses on fitting the algorithm to a specific clinical objective, validation aims to estimate the prediction error of the model. The accuracy of the model is evaluated by comparing the predicted output labels with the gold standard data. On the other hand, model testing is concerned with evaluating the generalizability of the model on unseen data. Overfitting can occur if a model performs well on the training data but fails to generalize to the testing dataset and may sometimes be detected during model training using the validation/tuning set.

The process of splitting WSIs for training and validation is a common approach to model validation. A certain ratio of all available WSIs is divided into those for training and model tuning, i.e., ML "validation" (it is of importance to distinguish model tuning / ML validation from clinical validation of an ML model as the terminologies reflect entirely different processes), with the remainder for testing. The ratio of data used for each set depends on the size of the available data. For example, a 70/30 split may be applied, with 70% of WSIs used for training and tuning and the remaining 30% for testing. The model is then evaluated using the testing set.

Training with a heterogenous dataset provides a holistic composite of WSIs representing a diverse demographic of patients from multiple institutions offering digitized images from a variety of different scanners. Dataset heterogeneity leverages the potential for improved model generalizability and is of important consideration in data splitting. Consideration of fair distribution of WSIs from different institutions, scanners, and patients within training, validation, and testing datasets is of importance when seeking to apply a model to multiple laboratories with support for different WSI scanners. However,

generalizing to other institutions adds in complexity, making data sourcing more difficult and may not always be relevant to how a model is applied in practice.

K-Fold cross validation is a powerful and widely used model validation approach that can improve model generalizability to new data, especially for smaller WSI datasets. In this process, WSIs for model development are randomly shuffled and divided into a specified number of groups. One group is then used as the testing dataset, while the other groups are combined for model training. This process is repeated with a different group designated for testing until all groups have been used for testing. Performance metrics from each testing cycle are then combined and confidence intervals generated to determine overall model performance.

1.3.1.3.5 Model Performance Evaluation

The final step in developing an image classifier is the evaluation of model performance, which is critical for assessing the classifier's validity and utility. There are nine commonly used statistical metrics for evaluating model performance, including confusion matrix, sensitivity, specificity, precision, accuracy, balanced accuracy, area under the receiver operating characteristic (ROC) curve (ROC-AUC), area under the precision-recall (PR) curve (AUC-PR), and F1-score. Notably, these metrics tend to distill a model's performance into a single value which provides a concise depiction of a model's performance but may fail to capture nuances that are important for evaluating the behavior of a model in a "real world" setting. Other aspects to consider include classification or prediction certainty, confidence, and performance on subsets of cases that are particularly relevant to a use case. It is important to consider that no individual metric entirely accounts for model performance and the appropriateness of a metric is constrained by clinical context.

Confusion Matrix

Confusion matrix comprises true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values calculated per pathologist ground truth labels and labels predicted by a classification model. The confusion matrix is simply the representation of FP, FN, TN, and TP results from which all of the aforementioned metrics are calculated. Multiple-classification scenarios require an increase in dimension and complexity of a 2 x 2 binary confusion matrix (**Figure 1. 4 A**) to a N x N matrix (**Figure 1. 4 B**) where N equals to the number of classes with TP, FP, TN, and FN from each of the N respective classes. TP values of each of the N respective classes in the multi-class classification confusion matrix are identifiable by their characteristic diagonal pattern of distribution leading from the uppermost left-lateral cell to the bottommost right-lateral compartment. Performance bias can also more easily be visualized using a confusion matrix.

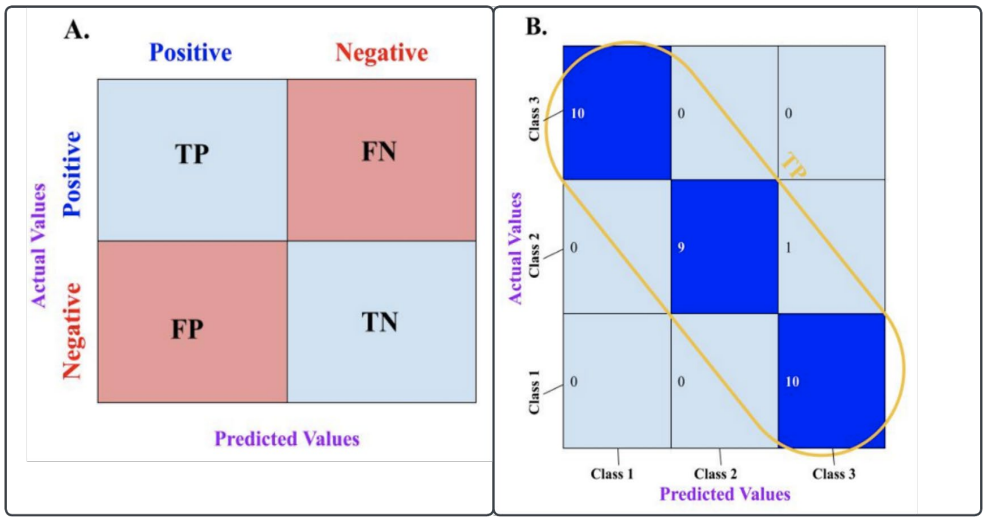


Figure 1. 4 Confusion matrix. A). Binary confusion matrix; B). Multi-class confusion matrix.

Sensitivity

Sensitivity, also known as recall, hit rate, or true positive rate (TPR), is a measure of the proportion of positive cases that are correctly identified as positive by a classification model. Maximizing sensitivity

reduces the chances of FN predictions but also reduces the true negative rate (TNR). In imbalanced classification scenarios where the target class is the non-dominant class, maximizing sensitivity may result in false positive predictions for negative cases (**Table 1. 3**).

CADs for frozen section analysis, i.e., rapid histological analysis on a mass during surgery, are in nascent development for assisting with immediate surgical consultation[82,83], a scenario whereby false positive predictions may levy substantially catastrophic outcomes. Women with suspicion for early-stage ovarian cancer require surgical staging that can be rapidly executed through use of frozen section techniques.[84] Though quickly rendered, the crude nature of frozen section samples lends them to difficult interpretability in comparison to those sectioned and embedded in paraffin (the most frequently used method for exhibiting well preserved morphology[85] and commonly selected for WSI transformation[8]). The precarious nature of frozen section interpretability (for both practitioners and AI) levies an increased risk for over-staging women without malignancy. Risk may be further compounded by CADs calibrated for maximal sensitivity that mistakenly confirm a false-positive diagnosis rendered by a surgical pathologist, leading to surgical over-treatment and patient complications ranging from loss of fertility to mortality.

Specificity

Specificity is also used to measure the ability of a classification model to correctly identify TN cases as negatives, which is particularly important in medical diagnoses where a false positive result can lead to unnecessary medical procedures or treatments. However, maximizing specificity should be balanced with other performance metrics, such as sensitivity and overall accuracy, to ensure that the model performs well for all classes in the dataset (**Table 1. 3**).

A CAD calibrated for maximal specificity during frozen section analysis may avoid precarious over-staging and surgery during surgical investigation of ovarian cancer presence. However, maximal specificity predisposes to increased risk for false-negative diagnosis and may lead to the missed detection of ovarian cancer that is present in a patient.

Precision

Precision is a statistical metric used to evaluate the performance of a classification model. It measures the proportion of true positive cases out of all cases that are positively predicted by the model. In other words, precision reflects the likelihood that a predicted positive case is truly positive. Precision is an important metric when the focus is on minimizing false positive predictions. However, it does not consider true negative values and may not be suitable for classifiers intended to identify negative samples (**Table 1. 3**).

A CAD calibrated for maximal precision is clinically irrelevant for scenarios whereby confirmation of lesion presence or absence is vital.

Table 1. 3 *Sensitivity, Specificity, and Precision.*

	Target Lesion Present	**Target Lesion Absent**
Positive Identification	TP	FP
Negative Identification	FN	TN
Sensitivity	$TP\ /\ (TP\ + FN)$	
Specificity	$TN\ /\ (TP + FP)$	
Precision	$TP\ /\ (TP + FP)$	

Accuracy & Balanced Accuracy

Accuracy reflects the likelihood that a model will correctly classify an image based on the proportion of images it has correctly classified in relation to its total predictions. Accuracy is an ideal statistical metric for evaluation of models formulated from balanced classification scenarios. Balanced accuracy, however, is calculated by taking the mean of sensitivity and specificity, making it a more reliable metric for evaluating models trained on imbalanced datasets. It reflects the overall accuracy of the model while considering the predictive performance on both dominant and non-dominant classes. However, it is important to note that even with balanced accuracy, it is still possible for a model to perform well on dominant classes and poorly on non-dominant classes, highlighting the importance of other evaluation metrics such as precision and recall.

Area Under the Receiver Operating Characteristic Curve

The ROC-AUC is a widely used metric for evaluating classifier performance (**Figure 1. 5 A**). It measures the performance of a classifier across different thresholds and is based on the receiver operating characteristic curve, which visualizes the classifier performance with sensitivity on the y-axis and false positive rate (FPR) on the x-axis. The ROC-AUC is the area under the ROC curve and is threshold-invariant, meaning it considers both sensitivity and specificity. However, maximizing ROC-AUC can lead to an imbalance between true positive and true negative predictions, resulting in more false positives when the threshold is lowered. As such, the ROC-AUC has limited interpretability for clinical use cases where accurate predictions on both positive and negative cases are important.[86]

Area Under Precision-Recall Curve

Like ROC-AUC, AUC-PR, also measures the classifier performance across multiple thresholds (**Figure 1. 5 B**). Different from ROC, the PR curve keeps sensitivity in the horizontal axis, and precision in the

vertical axis. Maximizing AUC-PR will reduce the number of positive cases that are missed by a classifier, also could minimize the risk by calling a negative sample as positive. Therefore, compared to ROC-AUC, AUC-PR is more appropriately applied to imbalanced classification scenarios, since ROC-AUC could be misleading by sacrificing more negative predictions.[87] F1-score, also known as traditional F-measure, balanced F-measure, Sorensen-dice coefficient, or dice similarity coefficient, is defined as the harmonic mean of precision and sensitivity. Since F1-score is determined by both precision and sensitivity, it shares the same strength and weakness of precision and sensitivity as these metrics.

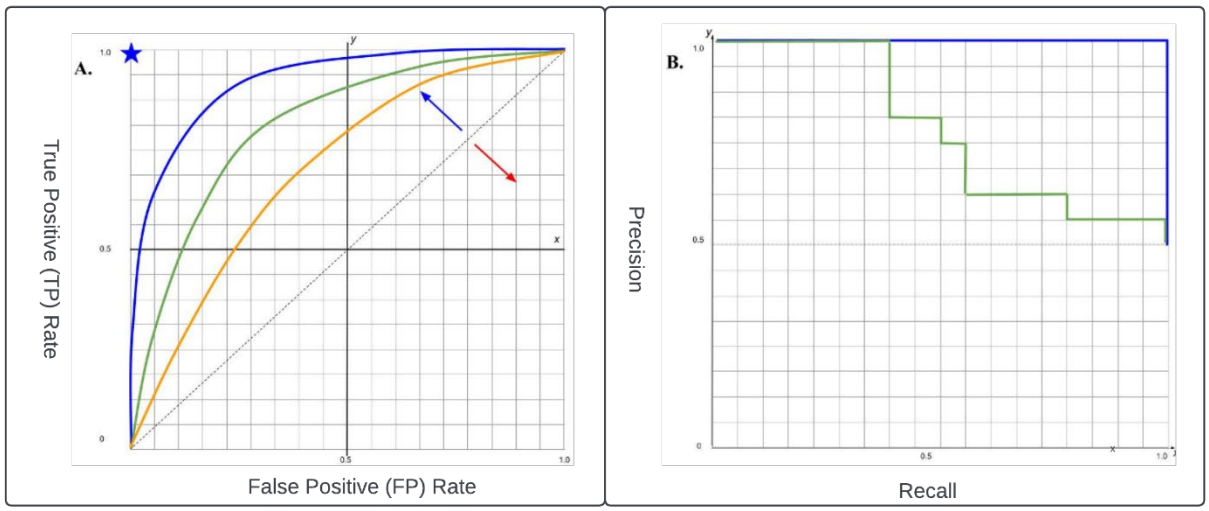


Figure 1. 5 *Diagrams of ROC-AUC and AUC-PR. A). ROC-AUC, with true positive rate on the y-axis and false positive rate on the x-axis; B). AUC-PR, with precision on the y-axis and recall on the x-axis.*

1.3.2 Likelihood Measurement

1.3.2.1 Definition

Likelihood measurement provides a quantitative representation of the confidence associated with a model's predictive classification or outcome.

1.3.2.2 Clinical Use Cases of Likelihood Measurement

The evaluation of disease progression after drug therapy is a common clinical concern. To translate this into a computer vision task, the clinical question is divided into two parts: 1) classification of lung adenocarcinoma into the epidermal growth factor receptor (*EGFR*)-mutated and non-*EGFR* mutated and 2) determining the likelihood of drug resistance in patients with *EGFR* mutation status. Molecular testing for *EGFR* mutation status is typically performed in lung adenocarcinoma patients, and pathologists must label associated WSIs accordingly to serve as ground truth. Using these labels in combination with corresponding WSIs, a classifier may be trained to predict *EGFR* mutation status. In cases where drug response is unknown, a computational model may be developed using ground truth provided by pathologists to train an algorithm to accurately predict drug resistance likelihood based on *EGFR* mutation status.

1.3.2.3 Considerations in Model Construction for Likelihood Measurement

As likelihood measurement is reliant on accurate image classification, considerations for this task remain the same as those for image classification. This is because accurate classification is a prerequisite for accurate likelihood measurement, and any errors in classification will propagate to likelihood measurement results. For more detailed information, please refer to **section 1.3.1.3.1 – 1.3.1.3.5**. These sections provide a comprehensive overview of the key challenges involved in developing accurate and reliable computer vision models for both image classification and likelihood measurement and offer valuable insights into potential solutions to these challenges.

1.3.3 Object Localization

1.3.3.1 Definition

An image localization model is a type of computer vision model that is designed to identify the coordinates of objects of interest within an image by applying bounding boxes around the objects. This can be particularly useful in pathology, where pathologists may need to analyze specific regions of tissue within a larger image. The model is trained using labeled images where the coordinates of the objects of interest are known. Once trained, the model can automatically identify the objects and their locations within new images. This can save pathologists a significant amount of time and improve the accuracy of their analyses.

1.3.3.2 Clinical Use Cases of Object Localization

Computational image localization can aid pathologists in identifying target objects of interest in WSIs. One example of this is the identification of mitotic figures in breast cancer, which can be transformed into a computer vision task.[88] However, accurate identification of mitosis in breast cancer WSIs is challenging and is plagued with low concordance rates,[89] while demanding significant time commitments from pathologists for provision of ground truth annotations. An assurance of high-quality annotations may be upheld through multi-rater collaboration given a number of raters sufficient to ensure maintenance of appropriate statistical power. Though burdensome, certain circumstances may require the recruitment of additional raters to ensure high-quality annotations and maintenance of appropriate statistical power.

1.3.3.3 Considerations in Model Construction for Object Localization

The successful implementation of image localization models in the field of digital pathology requires close collaboration between computational scientists and pathologists. Key challenges that need to be addressed in this context include model use case, how model data will be presented visually, i.e.,

visualization, determining the appropriate level of supervision, selecting an appropriately sized and diverse training dataset, dealing with imbalanced datasets, designing robust validation experiments, and evaluating model performance. A thorough examination of potential solutions to these challenges is provided earlier (**section 1.3.1.3.1 – 1.3.1.3.5**). By working together, computational scientists and pathologists can leverage their respective expertise to overcome these obstacles and develop accurate and reliable image localization models for use in clinical practice.

1.3.4 Object Counting

1.3.4.1 Definition

Object counting models are a crucial component of computer vision systems designed to accurately quantify the number of target objects present in an image. When developing such models, it is important to first develop an image localization model, which is responsible for identifying the locations of objects within the image. Once this is achieved, the object counting model can be trained using the localized object information to accurately determine the exact number of objects of interest present in the image sample. This sequential approach is critical to achieving precise and accurate object quantification in computer vision systems and has important implications for a wide range of biomedical applications, from cancer diagnosis to drug discovery.

1.3.4.2 Clinical Use Cases of Object Counting

The quantification of metastatic lymph nodes by pathologists is a common clinical task that can be significantly aided by computer vision. To achieve this goal, the clinical counting problem must be divided into two computational tasks. The first step is to develop an image localization model capable of accurately detecting metastatic lymph nodes, which requires the inclusion of pathologist ground-truth annotations of metastatic lymph nodes in the training dataset of WSIs. The second step is for pathologists

to manually quantify the true counts of metastatic lymph nodes in the testing dataset. To ensure the accuracy of ground-truth labels in both the training and testing datasets, all annotations must be reviewed by pathologists who are not involved in the annotation process. By following this two-step approach, it is possible to achieve highly accurate and reliable quantification of metastatic lymph nodes using computer vision, which can have significant implications for cancer diagnosis and treatment.

1.3.4.3 Consideration in Model Construction for Object Counting

As object counting is reliant on accurate image localization, considerations for this task remain the same as those for image localization. This is because accurate localization is a prerequisite for accurate counting, and any errors in localization will propagate to counting results. These sections provide a comprehensive overview of the key challenges involved in developing accurate and reliable computer vision models for both image localization and object counting and offer valuable insights into potential solutions to these challenges.

1.3.5 Image Segmentation

1.3.5.1 Definition

Image segmentation is a fundamental computer vision task that involves partitioning a digital image into multiple segments, where each segment comprises a set of pixels that share common characteristics or represent a similar, identical, or connected image object. The goal of image segmentation is to simplify and/or change the representation of an image into a form that is more meaningful and easier to analyze. Image segmentation is widely used in a range of biomedical applications, such as medical imaging and pathology, where it can be used to accurately identify and analyze different regions of interest within an image. By breaking down complex images into smaller, more manageable segments, image segmentation

enables more precise and accurate analysis of digital images, leading to improved diagnostic accuracy and better patient outcomes.

1.3.5.2 Clinical Use Cases of Image Segmentation

Identifying chemotherapy-induced necrosis in WSIs of osteosarcoma is a critical clinical task that can be visualized computationally for the benefit of pathologists.[90] To achieve this goal, pathologists must provide pixel-level ground-truth annotations of both necrotic and non-necrotic areas in the WSIs to computer scientists, who will train an image segmentation model to differentiate between chemotherapy-induced necrotic regions and non-necrotic regions in osteosarcoma WSIs. Careful selection of a sufficient and balanced training sample size is essential, with pathologists choosing enough WSI samples of osteosarcoma containing both chemotherapy-induced necrosis and non-necrotic areas. Since osteosarcoma has the potential for systemic metastasis, computational scientists must also consider the possibility of developing organ-specific segmentation models to differentiate between chemotherapy-induced necrosis and non-necrotic examples of osteosarcoma tissue that present outside the bone, to ensure model generalizability. Through collaboration between pathologists and computational scientists, accurate and reliable image segmentation models can be developed that aid in the diagnosis and prognosis of osteosarcoma, leading to improved patient outcomes.

1.3.5.3 Consideration in Model Construction for Image Segmentation

The questions that need to be addressed by computational scientists and pathologists when developing image segmentation models are like those in image classification problems. Therefore, they remain the same as those described previously, which provide a comprehensive framework for tackling the key challenges involved in developing accurate and reliable computer vision models for a range of biomedical applications, including image segmentation. By following these outlines, computational scientists and

pathologists can collaborate to develop highly accurate and effective image segmentation models that have important implications for disease diagnosis, prognosis, and treatment.

1.3.6 Image Visualization

1.3.6.1 Definition

Image visualization is a crucial computer vision task that involves the computational alteration of scanned WSIs. By modifying and enhancing the appearance of WSIs, image visualization can aid pathologists in making more accurate diagnostic and prognostic deductions, as well as improve the design of CAD systems by providing digitally altered WSIs for training. Using advanced image processing techniques, such as color enhancement, contrast adjustment, overlays and heatmaps, image visualization can reveal previously hidden features and structures within WSIs, providing pathologists with a more detailed and nuanced view of tissue morphology and cellular architecture. This, in turn, can lead to more accurate diagnoses and improved patient outcomes. By incorporating image visualization techniques into the development of CAD systems, computational scientists can also improve the accuracy and reliability of these systems, leading to better disease detection and diagnosis. Human oversight offered by "human-in-the-loop" (HITL) AI systems that incorporate pathologist interactions may further aid accuracy and reliability while bolstering safety and quality control. Overall, image visualization is a critical task in the field of biomedical imaging, with important implications for a range of clinical and research applications.

1.3.6.2 Pitfalls of Image Visualization for Clinical Use Cases

Image visualization algorithms aid both human and computational diagnostic faculty. Algorithms designed for color normalization may improve standardization of staining irregularities in digitized hematoxylin and eosin (H&E) slides, thereby leading to improved accuracy and reliability of CAD

systems. Enhancements in contrast and color balance (in addition to reduced staining variability) have been found corollary to greater refinement in disease detection and other diagnostic applications.[91–103]

For human interpretation of heterogeneous tumor presentations, a foundationally complex task often further confounded by single-instance representations of tumor histomorphology present on one WSI, image visualization algorithms may improve model training and pathologist interpretation accuracy by employing multiplex-detection based multiple instance learning.[104] Multiplex detection strategies facilitated using image visualization algorithms may highlight critical characteristics of tumor presentation through leveraging memory-based learning capable of portraying various phenotypes within the tumor feature space.[104]

As the field of biomedical imaging continues to evolve, it is likely that new and improved methods for evaluating the effectiveness of image visualization algorithms will emerge, enabling computational scientists to develop more accurate and reliable models for a range of clinical applications. Development of image visualization algorithms is critical for improving the inherently subjective nature of pathologist-dependent model performance evaluation.

1.3.6.3 Consideration in Model Construction for Image Visualization

When implementing image visualization algorithms, two important questions that need to be considered are the size and diversity of the training dataset, and the issue of imbalanced datasets. These questions are addressed in detail in the literature, and it is important to carefully consider potential solutions to these challenges when developing image visualization algorithms.

To ensure the effectiveness of image visualization algorithms, it is essential to use a large and diverse training dataset that includes a broad range of different types of images and tissue samples. This will help to ensure that the algorithm is able to generalize to a wide range of imaging scenarios and produce accurate and reliable results in a clinical setting. In addition, addressing issues related to imbalanced datasets is also critical to the success of image visualization algorithms, as these algorithms often rely on large, labeled datasets to achieve high levels of accuracy. To address these challenges, researchers may consider approaches such as data augmentation, oversampling, or under sampling to ensure that the training dataset is balanced, and representative of the population being studied.

By carefully considering these factors and leveraging the latest advances in machine learning and computer vision, it is possible to develop highly effective and reliable image visualization algorithms that have important implications for a range of biomedical applications.

1.3.7 Image Generation

1.3.7.1 Definition

Image generation is a type of computer vision task focusing on creating synthetic photorealistic images. Image generation be used for a variety of applications, e.g., three-dimensional (3D) reconstruction of two-dimensional (2D) WSIs,[105] data augmentation from synthetic images,[106] virtual staining, and computational removal of inked pathologist annotations that are digitized into the composite WSI specimen.[107]

1.3.7.2 Pitfalls of Image Generation for Clinical Use Cases

Three-dimensional visualization of two-dimensional WSIs using 3D image reconstruction algorithms offers revelatory insight from the enhanced spatial latitude of lesion terrain afforded during

examination.[108] When compared to conventional 2D histology, the data-rich nature of 3D histology presents the opportunity for greater comprehensive analysis of 3D microstructures of prognostic significance. Such is demonstrated in 3D visualization of digitally reconstructed core-needle biopsies (CNB) of the prostate resulting in greater detection of over graded, under graded, and missed cases of prostate cancer in addition to superior prognostic stratification and grading precision than yielded from analysis of 2D CNB sections.[16] Though in nascent development, advances in high-throughput 3D microscopy have begun to pique adoption interest from clinicians.[16] Synthetic photorealistic 3D images must be subjectively evaluated by pathologists in the same manner as all other image visualization algorithms as no standardized statistical measurement criterion currently exists.

1.3.7.3 Consideration in Model Construction for Image Generation

Like image visualization, image generation also requires careful consideration of the size and diversity of the training dataset, as well as issues related to imbalanced datasets. To achieve accurate and reliable results, it is essential to use a large and diverse training dataset that includes a broad range of different types of images and tissue samples. Researchers may also consider approaches such as data augmentation, oversampling, or under sampling to address imbalanced datasets and ensure that the training dataset is representative of the population being studied. Decisions must also be made regarding the need for pixel-level precision per specific use case vs. free-reign creation of generated images.

Also of consideration are the many additional challenges in existing approaches for image generation, including realism and unseen correlations between image features that constrain generated images to non-incorporable effigies of corporeal tissue traits under study. "Hallucinations", i.e., generated images that are derived from alterations of real images, e.g., generated images of colon cancer derived from real

images of normal colonic mucosa, are susceptible to potentially catastrophic real-world diagnostic pitfalls as they contain patterns that are non-existent within input images, though are present in generated outputs.

By carefully addressing these factors and leveraging the latest techniques in machine learning and computer vision, it is possible to develop highly effective and accurate image generation algorithms that have important implications for a range of biomedical applications. As these techniques continue to evolve, it is likely that new and improved approaches for addressing these challenges will emerge, leading to even more accurate and reliable results in the future.

1.4 Discussion and Conclusion

The field of CP may be eponymously characterized as the merger of clinical and computational thought hemispheres for the creation of a unified, synergistic brain. CAD tools have the potential to revolutionize clinical pathology practice by providing accurate, efficient, and reproducible diagnostic solutions offering enhanced productivity and improved diagnostic capabilities.[109] Furthermore, diagnostic and prognostic insights rendered from CP tools may impact downstream improvements throughout multidisciplinary care settings.[33,110] Yet, fractionated communication between computer scientist and pathologist minds currently leaves the full potential of CADs unrealized.

Complete clinical realization of CP is founded upon effective and clear communication amongst pathologists and developers. Pathologist understanding of computational terminologies and processes provides context of algorithmic limitations and capabilities that may be leveraged in leadership of algorithm construction for clinical use-case endpoints. The core ensemble of CAD development includes pathologist, data scientist, and computer engineer cohorts, in which the pathologist role is integral from initiation through completion.[16] A pathologist-identified clinical need for a specific patient population,

laboratory process, or end-user marks the origin of the algorithm development. The process concludes with pathologist application of the algorithm within a real-world clinical setting, where monitoring of the algorithm continues for relaying feedback to the CAD development team for further optimization.[111] Appropriately conveying the clinical diagnostic problem and the specific use-case for which an algorithm will be applied is tantamount to the success of all training and development steps which follow. As stewards of laboratory information,[112] pathologists are purveyors of data for algorithmic training and therefore have a duty to ensure the provision of data curated for algorithmic generalizability. When insufficient training data is available to ensure generalized applicability for a wide range of patients, pathologist awareness of general computational concepts and firm understanding of how much time they may offer to the algorithm development process are critical elements that shape the direction of CAD development, e.g., machine learning method employed, and therefore must be communicated clearly by the pathologist to the rest of the CAD development team. Our standardized framework of approach to clinical objectives and their computational execution provides both computational scientists and pathologists with a shared language and understanding that may be used to bridge the clinical-computational translation gap.

The benefits of clinically useful CADs for pathology are being documented with increasing frequency, with some CAD solutions deployed in anatomic laboratories for specific use-cases including quality control and first-read applications.[16] FDA approval has recently been granted for in-vitro diagnostic use of CAD software designed to supplement pathologist detection of suspicious carcinomatous regions in prostate needle biopsy WSIs (Paige Prostate Detect, Paige, New York).[113] Such milestones are marked within an increasing number of peer-reviewed publications chronicling AI implementation within clinical pathology laboratory settings.[26]

Limitations in pathologist understanding of ML model purpose, development, and output may result in inappropriate use or interpretations that levy profound clinical ramifications. Computer scientist limitations in understanding true algorithmic use cases may lead to algorithmic production of high scoring yet clinically meaningless metrics with harmful outcomes. Studies have shown that the AI models could outperform the pathologists for subspecialty diagnostics. When comparing the diagnostic performance of pathologists with AI models, it is important to note that these studies often suffer from limitations. They either include only a small number of pathologists, or they marginally include or completely exclude subspecialty pathologists.[114] Such studies may be heralded for their groundbreaking results yet lack true clinical applicability. Additionally, when developing an algorithm for a subspecialty use-case, the creation or procurement of annotated training data should be ideally executed and/or facilitated by pathologist practitioner(s) of the same subspecialty expertise. Holistically, a shared foundational understanding of algorithmic function, use, and interpretation is tantamount to achieving clinical translatability in CP.

1.5 Outline of the Dissertation Research

This dissertation focuses on enhancing cancer diagnostics through the use of AI technologies. The research efforts described herein involve the creation of an innovative anomaly detection pipeline, applied to colorectal cancer (CRC) tumor segmentation on WSIs. Furthermore, the importance of hyperparameter configurations and dataset variations in aiding model architecture selection for breast cancer diagnostics is emphasized. The contributions of this dissertation research go beyond the development of novel AI algorithms for WSI analysis. Valuable insights have also been provided regarding the standardization of close collaboration between pathologists and computational scientists, aiming to facilitate the effective development and evaluation of AI algorithms in the field of DP.

In Chapter 1, an overview of seven commonly encountered WSI analysis tasks was presented. These tasks encompass image classification, likelihood measurement, object localization, object counting, image segmentation, visualization, and generation. The chapter further provided a standardized pipeline aimed at bridging the knowledge gaps between clinical pathologists and computational scientists. This pipeline facilitated the collaborative development and evaluation of AI algorithms specifically tailored for DP.

In Chapter 2, the research presented focused on the application of both non-specialized and DP-specialized model architectures for the binary classification of normal and malignant breast cancer tissues on WSIs. The chapter detailed the conducted experiments, which aimed to explore the effects of hyperparameter configurations on the performance of breast cancer image classifiers. The findings highlighted the importance of hyperparameter tuning in conjunction with the development of specialized model architectures in the field of DP. Additionally, Chapter 2 examined the impact of dataset variations on the performance of classification models, emphasizing the crucial aspect of considering dataset differences when making decisions about model architecture selections.

Chapter 3 presented a detailed explanation of the innovative Progressive Context Encoders for Anomaly Detection (P-CEAD) model for WSI analysis, focusing on its performance in detecting anomalies in melanoma WSIs. The chapter covered various aspects, including data preprocessing, the design of model architectures, the three phases of model training, and the model inference procedure. Furthermore, a qualitative assessment of the model performance was conducted by a senior anatomic pathologist. The results of this evaluation indicated the potential of applying the anomaly detection approach for segmenting malignant tumors from WSIs, which subsequently led to the research efforts described in **Chapter 4**.

Chapter 4 delved into the research endeavors involving the utilization of the P-CEAD approach for segmenting malignant tumors on CRC WSIs. The chapter provided a succinct summary of the model

training and inference procedures, along with the comprehensive quantitative and qualitative assessments conduced to evaluate the performance of the model.

Chapter 5 commenced with a summary of the main contributions made by the dissertation research conducted in Chapters 2 - 4. Subsequently, the chapter proceeded to outline the research limitations and present future directions aimed at further enhancing the research outcomes.

1.6 Publications, Contributions, and Declarations:

Chapter modified with permission from the following manuscript (submitted to the Archives of Pathology & Laboratory Medicine, currently under peer-review process):

Gu Q, Patel A, Hanna MG, Lennerz JK, Garcia C, Zarella M, McClintock D, Hart SN. Bridging the Clinical-Computational Transparency Gap. 2023.

The re-print request has been approved with a written consent from the Archives of Pathology & Laboratory Medicine.

QG, AP, and SNH designed and conducted the literature review. QG and AP wrote the manuscript. MGH, JKL, CG, MZ, DM, and SNH edited the final version of the manuscript. QG partially secured funding. SNH supervised the study.

This study is partially funded by the University of Minnesota Graduate School 2022-2023 Doctoral Dissertation Fellowship.

The authors have declared that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

The authors have declared that none of the known generative AI tools are used to assist scientific writing for this manuscript.

CHAPTER 2: MODEL ARCHITECTURE AND HYPERPARAMETER CONFIGURATION IN ASSISTING BREAST CANCER DIAGNOSTICS FROM WHOLE SLIDE IMAGES

Abstract

Breast cancer is one of the most common cancers in women. With early diagnosis, some breast cancers are highly curable. However, the concordance rate of breast cancer diagnosis from histology slides by pathologists is unacceptably low. Classifying normal versus tumor breast tissues from microscopy images of breast histology is an ideal case to use for DL and could help to diagnose breast cancer more reproducibly. Since data preprocessing and hyperparameter configurations have impacts on breast cancer classification accuracies of DL models, training a DL classifier with appropriate data preprocessing approaches and optimized hyperparameter configurations could improve breast cancer classification accuracy.

The experiments involved training and testing 12 combinations of DL model architectures, comprising five non-specialized and seven digital pathology-specialized models. These experiments also encompassed image data processing and various hyperparameter configurations. The validation accuracy of tumor versus normal classification were calculated using the BreAst Cancer Histology (BACH) dataset.

The DenseNet201, a non-specialized model architecture, with transfer learning approach achieved 98.61% validation accuracy compared to only 64.00% for the digital pathology-specialized model architecture.

The combination of image data preprocessing approaches and hyperparameter configurations have a profound impact on the performance of deep neural networks for image classification. To identify an effective deep neural network for classifying tumor versus normal breast histology, researchers should not solely concentrate on developing new models exclusively for digital pathology. This is because optimizing hyperparameters of existing deep neural networks from the computer vision field could often achieve a high (and sometimes even superior) prediction accuracy.

2.1 Introduction

Breast cancer is one of the leading cancer-related causes of death in women.[115] Early-diagnosis for breast cancer can reduce the mortality rate for breast cancer patients given that 70-80% of patients with early diagnosis of non-metastatic breast cancer are curable.[116]

Breast biopsy is the definitive way to diagnose breast cancer,[117] however, the concordance rate between different pathologists in interpreting breast biopsies is relatively low (overall concordance rate is 75.3% with 48% concordance rate for atypia).[118] To improve agreement, DL has shown success in solving broader computer vision problems,[119] particularly in the medical image analysis field.[120]

The advent of whole slide imaging[121] has heralded a new era in pathology research, enabling the detailed analysis of histological images through DL methodologies.[122] This is highlighted in the work of Iizuka et al.,[123] who successfully employed DL algorithms to identify gastric and colonic epithelial tumors within histological slide preparations. Their approach achieved remarkable levels of accuracy, as demonstrated by Area Under the Curve (AUC) values of 97% and 99% for the prediction of gastric adenocarcinoma and adenoma, respectively. Likewise, colonic adenocarcinoma and adenoma prediction achieved AUC values of 96% and 99%, respectively. These findings underscore the potential of DL-based image

classifiers to enhance diagnostic precision, positioning it as a promising approach for distinguishing normal tissues from malignant neoplasms.

Differentiation of malignant tumors and normal tissues on histology slides can be achieved by two DL-based image classification approaches. First, non-specialized deep neural networks have been applied to group different classes of histology from microscopy images. Transfer learning[124] is a popular non-specialized approach, which uses either the last layer or all layers of the pre-trained networks, including InceptionV3,[125] DenseNet201,[126] ResNet152,[127] and VGG19[128] models for image classification. One-shot learning,[129] a distance-based classification model, is another non-specialized approach to predict the object categories from a few training samples. Koch, et al.[130] adopted the one-shot learning model for image classification[131] achieving near-state-of-the-art classification accuracy. Aside from general use networks, specialized deep neural networks have also been developed for microscopy images. The clustering-constrained attention multiple instance learning (CLAM) model[54] is a digital pathology specialized multi-class image classifier. CLAM is an attention-based weakly-supervised learning model that does not require large amounts of well-annotated training samples. CLAM is a unique approach in digital pathology, that ranks the patch-level feature importance by attention scores, then ranks information to train the final classifier.

Different DL models could affect the classification performance. However, hyper-parameter configurations[132] and data preprocessing[133] also have impacts on the performance of image classifiers. Zhou et al.[134] proposed a comparative experiment to study the impacts of hyperparameters on DL model performance. They found the classification precision scores varied from 84.8% to 99.5% for a number of 36 combinations of deep CNNs (DCNN)-based a roadway crack classification problem. They tested various hyperparameter configurations, including learning rate, dropout, and batch size on 10,000 test

images from laser-scanned roadway range image dataset (LRRD).[135] In addition, Heidari et al.[136] proposed a study to compare the performance of VGG16-based transfer learning approach with or without image preprocessing in classifying the Coronavirus Disease 2019 (COVID-19), non-COVID-19 pneumonia, and non-pneumonia cases from 8,504 2D X-ray images. The authors yielded a 7.4% increase in overall classification accuracy of the VGG16-based classifier with image preprocessing compared with the model without pre-processing steps. This indicates that the image preprocessing could also alter the DL model performance. Therefore, the standard deep neural networks could achieve a better classification performance by hyperparameter tuning and selecting appropriate data pre-processing techniques. What is not known is how much of a difference hyperparameters, model architectures, or general versus domain specific architecture make on medically relevant images like those in digital pathology.

The BACH dataset[137] is a publicly available dataset of H&E-stained microscopy images of breast histology labeled into four classes (i.e., “normal”, “benign”, “in situ carcinoma” and “invasive carcinoma”). An ensemble network-based image classifier proposed by Marami et al.[138] was the best performing model on the BACH dataset with the highest prediction accuracy. Their model was able to achieve an 84% accuracy for the four-class classification required by the BACH Challenge, but also achieved a 91.7% accuracy in classifying carcinoma versus non-carcinoma breast histology. The carcinoma versus non-carcinoma classification was made possible by using a binary classification model in which the images from “normal” and “benign” classes were reassigned into a single “non-carcinoma” class and images from “in situ carcinoma” and “invasive” carcinoma classes were reassigned into a single “carcinoma” class. However, the proposed approach by Marami et al. was to build a de novo algorithm using an ensemble of CNNs rather than fine tuning the conventional deep neural networks (i.e., ResNet,[127] and InceptionResNet[139]). Therefore, the proposed study compared the performance of models with

different combinations of hyperparameters and data preprocessing techniques, including custom versus purpose-built models.

2.2 Subjects and Methods

2.2.1 Data Preparation

Four hundred microscopy images of breast histology in ".tif" format were downloaded from the BACH dataset. Out of the 400 images, there are 100 microscopy images from each of the "benign", "normal", "in situ carcinoma" and "invasive carcinoma" classes. To reorganize the images from the BACH dataset for binary carcinoma versus non-carcinoma classification, images in the "benign" or "normal" BACH classes are labeled the "non-carcinoma" class (i.e., class zero) and images within the "in situ carcinoma" or "invasive carcinoma" BACH classes are labeled the "carcinoma" class (i.e., class one).

To create a five-fold cross validation dataset, all 400 images were first randomly shuffled and divided into five groups. To maintain a balanced dataset in each of the five groups, each group ended up with 80 images, including 40 images each from the carcinoma and non-carcinoma classes. For each of the five-folds, one of the five groups is selected as the validation set, while the remaining four groups are selected as the training set. The five-fold cross validation dataset preparation was implemented using the Scikit-Learn Python package.[140] Therefore, in each of the five-folds, there are 320 images with 160 images from each of the carcinoma and non-carcinoma classes used for training, and 80 images with 40 images each from the carcinoma and non-carcinoma classes used for validation. Patches from 400 microscopy images were extracted and saved in the TFRecords file format with each TFRecords file including the image patch array, file name, width, and height of the image patch.[141]

CLAM required image patch-level feature vectors as the model training input data -rather than images - while the pre-trained InceptionV3, DenseNet201,[126] ResNet152, VGG19, and one-shot learning model only required pixel data as input. **Sections 2.2.1.3 - 2.2.1.4** details the image feature extraction and normalization, specific for CLAM, while the **sections 2.2.1.1 - 2.2.1.2** describe patch extraction, image standardization, and scaling - all of which are identical for all deep neural networks.

Of note, it was also necessary to re-implement CLAM as it did not support the BACH files and some of the standardized profiling that are needed to perform. Comparing the re-implemented CLAM with the original source code confirmed there was no difference in classification outcomes. To make the comparison, A number of 40 H&E-stained malignant breast histology WSIs were downloaded from the Cancer Genome Atlas (TCGA) database.[142] These 40 WSIs include 20 *BRAF* mutated and 20 wild-type malignant breast histology WSIs. Then, 10 cross-validation sets were created by randomly selecting 35 out of the total 40 WSIs for each of the 10 folds, and split into training, validation, and testing sets. In each cross-validation set, there were 15 WSIs in the training set, 10 WSIs in the validation set, and 10 other WSIs in the testing set. The extracted image patches were used to create the image feature vectors for all the WSIs in each of the 10 cross-validation sets without any image preprocessing.

2.2.1.1 Image Patch Preparation

Each microscopy image from the BACH dataset has 2,048 x 1,536 x 3 pixels with a pixel scale of 0.42 μm x 0.42 μm.[137] JPEG format images (n=19,200) of 256 x 256 x 3 pixels in were split into 5-fold cross validation sets, with 15,360 image patches in the training (Class0: n=7,680; Class1=n=7,680) and 3,840 patches (Class0: n=1,920; Class1: n=1,920) in the validation (**Figure 2. 1**).

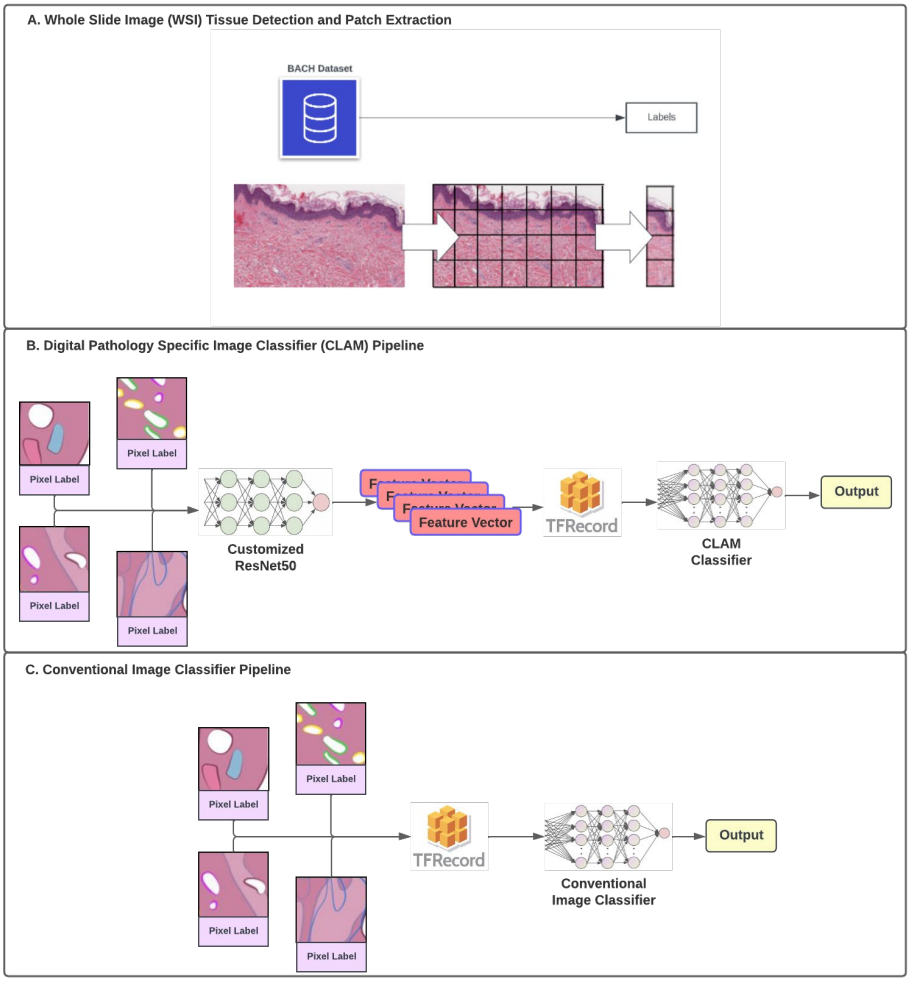


Figure 2. 1 *Pipeline Diagram for Digital Pathology-Specialized and Non-Specialized Image Classifiers. A). Whole Slide Image Tissue Detection and Patch Extraction; B). DP-Specialized Image Classifier (CLAM) Pipeline; C). Non-Specialized Conventional Image Classifiers (i.e., DenseNet201, InceptionV3, One-Shot Learning, ResNet152, and VGG19) Pipeline.*

2.2.1.2 Image Standardization

Image standardization is an image rescaling technique that linearly scales each of the 3 RGB-channel (i.e., red, green, blue) image patches to a mean of 0 and variance of 1. The formula of this technique to compute the standardized image patch array $\hat{x}$ is:

$$\hat{x} = (x - \bar{x}) \,/\, max(\sigma, (1.0 \,/\, sqrt(N)))$$

where,

$$\bar{x} = \sum_{i=1}^{N} x_i \ , \sigma = sqrt((\sum_{i=1}^{N} (x_i - \underline{x})^2) \ / \ N)$$

and N is denoted as the number of elements in each of the image patch x. An additional image rescaling technique is also applied in one of the experiments in this study. The formula used to compute the rescaled image patch array $\hat{x}$ from the original image patch array x is:

$$\hat{x} = abs(x_i \ / \ 255) \ \in \ [0,1], i = 1,2,3,\ldots,N$$

where N is denoted as the number of elements in each image patch x. Details of the combinations of different image scaling methods and experiments were listed in **Table 2. 1**.

Table 2. 1 *Data Preprocessing and Hyperparameter Configurations Summary Table for the DP-Specialized (CLAM) and Non-Specialized Image Classifiers (i.e., DenseNet201, InceptionV3, One-Shot Learning, ResNet152, and VGG19).*

Model Index	Study (Author, Year)	Model Name	Model Category	Data Preprocessing Technique	Optimizer Option	Loss Function	Learning Rate	Dropout Rate	Batch Size	Number of Epochs
C1	Lu et al., 2021	CLAM	DP-Specialized	Image Standardization; Image Feature Normalization	Adam	BinaryCross Entropy	2E-04	0.5	48	20
C2	Lu et al., 2021	CLAM	DP-Specialized	Image Standardization	Adam	BinaryCross Entropy	1E-04	0.25	48	20
C3	Lu et al., 2021	CLAM	DP-Specialized	Image Standardization	Adam	BinaryCross Entropy	1E-05	0.25	48	50
C4	Lu et al., 2021	CLAM	DP-Specialized	Image Standardization	Adam	BinaryCross Entropy	5E-05	0.25	48	20
C5	Lu et al., 2021	CLAM	DP-Specialized	Image Standardization	Adam	Hinge	5E-05	0.25	48	20

C6	Lu et al., 2021	CLAM	DP-Specialized	Image Standardization	SGD	CosineSimilarity	2E-03	0.25	48	20
C7	Lu et al., 2021	CLAM	DP-Specialized	Image Standardization	SGD	BinaryCrossEntropy	3E-03	0.25	48	20
D1	Huang et al., 2017	DenseNet201	Non-Specialized	Image Standardization	Adam	BinaryCrossEntropy	1E-05	0.25	20	16
I1	Szegedy et al., 2015	InceptionV3	Non-Specialized	Image Standardization	Adam	BinaryCrossEntropy	1E-05	N/A	20	7
O1	Li, 2006	One-Shot Learning	Non-Specialized	Image Standardization	Adam	BinaryCrossEntropy	1E-04	N/A	32	5
R1	He et al., 2015	ResNet152	Non-Specialized	Image Standardization	Adam	BinaryCrossEntropy	1E-05	N/A	20	5
V1	Simonyan et al., 2014	VGG19	Non-Specialized	Image Standardization	Adam	BinaryCrossEntropy	1E-05	N/A	20	6

2.2.1.3 Image Feature Extraction

The pre-trained ResNet50 model on ImageNet[143] was employed to extract image feature vectors for the preparation of CLAM model training. RGB channel image patches with dimensions of 256 x 256 x 3 were fed into this pre-trained ResNet50 model. Following processing through the third residual block of the pre-trained ResNet50 model, a 1,024-dimensional patch-level image feature vector was obtained (**Figure 2. 1**).

2.2.1.4 Image Feature Normalization

Image patch-level feature vectors are the required input for CLAM training. The L2 normalization[144] was applied on the extracted 1,024-dimensional patch-level image feature vectors to generate the normalized patch-level image feature vectors.

Each of the L2 normalized patch-level 1,024-dimensional image feature vectors $\hat{x}$ was computed from each of the original patch-level 1,024-dimensional image feature vectors x by the following,

$$\hat{x} = x \,/\, sqrt(max(\sum_{i=1}^{N} x^2 \ , \varepsilon))$$

where ε has a default value of 1E-12, and N is denoted as the number of elements in each of the patchlevel 1,024-dimensional image feature vectors x.

2.2.2 Model Training

2.2.2.1 Transfer Learning with Pre-Trained DL Models

Transfer learning was applied with different non-specialized model architectures, including InceptionV3, DenseNet201, ResNet152, and VGG19. These models were first pre-trained on ImageNet, then the last layer of these pre-trained models was trained on the H&E microscopy images from the BACH dataset. Training details of these models with the corresponding combinations of data preprocessing (i.e., image standardization, and image feature normalization), and hyperparameter configurations (i.e., learning rate, dropout rate, optimizers, loss functions, number of epochs, and batch size) are listed in **Table 2. 1**.

2.2.2.2 One-Shot Learning

One-shot learning was applied to learn the domain features from microscopy images from the normal and tumor classes reorganized from the BACH dataset. This would have allowed the model to classify the normal versus tumor breast histology from microscopy images.

Training details of the combination of the one-shot learning model, image data preprocessing (i.e., image standardization, and image feature normalization), and hyperparameter configurations (i.e., learning rate, dropout rate, optimizers, loss functions, number of epochs, and batch size) are listed in **Table 2. 1**.

2.2.2.3 Clustering-Constrained Attention Multiple Instance Learning

Microscopy images of breast histology from the BACH dataset are in ".tif" format, which is not supported by the original CLAM implementation. A TensorFlow-version[141] CLAM was re-implemented with three jointly trained neural networks (i.e., attention network,[145] instance classifier, and bag classifier[146]). To ensure the re-implemented CLAM achieves a similar classification performance as the original CLAM, both the original and re-implemented CLAM were evaluated on 10 validation WSIs from each of the 10 cross-validation sets to compute the validation accuracy. Then a student's t-test ran on the AUC of the original and re-implemented CLAM on 10 validation WSIs from each of the 10 cross-validation sets to determine whether the re-implemented CLAM achieves a similar binary classification accuracy as the original CLAM.

Then, similar to the experiments that have been performed using the non-specialized classifiers as discussed on **section 2.2.2.1 - 2.2.2.2**, the validation accuracy of CLAM with seven different combinations of data preprocessing (i.e., image standardization, and image feature normalization), and hyperparameter configurations (i.e., learning rate, dropout rate, optimizers, loss functions, number of epochs, and batch size) are listed in **Table 2. 1**.

All code, including the implementations of non-specialized and digital pathology-specialized model architectures, is publicly available at https://github.com/quincy-125/DP_BACH.

2.3 Results and Discussion

2.3.1 CLAM Reimplementation Results on TCGA Data

The AUC scores returned from both the original and re-implemented CLAM on 10 validation TCGA WSIs from each of the 10 cross-validation sets are shown in **Figure 2. 2**. There was no significant difference between the performance of the original and re-implemented CLAM (p-value=0.67).

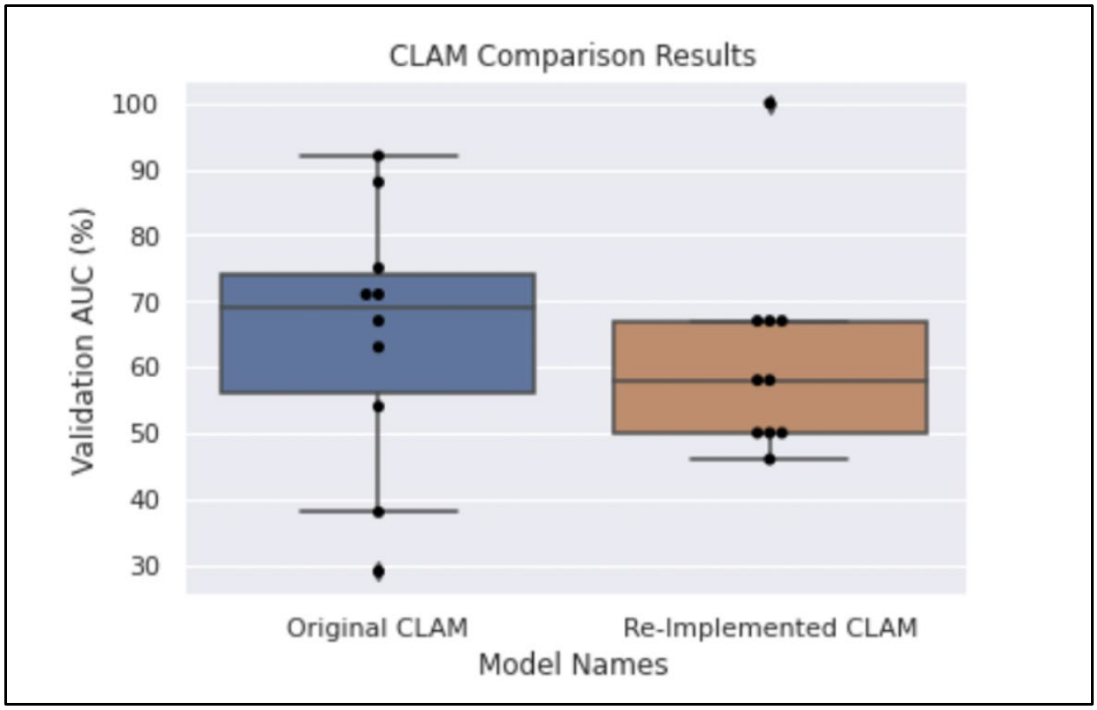


Figure 2. 2 *CLAM comparison box plot for the TCGA dataset. Each black dot represents the validation classification AUC scores from each of the 10-fold cross-validation sets. Left). Box plot for the original Pytorch-Version CLAM; Right). Box plot for the Tensorflow-Version re-implemented CLAM.*

2.3.2 Model Performance Comparison on the BACH Dataset

The validation accuracies of both the non-specialized classification models using DenseNet201, InceptionV3, One-Shot Learning, ResNet152, and VGG19 with each of their corresponding image preprocessing applied and optimized hyperparameter configurations, and the digital pathology-specialized CLAM models with seven different combinations of image preprocessing and hyperparameter configurations are listed in **Table 2. 2**. Among the results returned by the experiments, the DenseNet201

model (indexed as D1 in **Table 2. 1**), was the best performing model in classifying normal versus tumor breast tissues from the BACH dataset with a 98.16% validation accuracy. The optimal image standardization and hyperparameter configurations included the Adam optimizer, BinaryCrossEntropy as the loss function, learning rate=1E-05, batch size=20, and number of epochs=20.

Table 2. 2 *Results table including the validation accuracy of the non-specialized and DP-specialized model architectures with different hyperparameter configurations.*

Model Index	Model Name	Model Category	Validation Accuracy (mean ± std)
D1	DenseNet201	Non-Specialized	**98.61% ± 1.13%**
R1	ResNet152	Non-Specialized	97.08% ± 0.78%
I1	InceptionV3	Non-Specialized	95.29% ± 0.23%
V1	VGG19	Non-Specialized	89.48% ± 0.66%
O1	One-Shot Learning	Non-Specialized	82.40% ± 9.31%
C1	CLAM	DP- Specialized	60.00% ± 6.80%
C2	CLAM	DP- Specialized	64.00% ± 4.87%
C3	CLAM	DP- Specialized	64.00% ± 9.74%
C4	CLAM	DP- Specialized	63.00% ± 3.54%
C5	CLAM	DP- Specialized	50.00% ± 0.00%
C6	CLAM	DP- Specialized	51.00% ± 1.73%
C7	CLAM	DP- Specialized	56.00% ± 5.12%

2.3.3 Hyperparameter Tuning in Breast Cancer Classification Model Development

Hyperparameter tuning is critical to boost the classification performance, in addition to the model architecture. The results shown in Table 1 indicated that with the optimal hyperparameter configurations,

the non-specialized image classifiers, including the DenseNet201, ResNet152, InceptionV3, VGG19 with the transfer learning approach, and the One-Shot Learning approach, could outperform the digital pathology specialized model architecture, CLAM. This suggests that computational pathologists may need to focus more on hyperparameter tuning, rather than designing more complex digital pathology specialized model architectures. The learning rate has a higher impact on both the non-specialized and digital pathology-specialized classifiers performance compared with the rest of the hyperparameters (i.e., options of optimizers and loss functions, dropout rate, batch size, and number of epochs), and thus should be the first parameter to augment when optimizing models.

In addition to manual hyperparameter tuning, the automated hyperparameter searching algorithm is another option in selecting the optimal hyperparameter configurations. Therefore, future work could adopt automated hyperparameter tuning, which could improve the efficiency of the process to identify the optimal hyperparameter configurations.

2.3.4 Impacts of Dataset Differences on CLAM Performance

Dataset difference could affect the classification model performance, in addition to model architecture, and hyperparameter configurations. The unique architecture of the CLAM model led to the performance gap of CLAM on the BACH and TCGA dataset. CLAM is an attention-based multiple-instance learning image classifier, the attention module in the CLAM architecture first assigns attention scores to each of the patches from a certain WSI, then use the top- and least- k patches sorted from their corresponding attention scores as the positive- and negative- examples of the slide-level label. Since all patches in the BACH dataset are only informative tissue, each contributes equally to the slide-level label. This deviation violates the expectation of the CLAM model, that weights informative and non-informative patches -

inherently assuming that some of the images are non-informative. Therefore, CLAM should only be used when slides contain both informative and non-informative features.

DenseNet201, a non-specialized image classification model, had the highest validation accuracy (98.16%) in the breast cancer classification in this cohort. This study also indicates the impacts of hyperparameter configurations, and dataset differences, have a significant impact on image classification model performance. This suggests that digital pathology researchers must be careful to understand the strengths and limitations of choosing a model that is suited to the task at hand.

2.4 Publications, Contributions, and Declarations:

Chapter modified with permission from the following published article which is publicly available on medRxiv (submitted to the Journal of Pathology Informatics, currently under peer-review process):

Gu Q, Prodduturi N, Hart SN. Deep Learning in Automating Breast Cancer Diagnosis from Microscopy Images. Published online June 16, 2023:2023.06.15.23291437. doi:10.1101/2023.06.15.23291437

QG and NP implemented the model architectures and conducted the experiments. SNH and QG designed the experiment. QG partially secured the funding. SNH supervised the study.

The results shown in **Figure 2.2** are in whole based upon data generated by the TCGA Research Network: https://www.cancer.giv/tcga.

This work was funded by the University of Minnesota Graduate School Doctoral Dissertation Fellowship for the year of 2022-2023, and the Department of Laboratory Medicine and Pathology at the Mayo Clinic.

Prof. Shery L. Holt from the University of Minnesota, and Dr. Kristin Cardiel Nunez from the Children's Hospital of Philadelphia provided the support of English language editing and review services.

The authors have declared that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors have declared that the ChatGPT was used to assist language editing in the writing process. The use of ChatGPT was done with the authors oversight, control, and was carefully reviewed and edited by the authors.

CHAPTER 3: MELANOMA TUMOR SEGMENTATION FROM WHOLE SLIDE IMAGES USING PROGRESSIVE CONTEXT ENCODERS

Abstract

WSI is transforming the practice of pathology, converting a qualitative discipline into a quantitative one. However, one must exercise caution in interpreting algorithm assertions, particularly in pathology where an incorrect classification could have profound impacts on a patient, and rare classes exist that may not have been seen by the algorithm during training. A more robust approach would be to identify areas of an image for which the pathologist should concentrate their effort to make a final diagnosis. This anomaly detection strategy would be ideal for WSI but given the extremely high resolution and large file sizes, such an approach is difficult. Here, we combine progressive generative adversarial networks (GANs) with a flexible adversarial autoencoder architecture capable of learning the "normal distribution" of WSIs of normal skin tissue at extremely high resolution and demonstrate its anomaly detection performance. Our approach yielded pixel-level accuracy of 89% for identifying melanoma, suggesting that our label-free anomaly detection pipeline is a viable strategy for generating high quality annotations -without tedious manual segmentation by pathologists. The code is publicly available at https://github.com/quincy-125/P-CEAD.

3.1 Introduction

Skin cancer is the most common of all human cancers, with one million people in the United States diagnosed each year with some type of the disease. Most skin cancers are basal and squamous cell carcinomas. While malignant, these types are relatively easily cured with minimal surgical intervention. Malignant melanomas, however, account for about 1% of all skin cancers in the United States but cause the majority of skin cancer deaths. The number of people diagnosed with melanoma has risen sharply

over the past three decades. In men and women ages 50 and older, the number of people diagnosed with melanoma increased 3% per year from 2006 to 2015. Identification and validation of melanomas are of critical importance, as patients with dermatologist-detected melanomas have better survival, lower overall mortality, and lower cancer-related mortality.[147]

The most important recent advances in microscopy for surgical pathology were the invention of the digital microscope[148] and WSI.[149] The current generation of high-speed, high-capacity whole slide scanners can process between one and 1,000 slides at multiple resolutions and different image planes (i.e. z-stacks). The quality of the images has been steadily increasing over time. Several recent comparisons have been made between rendering a diagnosis on a glass slide and a digital assessment, with concordances reported between 75-100%.[150–152] The image files themselves are quite large, between 2-5 GB each, requiring analysis to be conducted on much smaller regions (a.k.a. "patches").[153]

Perhaps the most exciting opportunity resulting from the digital pathology transition to WSI is the potential utility of applying AI allowing for CP.[109,154,155] We previously showed that AI was capable of differentiating between Spitz and Conventional Nevi (benign skin lesions).[156] Later, Hekler *et al*.,[157] trained an AI to differentiate between compound or junctional nevi and melanoma. However, both approaches suffer from limitations of the training set. The World Health Organization (WHO) recognizes nine different types of melanoma,[158] but there are also significant numbers of benign lesions that mimic melanomas or are detected at premalignant stages.[159] Models like these that only account for two possible outcomes (Spitz / Conventional or benign / malignant), necessarily means that if the input image is not from either class, then the model will always make an incorrect diagnosis. Given some of the rarer examples of benign and malignant lesions, it may not be possible to accumulate enough examples from each class to build a model that captures everything one would see in dermatopathology practice.

A more practical approach would be to convert the "classification" problem into an "anomaly detection" problem. The major difference is that during training, the anomaly detection approach only sees one class (e.g., normal) whereas the classification approach needs examples from each possible class. In DP, the former is ideal since examples with negative findings are easily acquired. Rather than learning how to differentiate between all possible classes, the model instead learns to weigh each pixel as to how likely it belongs to the normal class, and if that score lies outside some expected distribution, then it gets flagged as an anomaly - without the need to say what class that anomaly belongs to. The advantage for digital pathology is that the areas of interest can quickly be identified and carefully scrutinized by the pathologist - who can more carefully consider the question of "what" the anomaly actually is.

3.1.1 Anomaly Detection Using Generative Adversarial Networks (GANs)

GANs are well suited for anomaly detection problems. A generative adversarial network (GAN) consists of two adversarial modules, a generator (G) and a discriminator (D). G typically learns to generate realistic looking images ($\hat{x}$) from latent-space vectors (z), which are then served to the discriminator for determining their real or fake status (q). However, the mapping from $z \rightarrow \hat{x}$ is different than $\hat{x} \rightarrow z$ which is important for understanding how and where the vector contributes to the generated image. AnnoGAN[160] requires an additional step to learn this mapping. To avoid this, EBGAN[161] simultaneously added an encoder (E) model for joint training of three networks (G, D, E) and added the latent/pseudo-latent variable (z / $\hat{z}$) as an input into D as in BiGAN.[162] GANomaly[163] rearranged the models into an adversarial autoencoder wherein $x \rightarrow z \rightarrow \hat{x} \rightarrow \hat{z}$, thereby removing the true G and leaving a bowtie architecture generator that both encodes (G_E) and decodes (G_D). An adversarial loss ($\mathcal{L}_{ADV}$) is calculated from a D, a contextual (a.k.a. reconstruction) loss ($\mathcal{L}_{CON}$) ($\hat{x} - x$), and an encoder loss ($\mathcal{L}_{ENC}$) measuring the difference of latent space mappings ($\hat{z} - z$). Di Mattia et al.,[164] recently reviewed the use of each of these GANs for use in anomaly detection. The

limitation to each of these methods is that they have been applied only to low resolution images. Berg *et al.,*[165] proposed GANanomalyDetection (by combining ProGAN[166] with a more updated version of AnnoGAN[160]), which theoretically could work for high resolution images, but was only tested on low resolution images. The main novelty of the ProGAN approach was the concept of training the model on smaller representations of the source images, while gradually increasing the image size and model architecture to very high resolution (1024 x 1024 pixels). The question as to whether the integration of progression learning in GANanomalyDetection for ultra-high-resolution images in digital pathology remains unanswered.

The work in GANomaly[163] showed that one could ignore the complication of GANs altogether and use autoencoders (AEs). Autoencoders compress an image x to a small latent space z which is then transformed into a reconstructed image $\hat{x}$, while minimizing the reconstruction error between x and $\hat{x}$.[167] Naturally, this ideal property led to the use of AEs for image compression.[168] Adversarial autoencoders however, take this one step further by drawing samples from the latent distribution z and combining the original reconstruction loss ($\mathcal{L}_{CON}$) with an adversarial loss ($\mathcal{L}_{ADV}$). Lazarou[169] combined both the generative and autoencoding aspects into an Autoencoding Generative Adversarial Network (AEGAN). The major limitation of AEGAN however, is that it requires two discriminators, an encoder, and a generator, so supporting large model architectures becomes difficult to implement in practice since the memory and compute required to train becomes technically infeasible and/or too expensive.[170]

An innovative way to take advantage of the ideal properties of both GANs and AEs is to combine their functionality. Pathak *et al.,*[171] called this a Context Encoder (CE), whereby an image is first augmented to remove blocks of pixels, then processed through an AE, with $\mathscr{L}_{\text{CON}}$ relative to the unmodified input image. This forces the model to create a semantically meaningful representation of the missing data (*i.e.,* GAN)

while learning a latent representation (*i.e.,* AE). However, CEs have only been applied to smaller images (512 x 512)[172] that are unrelated to the ultra-high resolution required for CP.

Here, we combine CE with the progressive framework to encode high resolution images from digital pathology and bias this learning toward encoding only non-diseased skin tissue so that the model will not correctly encode abnormal tissue, forcing high reconstruction error that can be exploited for anomaly detection. We are able to show that our P-CEAD model can segment tumor regions with high accuracy without the need for manual segmentation by a pathologist, representing a major step forward for more complex digital pathology workflows. This new method was able to achieve 89% pixel-level accuracy for anomalous regions of interest when compared to manually segmented melanomas.

3.2 Materials and Methods

The training dataset consists of a total of 200 slides from the Department of Laboratory Medicine and Pathology at Mayo Clinic. A senior dermatopathologist selected them from skin excisions based on the absence of inflammation, neoplastic process, and quality of glass slides. The testing dataset includes eight skin slides with definitive invasive melanoma. All slides are anonymized and scanned at 40X with Aperio ScanScope V1. The invasive melanoma in the test dataset was annotated by a pathologist using QuPath 0.2.3.[173]

Due to the complexity of the task and allowing for modularity, P-CEAD involves several phases of processing and model training (**Figure 3. 1)**. Briefly, in Phase 0, the preprocessing step identifies and removes image patches that are highly similar in image content. In Phase 1, the progressive autoencoder with inpainting is trained using 150 normal WSIs. During Phase 2, inference is run on 25 normal WSIs to calculate the normal error reference distribution. Next, in Phase 3, reconstruction error profiles from 25

additional normal WSIs were compared to the reference distribution to determine the value of three standard deviations from normal, which is used as a binary threshold to flag pixels as anomalies. Finally, inference uses a smoothed and filtered kernel density estimator (KDE) on the binary pixel flags in eight tumor-containing WSIs to determine an appropriate KDE threshold for defining an anomalous region.

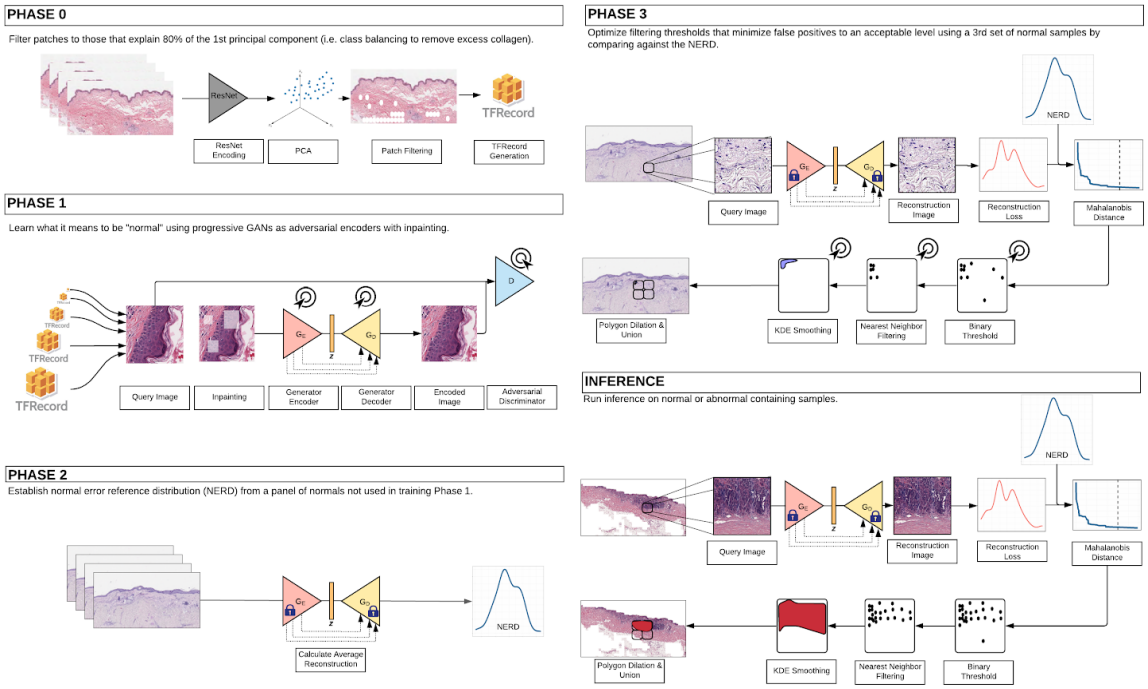


Figure 3. 1 *Overall architecture of the multiple components of P-CEAD.*

3.2.1 Data Preprocessing

The goal of the P-CEAD is to learn the manifold of normal images so it can identify outliers on that manifold for anomaly detection. Phases 1, 2, and 3 of P-CEAD all require normal images, so a significant amount of training data is necessary. However, many of the image patches contain redundant information. Skin sections are made of three compartments: epidermis, dermis, and subcutis. The subcutis is made of white fat appearing as mostly empty vacuoles in the H&E slide. The dermis is primarily made of collagen and represents an overwhelmingly large contribution of the total tissue observed on each slide. However,

the epidermal layers (stratum corneum, lucidum, granulosum, spinosum, basale) represent the vast majority of diagnostically relevant regions. The class imbalance of these three layers means that the networks would mostly see collagen and fat rather than learning to focus on the epidermis. To overcome this limitation, training images are pre-filtered. First each image is passed through ResNet50 to obtain a vector representation of the image. Principal Component Analysis (PCA) is then applied to this collection of image vectors to project the vectors into an orthonormal basis. From there, the images corresponding to the first 80% of the first principal component were selected. As expected, manual inspection also confirmed that these selected patches captured mostly epidermis, borders, and examples from each of the layers, whereas many images with collagen and whitespace were removed. Across the 200 WSIs there were 534,531 image patches in the pre-filtering set and 427,625 post-filtering (80% of the original image patch count). However, the filtering was not uniformly 80% for each slide with some slides having very little filtering and others being extremely filtered. The minimum slide filtering had 99.29% of image patches remaining and the maximum slide filtering had 9.71% of image patches remaining, with an average of 76.1%.

Since P-CEAD involves three training phases, three distinct training sets were generated to ensure that the same image patches were not used for multiple phases. Training Phase 1 was the most computationally intensive, the majority of WSIs (n=150,330,415 patches) were used in this Phase. Twenty-five WSIs each were used for training Phases 2 and 3, corresponding to 44,693 and 52,517 patches, respectively. As mentioned previously, these “normal” slides were selected based on the absence of inflammatory and neoplastic processes.

3.2.2 Network Weight Training

Before training the network weights, images must be augmented. Otherwise, due to the architecture (and even more specifically the skip connections), the generator would simply become a compressive autoencoder. However, by masking the inputs the generator is forced to learn not only how to encode and decode the image, but also constrain the encoding and decoding to be context specific, in essence learning to encode images in a way that enforces an expected image type (in this case a normal skin WSI).

A percentage of pixels in each image (default 20%) are masked for each image patch. Random squares are generated using a halving geometric series such that the total sum of all of the masked pixels approximately matches the desired masking percentage. The squares are allowed to overlap, and if so, will randomly have their intersections unmasked, creating more complex shapes. This is uncommon enough that the effective mask percentage is still close to the setpoint, yet it does help break edge symmetries to discourage the model from just learning edge detection. Each mask block then randomly shuffles its pixels so that the original information is still within the mask region, albeit not in the correct spatial order. Randomly shuffled pixels produced more realistic inpainting than standard black pixel masking, but increased processing time when shuffling on-the-fly for each image patch. Adding randomly shuffled masked versions of the image patches to the TFRecords would decrease the amount of compute and memory cost but doing it on-the-fly ensures that each epoch will produce a different masking for each image which increases the diversity for learning the normal image manifold and reduces the likelihood of overfitting on a fixed set of masked images. Masked images are only used during the Phase 1 of training, whereas the other phases of training only use the unmasked image patches.

The model architecture at a high level consists of a generator and a discriminator. Having an additional network to act as an encoder (as in some GAN architectures) yielded negligible improvement and

increased training time so was excluded from the final model architecture. All subnetworks follow the general pattern of Kerras *et al.*'s Progressively Growing GANs such as the generator using pixel normalization, the discriminator using minibatch standard deviation, WGAN loss with gradient penalty, epsilon drift penalty, no batch normalization, leaky ReLUs for activations, grows in resolution by factors of two from 4x4 images to 1024 x 1024 images, Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.99$, etc.

The generator has a bowtie autoencoder structure consisting of an encoder that takes an image, x, as input and outputs a latent vector, z, and a decoder which takes a latent vector, z, as input and outputs an image $\hat{x}$. During Phase 1, the actual input to the bowtie is x', the masked version of image x. To help facilitate high-resolution image generation, the bowtie architecture is specifically a U-net with skip connections between the encoder and decoder's corresponding convolutional layers. Without skip connections, high resolution generated images were blurry from high resolution pixel information being lost at the U-net bottleneck. Skip connections were also pruned during testing, but much of the high-resolution detail was lost by not including all of them and thus all skip connections are retained in the final model.

The discriminator is a standard progressively growing GAN discriminator. However, conditional weights were added for reconstruction and adversarial loss terms. For the generator, the final weighting scheme was,

$$\mathcal{L}_{CON} = 1.0 * L2(x - G(x))$$

and

$$\mathcal{L}_{ADV} = 1.0 * (-q(D(G(x))))$$

giving the total generator loss of

$$\mathcal{L}_{GEN} = \mathcal{L}_{CON} + \mathcal{L}_{ADV} = L2(x - G(x)) - ExpVal(D(G(x)))$$

For the discriminator there were two adversarial loss terms

$$1.0 * D(x) - 1.0 * D(G(x))$$

In addition, there is a WGAN gradient penalty (GP) (γ)[174] and an epsilon drift penalty (EDP) (ε) as done in Kerras *et al.* giving the total discriminator loss of

$$\mathcal{L}_{DIS} = q(D(x)) - q(D(G(x))) + \gamma + \varepsilon$$

For each step in training Phase 1, for each image x in a minibatch of image patches X, random masked blocks were created for each patch using the shuffling method. These augmented images x' are the same tensor shape as the original input images x. x' is then passed through the bowtie, with intermediate tensors flowing along the skip connections. The final output of the generator's decoder will be an image $\hat{x}$, of the same tensor shape as the generator input x', and therefore the same as our original inputs from the TFRecords. The L2 norm between the original images x and the generated images $\hat{x}$ is the generator's reconstruction loss term.

Both the real images x and the generated images $\hat{x}$ then proceed to the discriminator network where each example will receive one logit indicating whether the discriminator thinks it is the original image or is augmented. The discriminator will then be updated using the mean of the difference between the real and fake logits in addition to the GP and EDP. The negative mean of the fake logits is then added to the generator network using alternating gradient descent, since that provides more stable learning than simultaneous gradient descent. Both the discriminator and the generator had only one network weight update at each step.

Phase 1 was trained for one epoch at each image size. The prediction outputs after training the Phase 1 model are the generated images $\hat{x}$ and the absolute errors between the input images x and the generated images $\hat{x}$.

3.2.3 Normal Error Reference Distribution Calculation

Once the progressive GAN's network weights have been trained, the model weights are fixed in the bowtie generator. Beginning here, images are no longer augmented with masks since the inpainting training through the learned network weights is complete. The purpose of this phase is to learn the Normal Error Reference Distribution (NERD). This comes from the maximum likelihood estimation of the errors which are assumed to be gaussian in nature. Therefore, the NERD is a multivariate gaussian distribution of the absolute errors between x and $\hat{x}$ with a mean vector for each color channel (RGB) and a 3 x 3 covariance matrix for the color channels. The NERD's mean vector and covariance matrix parameters are calculated from 25 WSIs not used in Phase 1.

The motivation for including the NERD is due to the fact that all imperfect models produce prediction errors, and those errors are samples from a distribution of errors. Normal images should produce very low prediction errors of a certain distribution due to the progressive autoencoding in Phase 1, assuming the learned manifold of normal images is accurate. On the other hand, anomalous images should produce higher prediction errors, representing a different distribution.

The prediction outputs after training this phase of the model are Mahalanobis distances, calculated using the color channel means and covariance matrix. This distance metric is commonly used for finding multivariate outliers, whereas unlike Euclidean distance treats each axis independently (a sphere), Mahalanobis distance considers the scales (and cross correlations) of each axis (an ellipsoid). Lower distances correspond to pixels that have low absolute errors between input and generated images x and $\hat{x}$, respectively, while larger distances correspond to pixels that have larger absolute errors.

3.2.4 Dynamic Distance Threshold, Kernel Density Estimator (KDE) Smoothing, and Dilation

Knowing the Mahalanobis distances, one can determine whether an observed value exceeds that from the expected distribution. Larger distances are indicative of higher reconstruction error and are more likely to be anomalous. To convert the linear measurement of distance into a binary classification metric such as anomalous or not, a threshold can be determined where anything above it is assigned "anomalous" and everything below it is assigned "normal". Using the fixed bowtie and the NERD, the optimal binary flag threshold is determined from a final catalog of 25 normal WSIs. Each image patch is run through the bowtie generator to yield the reconstruction error profile and compared with the NERD to calculate each pixel's Mahalanobis distance. Just like for the NERD, these Mahalanobis distances form a distribution across all of the images in the Phase 3 training dataset, so the mean and standard deviation of Mahalanobis distances is calculated for this set. The threshold is set to be equal to the mean plus a number of standard deviations for the Phase 3 training set.

At full resolution, each image patch is a matrix of 1024 x 1024 pixels, with each pixel having a value of zero (normal) or one (anomaly) which equates to over a million pixels. Up to this point, all pixels have been treated as independent. However, even at 40X magnification, each pixel represents a region smaller than the nucleus of a single cell - far below the resolution of the human eye, much less the resolution that a human would be capable of providing manual annotation for which to make benchmark comparisons. Moreover, there is nothing inherently intuitive or medically relevant for a single pixel. Instead, larger regions are of interest. Given these constraints, it is unreasonable to expect every pixel in a true anomaly to be flagged as an anomaly. Instead, spatial information can be leveraged to define the anomalous regions more accurately.

Since pixel-level flags can be noisy, spurious anomalous pixels in an otherwise large normal pixel region may incorrectly be classified as anomalies. To account for these spurious calls, two steps are performed: filtering and smoothing. To remove false positive pixels, we want to remove any small clusters of pixels. Looking at each flagged pixel, a cluster is removed if within a specified connectivity (default = 1) that the neighboring flagged pixel count is less than minimum adjacent pixel neighborhood size (default = 5). In other words, given an anomalous pixel *a*, if pixels connected to it and pixels connected to those and so on add up to less than five pixels surrounding *a*, then *A* is removed, otherwise it is retained. After regional filtering, a patch-level threshold is also applied, requiring that at a minimum of *A* (default = 10) anomalous pixels are found, otherwise all anomalous flagged pixels will be removed.

For the remaining pixel-level anomaly flagged images that need to be smoothed, a 2D gaussian KDE is used on the pixel-level anomaly flags. Other kernel types and several bandwidth values were explored, but the best results were obtained with a gaussian kernel with a bandwidth of 100. A minmax normalization then transforms those evaluations to be within [0., 1.] and then scaled by *(anomaly_flag_counts / scaling_factor) ^ scaling_power* for visual consistency. These will be used for thresholding at a specified value between [0., 1.] to create a boolean mask. Therefore, rather than using the scaled kernel densities with a color map, we directly use those as is, making a grayscale KDE image.

Finally, the scaled grayscale KDE images can be compared against a specified threshold (default = 0.2) to create a bitmask image. These are then rasterized and converted to polygon objects using Shapely[175] which are then dilated (default=4) to expand the shapely polygon beyond the image patch bounds. A final union of polygon objects is rendered across all image patches in a WSI. In this way, potential edge effects from neighboring patches will be eliminated. We then intersect the union polygons with a polygon of the original patch boundaries to constrain any dilated exteriors from extending outside the patch regions.

To measure performance relative to human annotation, the intersection, union, and difference between the two sets of polygons are used to create resultant polygon sets that represent different confusion metrics. TPs are measured as the area of intersection between the KDE and human-curated annotation polygons. FPs are measured as the area of difference between the KDE polygons and the true positive intersection polygons. FNs on the other hand are measured as the area of difference between the annotation polygons and the true positive intersection polygons. This leaves the TNs as the remaining area of the original patches. Sensitivity is calculated as TP / (TP + FN) and specificity as TN / (TN + FP).

3.3 Results

Phase 0 computation took 12.5 hours of computation to compress with ResNet and run PCA using 2GB of memory serially in a Google Colab notebook instance with 32 vCPUs and 208 GB RAM and 2 NVIDIA Tesla T4s. All distributed training was performed on n1-highmem-16 virtual machines with two NVIDIA Tesla V100 GPUS on the Google Cloud Platform (GCP). Phase 1 training required 129.5 hours using 16GB memory. Phase 2 and 3 both completed in under one hour with 16GB memory.

We initially attempted to simply adapt the GANanomalyDetection architecture to whole slide images. However, the approach failed to yield acceptable images and was discarded. When exploring images generated by the GAN, it was apparent that it had only learned to encode collagen and fat, which make up the vast majority of pixels in whole slide images of skin. We attempted to reduce the redundant images by selecting the images that most contributed (80%) to the overall variance using PCA (**section 3.2 Materials and Methods**). However, we could not filter too many images out or else the model would not be exposed to sufficient examples of normal tissue. Another issue we found with GANanomalyDetection was that it either learned to encode the real images or the fake images - but never both. As a result, either

the fake or the real images appeared blurry or would have a mode collapse (**Figure 3. 2**) when using the loss calculation (**Figure 3. 3**),

$$G(z) - (G(E(G(z))))$$

This could not be overcome despite applying multiple weights to each loss function or changing combinations thereof.

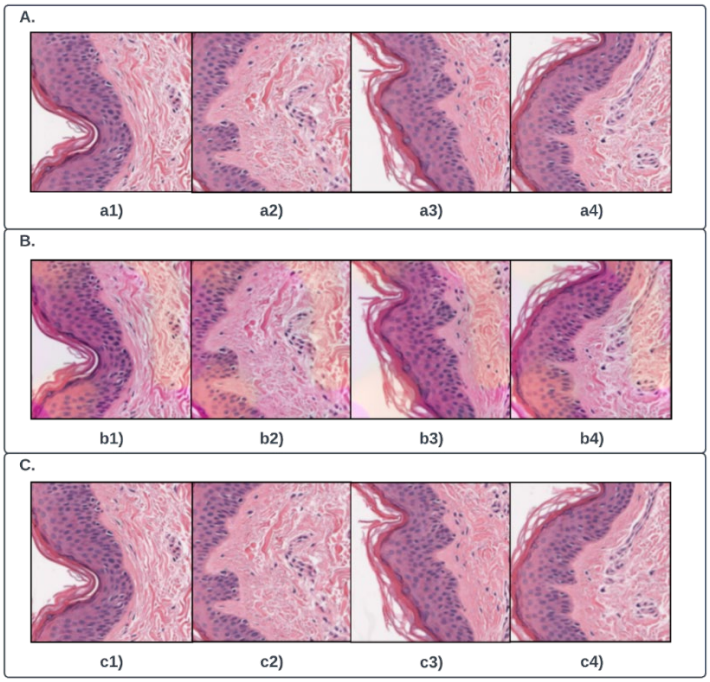
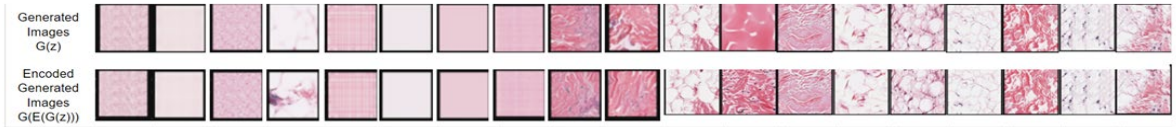


Figure 3. 2 *Effect of skip connections. A). Original image patch; B). Skip connections turned off; C). Skip connections turned on. No uniform noise was added to* $\boldsymbol{z}$ *or to fake images, and the loss for "*$\boldsymbol{x}$*_minus_*$\boldsymbol{G}$*_of_*$\boldsymbol{x}$*_L2_loss_weight" was zero.*



Exp ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
q(1 - D(G(E(x))))							1	1											
q(1 - D(G(E(G(z)))))																			
||z - E(G(z))||																			
||E(x) - E(G(E(x)))||						1	1	1											
||G(z) - G(E(G(z)))||	1	1	1	0.01	1	1	1	1			0.001			0.001	0.001	0.001	0.001	0.001	0.001
||x - G(E(x))||		1	0.01	1	1	1	1	1				0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
q(1 - D(G(E(x))))							1	1									50		50
q(1 - D(G(E(G(z)))))																		50	50
||z - E(G(z))||																			
||E(x) - E(G(E(x)))||					1	1	1	1											
||G(z) - G(E(G(z)))||	1	1	1	0.01	1	1	1	1	1	1	1			1	1	1	1	1	1
||x - G(E(x))||		1	0.01	1	1	1	1	1		1		1	1	1	1	1	1	1	1
q(1 - D(G(E(x))))								1						1		1	1	1	1
q(1 - D(G(E(G(z)))))															1	1	1	1	1

Figure 3. 3 *Loss function penalty exploration.*

As an alternative approach, we pivoted from a GAN-based architecture to an Adversarial Autoencoder-based architecture. Real images were encoded and decoded well in terms of histological structure, but distortions in color were also present (**Figure 3. 3**). This coloration artifact was overcome by adding skip connections to the autoencoder, producing a more reasonable reconstruction. The downside of the Autoencoder was that it simply learned to encode images - regardless of whether they contained anomalies, so input regions were masked from the image to learn context encoding. The rationale is that the context encoder would learn to preferentially compress and decompress images free of anomalies because it learned what normal should look like. When asked to compress and decompress a region with anomalies, the reconstruction error should be much higher in abnormal images because the model would not have learned how to encode and decode abnormal images (**Figure 3. 4**).

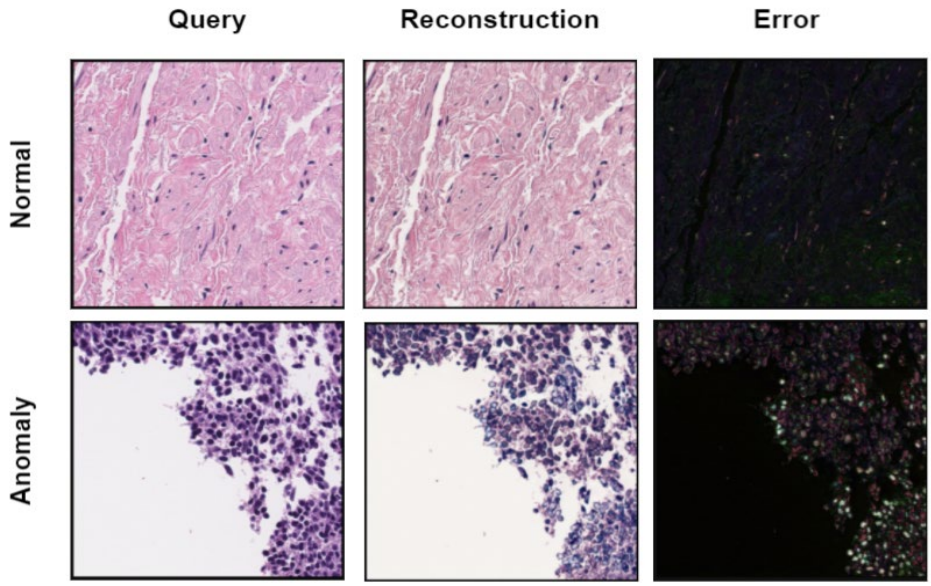


Figure 3. 4 *Examples of reconstruction error from image reconstruction.*

Once satisfied with the overall approach, we evaluated model performance on eight slides with anomalies that were not in any phase of training. A board certified pathologist created ground truth segmentation maps to indicate areas of melanoma on H&E slides. Using a KDE threshold of 0.21 and a polygon dilation factor of 4, on average the model was both sensitive (94% ± 8%), specific (87% ± 7%), and accurate (89% ± 7%) (**Table 3. 1**).

Table 3. 1 *Performance metrics for eight whole slide images containing melanoma.*

Slide Index	Sensitivity	Specificity	Accuracy
S1	99%	89%	90%
S2	100%	74%	75%
S3	97%	72%	86%
S4	78%	94%	94%
S5	100%	98%	98%
S6	83%	91%	88%

S7	97%	89%	89%
S8	97%	92%	93%

We should also note that inference parameters for identifying melanoma may not be the same for non-melanoma lesions or anomalies in other tissues. Rather, the inference module we constructed allows users to select, adapt, and change filtering criteria to suit the task at hand through configurations. Each modification to the inference filters would necessarily alter the performance metrics. For example, choosing an alternative dilation factor for the polygon expansion can have a dramatic effect on sensitivity and specificity (**Figure 3. 5**).

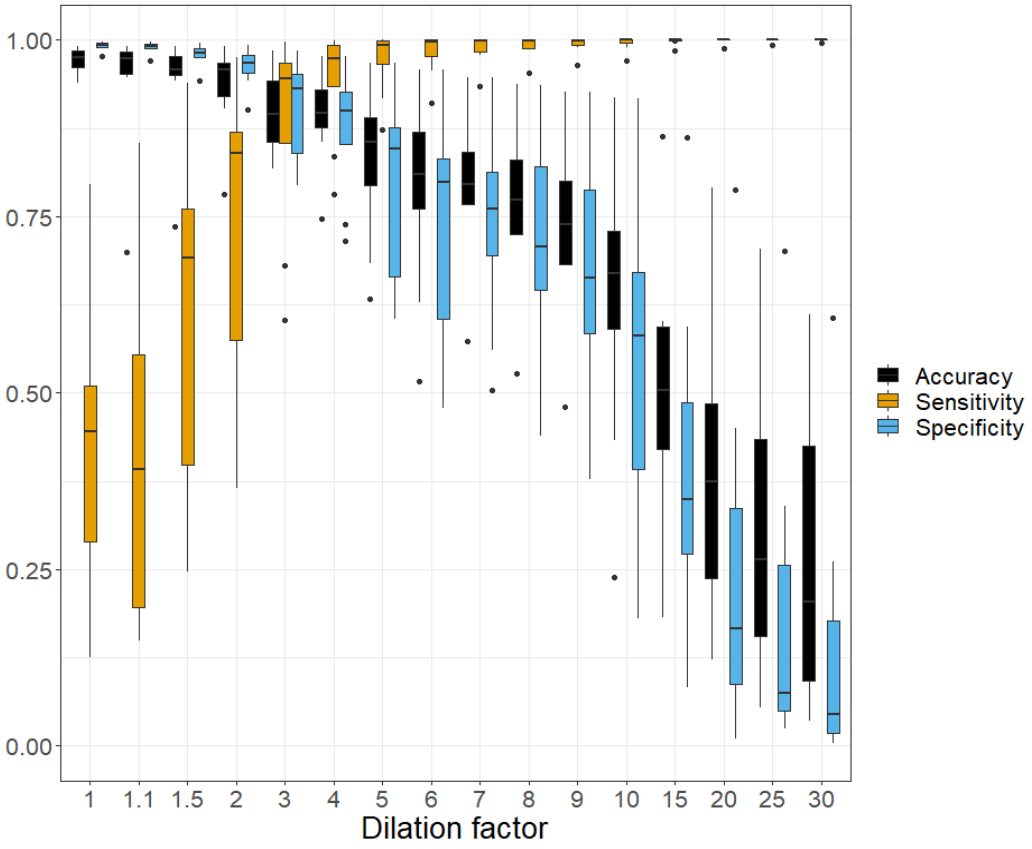


Figure 3. 5 *Performance metrics for eight whole slide images as a function of polygon dilation.*

3.3.1 Exploration of Predictions

To improve our understanding of the P-CEAD model, we explored predictions at the patch level. **Figure 3. 6** shows representative images that contain true and false positives and negatives. The top row is an example of normal whereas the bottom is completely involved with melanoma. In the normal image, most of the predictions were correctly identified as not containing anomalies, but pixels toward the top of the image patch were. Upon closer inspection, this is thought to be the result of epidermal lymphocytic infiltrate or presence of keratinocytes with abundant clear cytoplasm. Such changes, while not associated with tumors, were limited in our training set of histologically normal samples, for which very few were expected to have extensive immune infiltrate. Review of additional false positive regions suggested an enrichment in areas where the model expected - but did not identify - a preserved epidermis, due to the presence of parakeratosis, intraepidermal lymphocytes, knife artifact, epidermal denudation, or lumina of large arteries.

The case of false negatives is shown in the bottom row of **Figure 3. 6**. Here, the entire image is involved with melanoma, yet only a small portion of the image patch (Upper right corner) is flagged after filtering. Our interpretation of the missed pixels is due to an increase of degenerated tumoral cells, as well as increased extracellular matrix. Like the other example, reconstructed images are generally darker in color than the original query image - which ultimately results in higher reconstruction errors and flagged as anomalies.

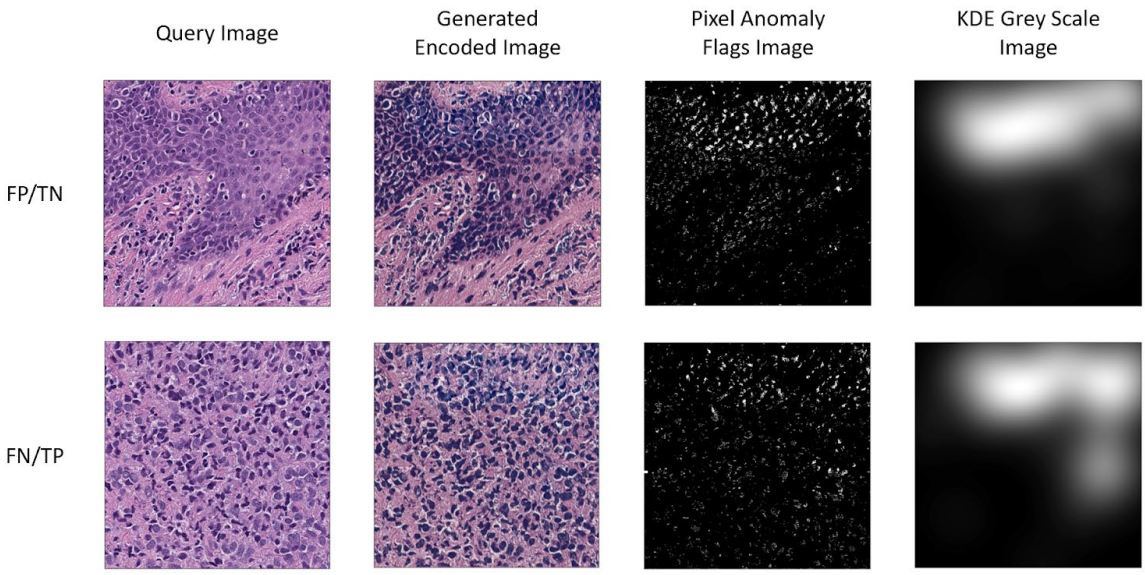


Figure 3. 6 *Examples of correct and incorrect predictions from P-CEAD. Top Row). The query image contains only normal tissue. Reconstructing the image through the P-CEAD model results in an overall darker image, which also corresponds to a higher error rate and subsequent flagging of individual pixels. The greyscale image defines the region for polygon creation, with white being used to call anomalies; Bottom Row). This query image contains 100% tumor, but only a portion was flagged as anomalous. This type of model exploration can inform users of how and where filters could be applied to refine final predictions. False positive (FP); True negative (TN); False negative (FN); True positive (TP).*

3.4 Discussion and Conclusion

The main contributions to P-CEAD are the diversity sampling for unsupervised patch selection, addition of inpainting, removal of unnecessary loss terms from previous architectures, and the development of a modular secondary process for fine tuning anomaly detection.

Image inpainting was required for G_E to preferentially encode patterns that were observed only in the training set. RGB values from blocks of randomly chosen coordinates were removed or shuffled before being auto encoded. The reconstruction loss (measuring the per-pixel delta of the auto-encoded image relative to the pre-augmented image) was thus not only measuring the quality of compression/decompression, but also the ability of G_E to "hallucinate" realistic patterns of normal.

Inpainting was imperative to bias the AE such that the AE would produce larger errors in the reconstructed images in the area of patterns it had not observed before (i.e., anomalies).

Multiple architectures and combinations of weighted loss terms were attempted before developing P-CEAD. Models, including GANanomalyDetection, were able to encode real or generated images, but never both. We attributed this effect to the latent-space loss term ($\mathcal{L}_{ENC}$, $\hat{z} - z$) and its dependency on a high quality generator architecture. Otherwise, as in the case of GANanomalyDetection, the training step is trying to minimize the loss when z' is dependent on the outputs of three independent models (generator-encoder, generator-decoder, and an additional encoder). In P-CEAD, the $\mathcal{L}_{ENC}$ is removed and thus obviates the need for the additional encoder network and now more resembles an autoencoder for the generator of the GAN system. However, unlike autoencoders (adversarial or otherwise), we are not interested in the latent space representation.[176] Instead, the generative component occurs though inpainting during G_E and G_D using the $\mathcal{L}_{CON}$ for optimization. The addition of skip-connections also had a profound influence on the reconstructed image quality and added to the model's generative ability.

P-CEAD defines anomalies in a more innovative and practical way than previous methods. GANomaly,[163] f-AnoGAN,[177] EBGAN,[161] and GANanomalyDetection[165] all define their anomaly scores as some derivation of reconstruction error relative to the query image. In contrast, P-CEAD measures the difference in reconstruction error relative to the NERD. This distinction is subtle, but important. In our approach, we separate the model training from error calculation. In an ideal world, one would have captured all representations of normal images during model training, but this is rarely possible in reality. The model we have presented here was trained using SVS files from an Aperio scanner with a JPEG2000 image compression that was sectioned, stained, and imaged at Mayo Clinic. We cannot be certain that the reconstruction error distribution from a slide processed by an external lab on a different scanner would be

the same as ours. However, having established the NERD on the internal dataset, one can easily compare it to the reconstruction error from the external laboratory's normal slides. If the distributions are the same, then no assumptions are violated, and the model should behave as expected. If they are not the same, then normal samples from the external lab could be used to calculate a new NERD for processing the external data, without retraining the computationally demanding Phases 0 and 1.

When exploring the triple network architectures (as in GANanomalyDetection) eight different weighted losses were attempted (**Table 3. 2**). The use of those losses could be applied to any or all of the independent networks and have a weight of 1, 0.01, 0.001, or 50. In general, we observed when L6 and L7 were greater than 0.001, then the generated images were visually similar to normal skin. Manual review of the generated images showed dominant representation of fat and collagen, despite being selected against in pre-training. L2 and L3 had little influence when applied to the encoder or discriminator architecture with respect to generating more realistic images. The real limitation was when trying to encode real images using the encoder network, particularly in images containing epidermal layers. Real images in this context were consistently highly blurred, rendering them useless for further investigation - despite a stabilization of all the loss terms.

Table 3. 2 *Loss Term Combinations.*

Loss Index	Loss Term	*G*	*E*	*D*
L0	$q(D(x))$			X
L1	$q(1 - D(G(z)))$	X		X
L2	$q(1 - D(G(E(x))))$	X	X	X
L3	$q(1 - D(G(E(G(z)))))$	X	X	X
L4	$\|\|z - E(G(z))\|\|$	X	X	

L5	$\|\|E(x) - E(G(E(x)))\|\|$	X	X	
L6	$\|\|G(z) - G(E(G(z)))\|\|$	X	X	
L7	$\|\|x - G(E(x))\|\|$	X	X	

A key limitation to P-CEAD is that it does not define what anomalies are. This is a separate task that should be performed after P-CEAD has segmented any anomalous regions. Another limitation is the nearest neighbor filtering during inference. In theory, requiring a minimum number of pixels to be co-located will decrease the analytical sensitivity of the method. Also, the costs and technical considerations required for Phase 1 training may be prohibitive for many clinical departments (**Table 3. 3**).

Table 3. 3 *Caveats of P-CEAD.*

Caveats Index	Caveats Title	Caveats Content
C1	Inappropriate image inputs	● If the network weights are trained on images that contain anomalies, then since the training is unsupervised, the model has no ability to know that these images aren't normal and therefore the learned manifold expands to encode those images within it. ● Likewise, the error distribution training in Phase 2 needs to contain only normal type images. This time, if anomalous images are within the training dataset, since the manifold is now fixed due to the first phase of training being complete, it is the error distribution's parameters that will now expand. This is due to the off-manifold anomalous images generating larger errors which then get encoded into the error distribution. ● Lastly, phase 3's training requires all images to be normal so that the tightest distance threshold can be set. If anomalous images were accidentally included in its training dataset, then larger distances would result, ergo raising the acceptable distance threshold. ● Thus, since all three training phases require the training datasets to have as little anomalous example contamination, it is crucial to ensure that any images that could be contaminated are filtered out before the training phases.
C2	Different PCA	● Other global filtering percentages were also assessed. ● Exceeding 80% resulted in many blurry and blank image patches

	thresholds	being included in the training datasets. ● Being stricter than 80% didn't inherently make the images unsuitable for training, but it did throw out too many useful images which could lead to our model not having enough unique data to learn what "normal" looks like across the three training phases. ● Thus, we stuck with keeping the top 80% of image patches.
C3	Mahalanobis distance thresholding	● One could directly use the Mahalanobis distance as a numerical score for pseudo likelihood of being normal. Distances are in the range of 0 to ∞ with lower scores meaning more likely normal. ● However, each pixel's distance is independent of its neighbors - even though there should be a very strong spatial correlation. This is the logical basis for the KDE smoothing since we would not expect pixels near each other to be representing different anomalies. ● However, the KDE expects a two-dimensional array binary mask, so distances do not satisfy this condition. This results in the KDE scoring each sample the same and therefore since there is no variance every sample result in a zero-valued grayscale pixel, thus the resulting image is entirely zeros. This can be remedied by converting the input images into binary masks via a threshold, clipping, etc. which is what we do with the dynamic threshold learned in training phase 3.

Despite its limitations, P-CEAD represents a significant step forward for applying AI to whole slide images within a clinically relevant context. Unlike challenge competitions or toy problems, P-CEAD is not based on a simple classification of tumor vs normal. Many clinically benign nevi mimic the architecture of pathogenic varieties - which could otherwise be classified as tumor in a binary prediction algorithm. P-CEAD allows the "gray area" to exist without overconfident claims of diagnostic accuracy. Furthermore, P-CEAD may have additional uses such as labeling anomalous regions for tissue scraping and downstream molecular testing.

3.5 Publications, Contributions, and Declarations:

Chapter modified with permission from the following published article which is publicly available on bioRxiv:

Gillard R, Meroueh C, **Gu Q,** et al. Using Progressive Context Encoders for Anomaly Detection in Digital Pathology Images. Published online July 4, 2021:2021.07.02.450957. doi:10.1101/2021.07.02.450957

RG, QG, SP, NP developed and tested the code, documentation, and experiments. CM and TF performed clinical annotation and case selection. SNH and TF secured the funding. SNH and TF designed the study, oversaw the study execution, and secured its funding. All authors discussed the results and commented on the manuscript and approved the final version.

This work was funded by the Leon Lowenstein Foundation (SNH), the Mayo Clinic / Google Joint Steering Committee, and the Mayo Clinic Center for Digital Health (SNH, TF).

RG and SP are employees of Google. QG, NP, CM, TF, and SNH have declared that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in the paper.

The authors have declared that none of the known generative AI tools are used to assist scientific writing for this manuscript.

CHAPTER 4: EXTENDING ANOMALY DETECTION BASED TUMOR SEGMENTATION ALGORITHM IN COLORECTAL CANCER USE CASES

Abstract

CRC is the 2nd most commonly diagnosed cancer in the United States. Genetic testing is critical in assisting in the early detection of CRC and selection of individualized treatment plans, which have shown to improve the survival rate of CRC patients. The tissue slide review (TSR), a tumor tissue macro-dissection procedure, is a required pre-analytical step to perform genetic testing. Due to the subjective nature of the process, major discrepancies in CRC diagnostics by pathologists are reported, and metrics for quality are often only qualitative. P-CEAD is an anomaly detection approach to detect tumor tissue from WSIs, since tumor tissue is by its nature, an anomaly. P-CEAD-based CRC tumor segmentation achieves a 71% ± 26% sensitivity, 92% ± 7% specificity, and 63% ± 23% F1 score. The proposed approach provides an automated CRC tumor segmentation pipeline with a quantitatively reproducible quality compared with the conventional manual tumor segmentation procedure.

4.1 Introduction

4.1.1 Background

CRC is the second most frequently diagnosed cancer in the United States for both sexes and is also the second most common cause of cancer-related deaths worldwide.[178,179] Genetic testing is the cornerstone of personalized medicine, and is rapidly becoming a necessary tool for prognostication and treatment selection, which have the potential to enhance the five-year survival rate of CRC patients.[180]

According to the most recent National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines in Oncology (NCCN Guidelines),[181] the most important factors that influence treatment selection include pathologic staging and prognostic markers, including, but not limited to, Mismatch

Repair (MMR) status (with reflex for *MLH1* promoter methylation or more expanded genomic testing), human epidermal growth factor receptor 2 (HER2) Immunostain / Fluorescent in-situ hybridization, and *KRAS, NRAS, BRAF, and NTRK* mutations. Next-generation sequencing (NGS) offers to investigate most of the above mutations / fusions.

It is important to conduct genetic testing in clinical CRC patient care, as 5% to 15% of cases are caused by inherited cancer susceptibility genes.[182,183] Identifying *TP53* mutation status can help to subtype and stage CRCs, leading to improved diagnosis.[184] *EGFR* inhibitor therapies are not effective for CRC patients with positive mutations in *KRAS*, *BRAF*, *PI3KCA*, and *PTEN*, highlighting the need for understanding genetic mutation status to select successful individualized therapeutic options. Different genetic mutation status also impacts CRC survival, where the CRC patients with a positive mutation of *LRP1B* have a higher recurrence rate and shorter progression-free survival (PFS) compared to those with a positive mutation of *FAT4*.[184] Therefore, CRC genetic testing is critical in improving predictions of CRC prognostics and survival rate.

In conventional clinical CRC patient-care pathways, tumor samples are formalin-fixed and paraffin-embedded into one or more tissue blocks. A guideline by Ballester and Cruz-Correa is used to determine if individuals should undergo genetic testing based on factors such as age at diagnosis of affected family members and personal and family history of colon polyps and extracolonic cancers.[185] If a patient meets the guideline for genetic testing, cytotechnologists will use the H&E-stained slide with circled tumor regions from pathologists to identify tumor tissue regions on an unstained slide. The tumor tissue is then sent to a molecular pathology laboratory for genetic testing.

The most important factors in ensuring a successful NGS testing are preanalytical variables, including selection of invasive tumor, size of invasive tumor, viability of tumor, and purity of tumor (i.e., minimal presence of benign cells, including inflammatory cells). The current clinical workflow for NGS testing, known as TSR, is completely manual and suffers from significant interindividual variation leading to discrepancies in diagnosis.[186] To improve on this process, we are proposing to employ an AI tumor segmentation algorithm to automatically detect tumor regions from digitized H&E-stained WSIs. This would allow control of multiple pre-analytical variables through selecting the block with the largest tumor surface area and segmenting that area for later tumor recovery for subsequent testing (**Figure 4. 1**).

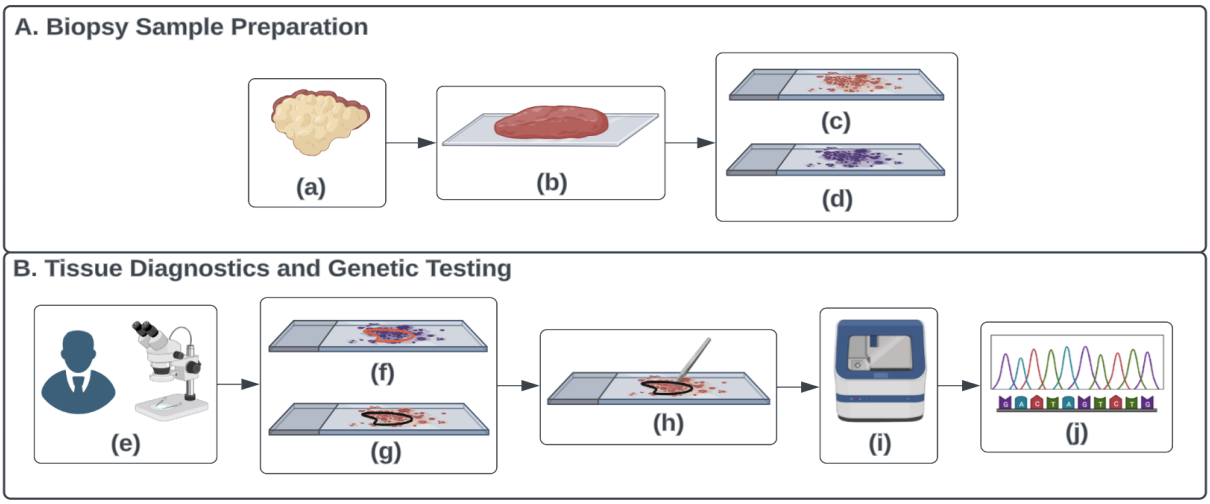


Figure 4. 1 *Diagram of manual workflow of TSR. There are ten components included in the figure. Component (a) is a CRC tumor tissue; (b) is a cut CRC tumor biopsy sample; (c) is a glass slide with the non-stained two-dimensional CRC tumor tissue block cut from (b); (d) is a glass slide with the H&E-stained two-dimensional CRC tumor tissue block section cut from (b), which is the adjacent two-dimensional CRC tissue block section to (c); (e) illustrates the general anatomic pathology practice workflow for pathologists to make cancer diagnostics using microscope on glass slides; (f) is the pathologists diagnostics with red polygon highlighting the CRC tumor tissue regions from (d); (g) is the black CRC tumor polygon on (c) that has been aligned with the red CRC tumor polygon on (d); (h) illustrates the clinical workflow for cytotechnologists to scrape the CRC tumor tissue on (g); (i) is the NGS device used for genetic testing; (j) is the genetic testing results from the NGS technology. Two sub-figures included in this figure, A). Biopsy Sample Preparation Pipeline; B). Tissue Diagnostics and Genetic Testing Pipeline.*

4.1.2 Related Work

Image classification is a widely used method for detecting tumor regions in WSIs. This approach labels a WSI as either CRC positive or negative. However, it does not provide the exact location of the tumor regions in the slide with their corresponding x- and y- coordinates.[59,187–189] On the other hand, image segmentation provides the x- and y- coordinates of tumor regions in CRC WSIs - which is necessary for the TSR process (rather than a yes no answer that tumor is present).[190] While a supervised image segmentation approach is promising, acquiring ground truth annotations from pathologists to train the supervised image segmentation model can be biased, expensive, and time consuming, making the training process impractical.

Attempting to find other approaches in order to mitigate the shortcomings of requiring pathologist-provided annotations, the use of a GAN is explored for unsupervised anomaly detection.[191] It is used to identify patterns of pixels that deviate from the established pattern in training images, without the need for high-quality pixel-level annotations from pathologists. This approach is particularly useful in tumor segmentation, as tumor tissue is a type of anomalous colon tissue.[192,193]

The GAN-based anomaly detection algorithm, referred to as GANomaly[163], is a commonly used unsupervised anomaly detection approach. However, the GANomaly approach is based on the deep convolutional GAN (DCGAN)[194], which is not meant for very high-resolution images, like colon WSIs. Different from DCGAN, the progressive GAN, also known as *p*GAN, is specifically designed for high resolution image data[166]. In *p*GAN, two major components, the generator (*G*) and discriminator (*D*), are trained gradually starting from 4 x 4 resolution. Image layers of increasing resolution are incrementally added to *G* and *D,* allowing the model to be progressively trained from 4 x 4 up to 1024 x 1024, increasing by a multiple of two, while keeping all the existing layers trainable during the entire training

process. In addition, to maintain a smooth transition from lower to higher resolutions during the training of *G*, new layers are faded in smoothly while doubling the current resolution of image features using nearest neighbor filtering. A newly added *toRGB* layer with weight α increases linearly from zero to one, which further projects the features to the R(red)G(green)B(blue) color channels. Reversely, another newly added *fromRGB* layer with the same weight α projects the RGB color images to the feature vectors. The features are further faded into a new convolutional layer to halve their resolution using the average pooling strategy. Similarly, a smoothed training process for *D* is performed. This process could downscale the input images to match the requirements for the current image sizes of the network. This unique progressive GAN architecture is able to outperform the other conventional GAN architectures in generating photorealistic high-resolution normal colon WSIs by providing a global view focus on the normal colon histology representation from the entire slide in a relatively lower resolution level, and a local view focus on the detailed nuclei morphology patterns in a relative higher resolution level. Therefore, applying P-CEAD[195] was proposed for CRC tumor segmentation.

4.2 Materials and Methods

The objective of this research is to automate the process of segmenting CRC tumor regions from WSI using P-CEAD. P-CEAD is a distinctive anomaly detection pipeline based on *p*GAN. Its training process consists of three phases (**Figure 4. 2**).[195] In Phase 1, a *p*GAN architecture is trained using an image inpainting technique[196] on normal colon WSIs exclusively, in order to produce photorealistic normal (non-diseased) colon WSIs. This training phase enables *p*GAN to learn a reliable reference distribution of normal colon tissue representations by minimizing the error distance values between the input real normal colon WSIs and the generated photorealistic colon WSIs. Since not all pixels in a WSI are part of the tissue regions, the *Otsu*[197] method was used to identify these regions and extract image patches from them. Image patches were extracted from tissue regions on WSIs in 1024 x 1024 pixels, then down

sampled to 512 x 512, 256 x 256, 8 x 8, and 4 x 4 pixels. The training data is saved in TFRecord files,[141] with each file containing binary image patch tensors and the corresponding file name, height, width, and number of channels for each patch respective to different resolution levels. After completing phase 1 of the training, the weights of *p*GAN are frozen.

The goal of phase 2 in the training process is to calculate NERD. NERD is a multivariate gaussian distribution of the absolute errors, also known as reconstruction errors, between the input real WSIs and the generated photorealistic WSIs. Because, during phase 1, *p*GAN is only trained on normal colon WSIs, the absolute errors between the input real normal colon WSIs and the generated photorealistic normal colon WSIs should be small. The reconstruction errors between the input real CRC and the generated photorealistic CRC WSIs are expected to be relatively large because the GAN never learned how to encode features present in anomalous tissues and is therefore more prone to create higher reconstruction errors.

During phase 3 of the training, the NERD and reconstruction errors are used to calculate pixel-level Mahalanobis distances. The goal of this phase is to identify a cut-off threshold to distinguish between normal and CRC tumor pixels in a WSI. If the Mahalanobis distance for a given pixel is higher than the threshold, it is considered an abnormal colon pixel; otherwise, it is considered a normal colon pixel.

After completing all three phases of training, the *p*GAN model was fed 1024 x 1024 resolution image patches extracted from tissue regions on a test set of WSI containing CRC. From this, the reconstruction errors between the input and generated images from the trained *p*GAN were calculated and binarized based on the Mahalanobis distance threshold. Using the *shapely* package for Python,[196] polygon objects were created around the identified CRC tumor pixels and saved into a GeoPandas dataframe.[198] The

comparison between predicted and pathologist-annotated CRC tumor polygon objects were used to calculate a confusion matrix, including pixel-level counts of TP, FP, TN, and FN areas of the WSI. TP was defined as the number of pixels of the areas that are within both the predicted and annotated CRC tumor polygons. FP was defined as the number of pixels of the areas that are within the predicted CRC tumor polygons but are not within the annotated CRC tumor polygons. TN was defined as the number of pixels of the areas that are outside of both the predicted and annotated CRC tumor polygons. FN was defined as the number of pixels of the areas that are outside of the predicted CRC tumor polygons but are within the annotated CRC tumor polygons. Sensitivity, specificity, and F1 score are derived from these values to provide a quantitative measurement of the model performance. The codebase, including the training and inference pipeline, is publicly available via https://github.com/quincy-125/tsr_crc_tumor_seg.

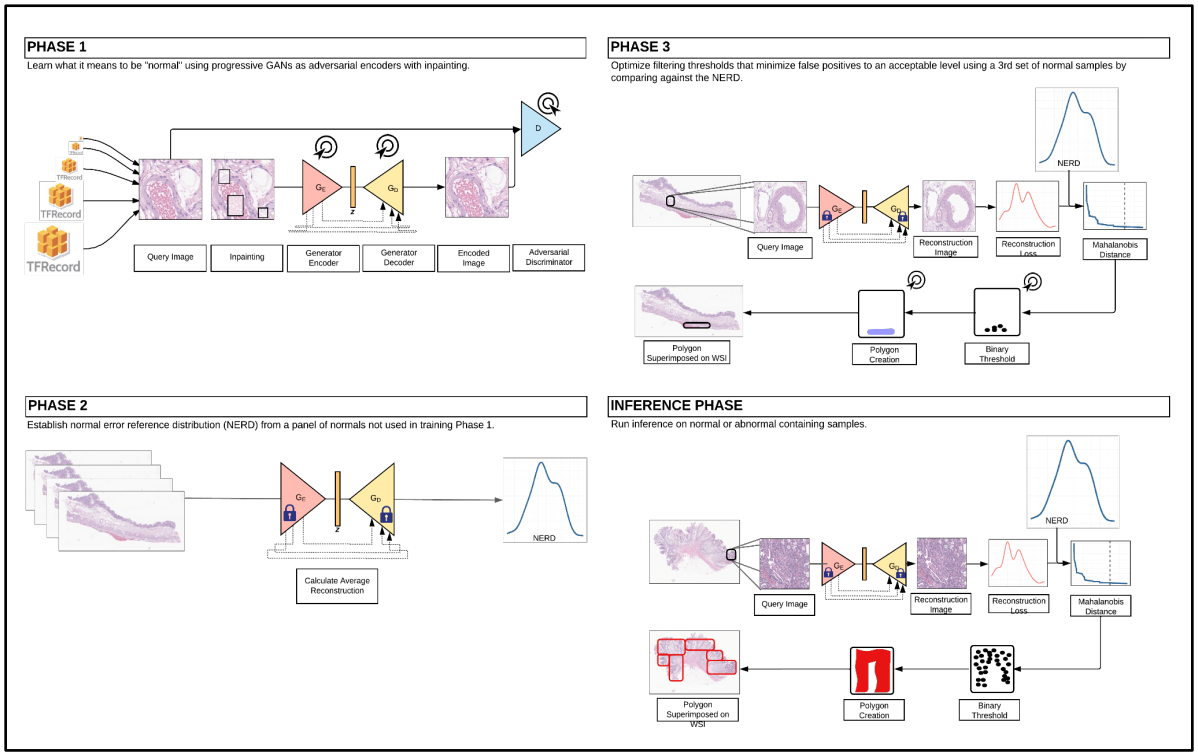


Figure 4. 2 *Training and Inference Pipeline Diagram of P-CEAD in CRC Tumor Segmentation. Phase 1). Phase 1 of the Training Pipeline, pGAN Training; Phase 2). Phase 2 of the Training Pipeline,*

Calculating NERD; Phase 3). Phase 3 of the Training Pipeline, Selecting Cut-Off Mahalanobis Distance Threshold; Inference Phase). Evaluating P-CEAD performance in CRC Tumor Segmentation.

A total of 277 WSIs scanned by the Aperio GT450 scanner[199] at the Mayo Clinic were used for training and inference (**Table 4. 1**). Out of these, 140 were normal colon WSIs and 137 were CRC WSIs. All WSIs underwent quality control examination by a senior cytotechnologist and a senior anatomic pathologist. During the training process, 140 normal colon WSIs were used. Out of these, 100 were used for Phase 1, 20 were used for Phase 2, and the remaining 20 were used for Phase 3. Model inference was performed using all 137 CRC WSIs. The manual annotations of CRC tumors were required to compute the statistical metrics (i.e., confusion matrix, sensitivity, specificity, and accuracy). Tumor annotations from all 137 CRC WSIs were drawn by pathologists using QuPath.[173]

Table 4. 1 *Data Information Summary Table with WSI Type and Number of WSIs Information Regarding Each of the Three Training Phases and One Inference Phase.*

Phase Name	Slide Type	Number of Slides
Training Phase 1	Normal Colon	100
Training Phase 2	Normal Colon	20
Training Phase 3	Normal Colon	20
Inference Phase	Colorectal Cancer	137

4.3 Results and Discussions

The sensitivity, specificity, and accuracy were calculated based on the confusion matrix values for each of the 137 CRC tumor WSI inference results. The proposed P-CEAD based CRC tumor segmentation model achieved 71% $\pm$ 26% sensitivity, 92% $\pm$ 7% specificity, and 63% $\pm$ 23% F1 score (**Figure 4. 3**).

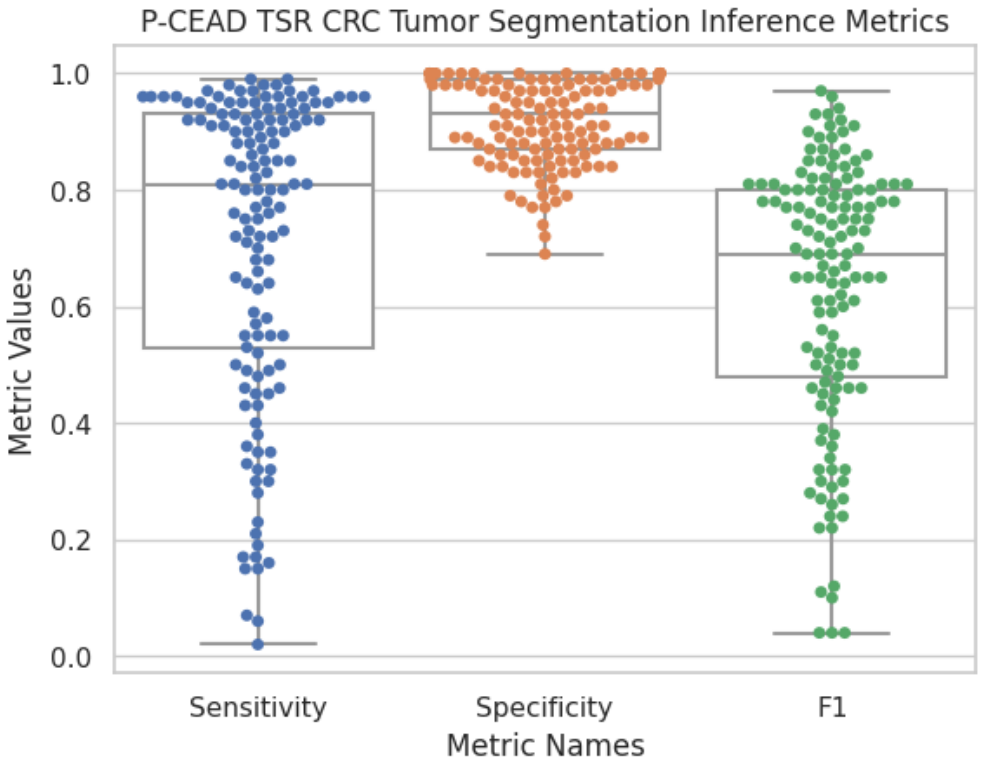


Figure 4. 3 *Quantitative Measurement Results of P-CEAD Inference Performance in CRC Tumor Segmentation on 137 CRC Tumor WSIs. The Statistical Metrics Including the Sensitivity, Specificity, and F1 Score. Each CRC WSI is a blue dot.*

4.3.1 Benefits of Applying Unsupervised Tumor Segmentation Approach

A notable advantage of the P-CEAD-based CRC tumor segmentation pipeline is its fully unsupervised nature. This eliminates the need for time-consuming and costly pathologist annotations during the training process, underlining one of the benefits of implementing the unsupervised P-CEAD approach for CRC tumor segmentation in WSI.

4.3.2 Qualitative Evaluation of P-CEAD based CRC Tumor Segmentation Performance

4.3.2.1 Impacts of Whitespace of WSIs on Model Performance Evaluation

The ground truth CRC tumor annotation is performed by a pathologist, and is a microscopic tumor-level type annotation, with considerable non-tumor areas surrounding the main lesion, inclusive of whitespace

regions. These anomalies are a source of model error since our model focuses only on tissue-containing patches, excluding the whitespace regions. Consequently, the manually annotated CRC tumor areas (TP) tend to be larger than the predicted tumor areas (TP+FP), leading to an increase in false negative predictions (**Figure 4. 4 A**). One potential solution could be to remove whitespace regions from the manually annotated areas to reduce the false negatives in future iterations.

4.3.2.2 Impacts of Predicted Artifacts on Model Performance Evaluation

Originally designed as an anomaly detection model, P-CEAD identifies all regions diverging from the norm, which includes inked tissue, inflamed tissue, and malignant areas from WSI. This could lead the model to classify artifacts such as on-slide annotations as anomalies, thereby increasing the false positive predictions (**Figure 4. 4 B**). To mitigate this, we propose adopting Jiang et al.'s[108] ink-removal technique as part of the data preprocessing procedure before model inference in future experiments.

4.3.2.3 Impacts of Predicted Non-Malignant Tumor Anomalies on Model Performance Evaluation

In our P-CEAD-based model, peritumoral changes were included in the predicted CRC tumor areas. As discussed earlier, P-CEAD aims to detect all anomalous tissues, not solely malignant CRC tumors. Hence, the model included benign stromal tissue connected to malignant CRC tumors within the predicted areas, a factor contributing to false positives. For model training, we relied on normal colon WSIs (**section 4. 2**). A potential amendment could be to introduce benign tissues into the training set to adjust the Normalized Error Rate Difference (NERD), thereby reducing false positive predictions from non-malignant tissues (**Figure 4. 4 C**)

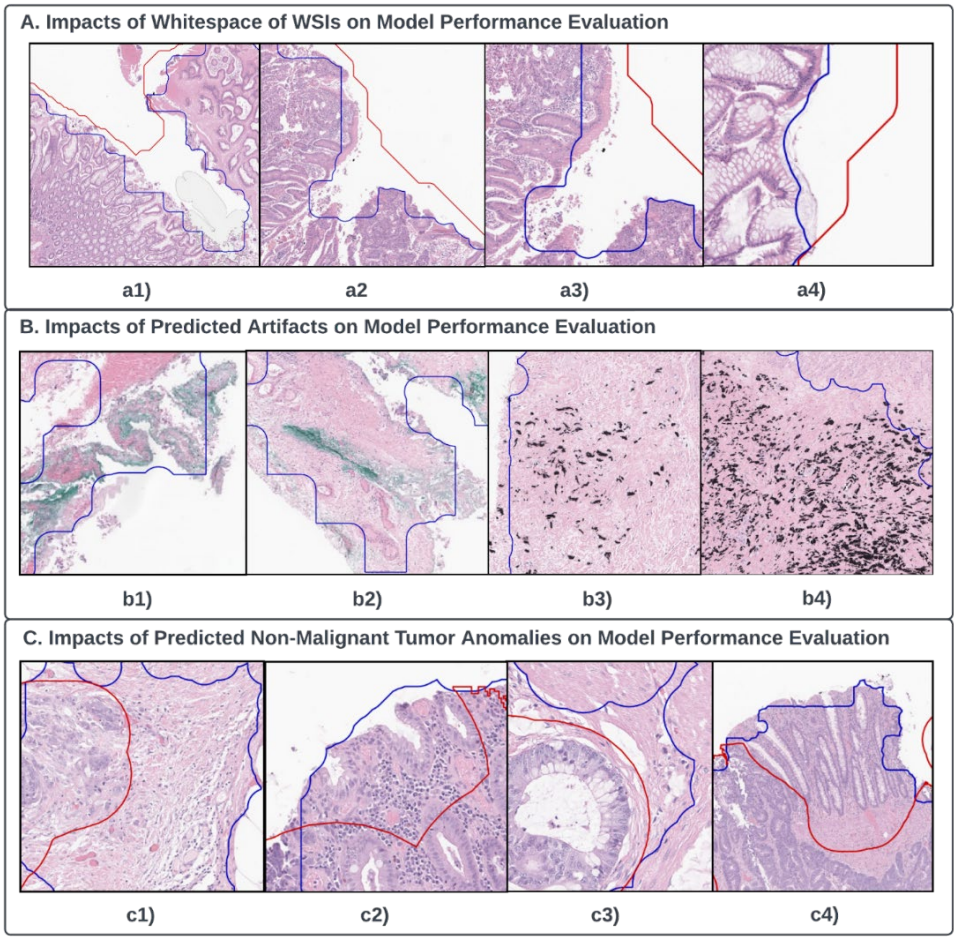


Figure 4. 4 *Qualitative Model Performance Evaluations. A). Impacts of whitespace of WSIs on model performance evaluation with a1) - a4) four example patches. All whitespace areas presented on a1) - a4) are all included in manual CRC tumor annotation regions, but not included in the model prediction regions. B). Impacts of predicted artifacts on model performance evaluation with b1) - b4) four example patches. b1) and b2) are example patches with green on-slide annotation inks that are within the model prediction regions, but outside the manual CRC tumor annotation regions. b3) and b4) are example patches with black on-slide annotation inks that are within the model prediction regions, but outside the manual CRC tumor annotation regions. C). Impacts of predicted non-malignant CRC tumor anomalous tissue on model performance evaluation with c1) - c4) four example patches. On each of the four example patches, tissues on the left to the red polygon boundary line are included in the manual CRC tumor annotations; tissues on the right to the red polygon boundary line are not included in the manual CRC tumor annotations but included in the model prediction regions.*

In summary, our P-CEAD model, an unsupervised anomaly detection-based tumor segmentation approach, yielded 71% ± 26% sensitivity, 92% ± 7% specificity, and 63% ± 23% F1 Score in segmenting CRC tumors from WSI. This underscores the value in further exploration of the P-CEAD-based tumor segmentation algorithm in other cancer types. To optimize model performance, we recommend adding WSIs with artifacts or non-malignant CRC tumor anomalous tissue to the training data set. This could reduce the misclassification of such tissues as malignant CRC tumors when utilizing the anomaly detection approach of P-CEAD. Further, image preprocessing approaches such as ink-removal and whitespace removal using the Otsu method could enhance both quantitative (i.e., reducing false positives and negatives) and qualitative model performance.

4.4 Publications, Contributions, and Declarations:

Chapter modified with permission from the following article which is publicly available on medRxiv (submitted to the Journal of Pathology Informatics, currently under peer-review process):

Gu Q, Meroueh C, Levernier JG, Kroneman TN, Flotte TJ, Hart SN. Using an Anomaly Detection Approach for the Segmentation of Colorectal Cancer Tumors in Whole Slide Images. Published online July 18, 2023:2023.07.17.23292768.
doi:10.1101/2023.07.17.23292768

QG, JL implemented the training and inference pipelines of P-CEAD and did all the experiments. CM, TK, and TF provided the data, and annotations for tumor ground-truth. QG and SNH secured funding. TF and SNH oversee the project.

This study is funded by the Mayo Clinic Digital Pathology Program, and the University of Minnesota Graduate School Doctoral Dissertation Fellowship for the year of 2022-2023.

The authors have declared that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

The authors have declared that ChatGPT was used to assist language editing in the writing process. The use of ChatGPT was done with the authors oversight, control, and was carefully reviewed and edited by the authors.

CHAPTER 5: CONCLUSION

5.1 Summary

The field of digital pathology has witnessed extensive utilization of AI, particularly DL techniques. The implementation of DL-based CAD tools holds great potential in assisting pathologists with cancer diagnosis on WSIs. It is crucial to establish efficient communication that effectively bridges the knowledge gaps between pathologists and computational scientists. This communication plays a pivotal role in the successful development of diagnostic tools on WSIs using DL approaches.

In this dissertation, the first significant contribution lies in the proposal of standardizing seven common tasks in WSI analysis. These tasks encompass image classification, segmentation, visualization, generation, likelihood measurement, object localization, and counting. The proposed standardizations address various aspects, such as the level of supervision, size and diversity of the training dataset, approaches to tackle imbalanced datasets, validation experimental design, and model performance evaluation metrics. By establishing these standardizations, a closer collaboration between pathologists and computational scientists can be fostered, leading to the development of effective AI algorithms for digital pathology.

The identification of the significant impacts of hyperparameter configurations and dataset variations on the binary breast cancer classification performance is a notable contribution. The observation that the digital pathology specialized model architecture, CLAM, exhibited varying classification performance on different datasets (i.e., TCGA and BACH datasets) emphasizes the importance of qualitative assessments of the dataset in the selection of model architectures. Additionally, the concept of the level of supervision (**section 1.3.1.3.1**), emerges as another crucial factor to consider when selecting model architectures.

Non-specialized model architectures (**section 2.2.2.1 - 2.2.2.2**), which represent fully supervised approaches, are preferred when pixel-level annotations are available. Alternatively, the digital pathology specialized model architecture known as CLAM (**section 2.2.2.3**), is recommended for application when only slide-level annotations are accessible, as it employs a weakly supervised approach.

The final significant contribution of this dissertation involves the development of the innovative P-CEAD model and its subsequent application in CRC tumor segmentations. Initially designed to identify various anomalous tissue types in melanoma WSIs, P-CEAD underwent a shift in its quantitative and qualitative assessments. These assessments focused on evaluating the model performance in melanoma tumor segmentations on WSIs since pathologists provided annotations solely for malignant melanoma tumors. The remarkable quantitative- (i.e., 94% $\pm$ 8% sensitivity, 87% $\pm$ 7% specificity, and 89% $\pm$ 7% accuracy) and qualitative- melanoma tumor segmentation performance of P-CEAD promoted its application in CRC tumor segmentations. Furthermore, qualitative evaluations by pathologists regarding the generated images of normal skin, colon, and lung tissues aided computational scientists in identifying challenges related to mode collapse during model training phases (**section 3.3 and section 5.2.2.1**). Pathologists-led qualitative assessments of P-CEAD-based CRC tumor segmentation model performance (**section 4.3.2**) provided insights into the reasons behind the false predictions (FP + FN), leading to targeted solutions to address these challenges and further improve the model performance (**section 4.3.2**). Thus, bridging the knowledge gaps between pathologists and computational scientists is critical in identifying and surmounting model development challenges, ultimately contributing to the successful development of clinical valuable CAD tools.

5.2 Lessons Learned and Future Directions

5.2.1 Lessons Learned and Future Directions for Study Summarized in Chapter 2

5.2.1.1 Insufficient Number of Digital Pathology-Specialized Classifiers Used to Compared with Non-Specialized Classifiers

In the experiments discussed in **section 2.2.2**, CLAM is the only digital pathology-specialized model architecture used to compare with five different non-specialized model architectures, including DenseNet201, InceptionV3, ResNet152, and VGG19 with transfer learning approach, and one-shot learning. There are other digital pathology-specialized image classifiers other than the CLAM, including the recalibrated multi-instance DL-based classifier (RMDL),[200] the weakly supervised based fast and effective classifier (WeaklyFEC),[201] the self-supervised contrastive learning-based classifier (SSCDP),[202] and the transformer-based pathology image classifier (TransDPC).[203] Adding more digital pathology-specialized classification model architectures to compare with the non-specialized model architectures could be considered for further improvements of this study.

5.2.1.2 Insufficient Number of Datasets Used in Comparison Experiments

Dataset variation is an important factor to be considered when selecting an appropriate image classification model architecture. This is a major conclusion stated in **section 2.3.4.** However, the BACH dataset is the only dataset used to conduct experiments in **Chapter 2**. There are four other widely used publicly available digital pathology WSIs dataset,[204] including the Cancer Digital Slide Archive (CDSA),[205] the Camelyon database,[206] including Camelyon16,[41] and Camelyon17[40] datasets, the TUPAC16 dataset,[207] and the Kimia Path24[208] database. Comparing the non-specialized with digital pathology-specialized classification model architectures on a different dataset other than the BACH dataset could be considered for further improvements of this study.

5.2.2 Lessons Learned and Future Directions for Study Summarized in Chapter 3

5.2.2.1 Mode Collapse in GAN Training

Mode collapse (**Figure 5. 1**), a scenario where the generator in the GAN architecture could only generate one or a limited type(s) of output, is a critical challenge in training GANs.[209–211] Approaches including the image inpainting[196] and applying Wasserstein loss[212,213] have already been experimented in the study discussed in **section 3.2.2** to successfully addressed the mode collapse challenges (**Figure 5. 2**) on melanoma use case. However, when extending the P-CEAD approach from detecting anomalies on melanoma WSIs to detect anomalies on lung cancer WSIs, mode collapse challenge was retained and resulted in a failed experiment on lung cancer use case (**Figure 5. 3**).

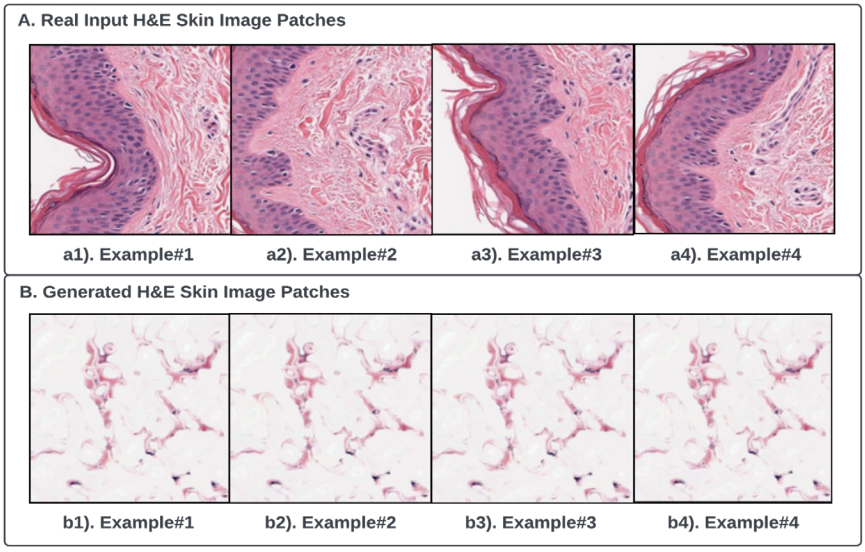


Figure 5. 1 *Visualization of Mode Collapse Examples of P-CEAD on Normal Skin WSIs. A). Real input H&E normal skin image patches used for P-CEAD training with a1)-a4) four example patches; B). Generated H&E normal skin image patches by P-CEAD during the training phases with b1)-b4) four example patches.*

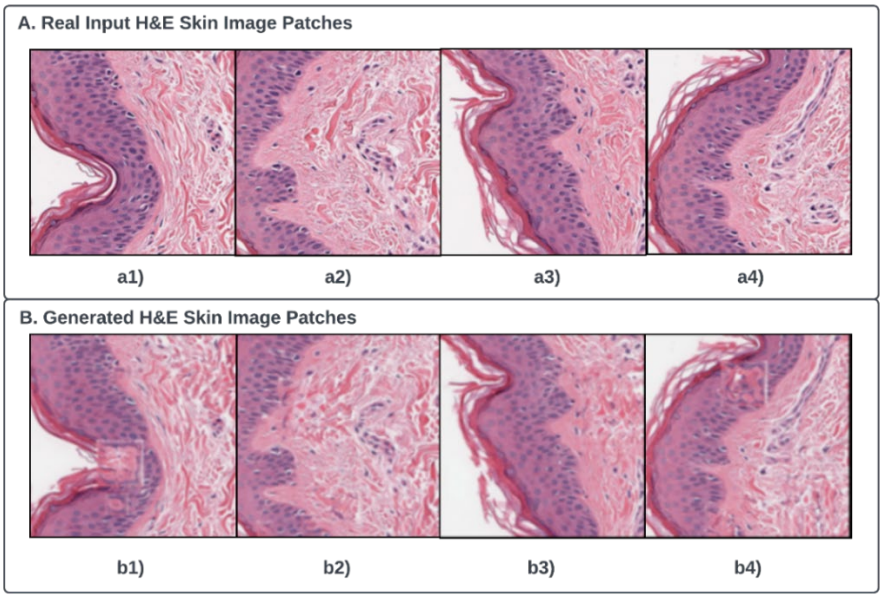


Figure 5. 2 *Visualization of real and successful generated example patches of P-CEAD on Normal Skin WSIs with the mode collapse challenge addressed. A). Real input H&E normal skin image patches used for P-CEAD training with a1)-a4) four example patches; B). Generated H&E normal skin image patches by P-CEAD during the training phases with b1)-b4) four example patches.*

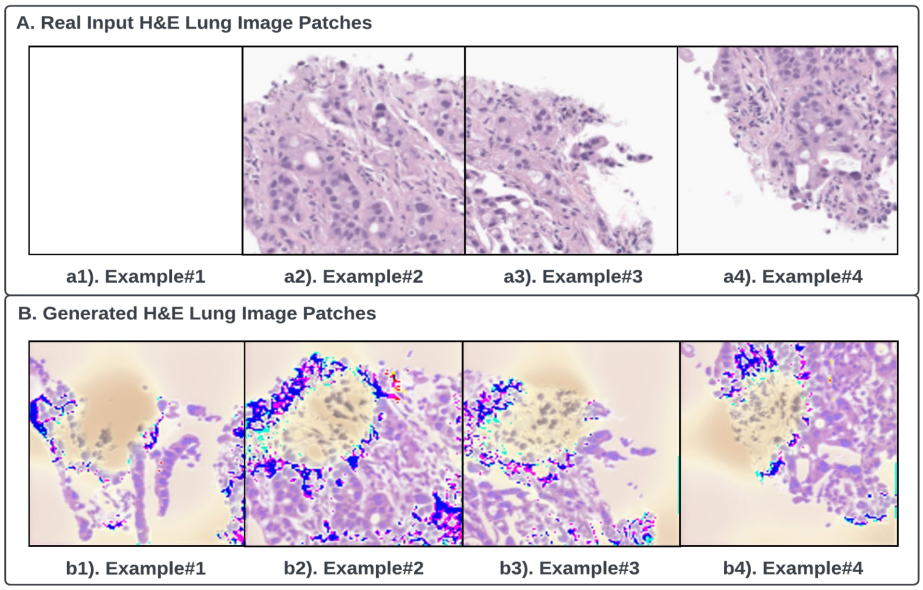


Figure 5. 3 *Visualization of Mode Collapse Examples of P-CEAD on Normal Lung WSIs. A). Real input H&E normal lung image patches used for P-CEAD training with a1)-a4) four example patches; B). Generated H&E normal lung image patches by P-CEAD during the training phases with b1)-b4) four example patches.*

To better address the mode collapse challenge in GAN training, investigating the alternative solutions by leveraging some latest research on addressing mode collapse, including Langevin Stein variational gradient descent,[214] manifold-guided training,[215] training multiple generators with Wasserstein GAN,[216] and Unrolled GAN,[217] could be considered. Alternatively, diffusion models, another type of generative AI, are proven to have a better performance than GAN in image synthesis, and not suffer from the mode collapse.[218–220] These approaches could be considered to further help address the mode collapse challenge, therefore, allowing extending the use cases of the P-CEAD based anomaly detection approach to cancer cases other than the melanoma, and CRC.

5.2.2.2 Challenges in Acquiring Fully Accurate Ground Truth Anomaly Annotations

There are a large number of skin conditions, melanoma is only one subtype of skin lesions.[221–224] Acquiring the ground truth annotations on WSIs with all different types of skin lesions from pathologists is impractical (i.e., time-consuming and expensive). However, the P-CEAD is an anomaly detection model. It is expected to detect all types of skin lesions on WSIs that are not limited to melanoma. The quantitative evaluation of the model by comparing the predicted anomaly regions with the ground truth melanoma annotation areas could result in more false positive predictions. However, those false positive predictions are true skin anomalous tissue, but not melanoma. To better evaluate the anomaly detection performance of the P-CEAD on skin WSIs, given the normal skin WSIs for model inference could be considered as an alternative approach to ensure the quality of the P-CEAD's anomaly detection performance.

5.2.3 Lessons Learned and Future Directions for Study Summarized in Chapter 4

5.2.3.1 Exploration of Different Model Architectures for CRC Tumor Segmentations

Tumor segmentation on WSIs is a standard task in the field of digital pathology. There are many other published works on tumor segmentation from WSIs other than the P-CEAD based tumor segmentation approach. Conventional DL approaches, including CNN,[225] DCNN,[226] fully CNN (FCNN),[227,228] U-Net,[227,229–231] a hybrid model architecture integrating the support vector machine (SVM) and CNN,[232] and multi-resolution encoder-decoder network[233] have been applied on segmenting regions of interest on WSIs.[204] In addition, more complex approaches, including the reinforced auto-zoom network (RAZN),[234] and multi-instance learning with attention mechanism,[235,236] have been leveraged to perform image segmentation tasks on WSIs.[204] These conventional or complex approaches could be considered as alternative approaches in addition to the proposed P-CEAD based tumor segmentation approach for further explorations.

5.2.3.2 Reducing the Needs of Image Registration in Tissue Slide Review Process

The TSR procedure requires cytotechnologists to scrape tumor tissue on the non-stained tissue slide based on the tumor annotation made on the H&E-stained adjacent tissue slide, which will require the additional image registration[237] (alignment) efforts. The challenge of image registration remained even with the proposed automated P-CEAD based CRC tumor segmentation approach. Classical image registration approaches, including the Scale Invariant Feature Transform (SIFT),[238] Speed-Up Robust Features (SURF),[239] Oriented FAST and Rotated BRIEF (ORB),[240] KAZE (i.e., a novel multiscale 2D feature detection and description algorithm in nonlinear scale spaces),[241] and Fast SIFT (F-SIFT),[242] which are robust and have been commonly applied in various fields.[243] Additionally, a most recent research proposed by Hoque et al.,[243] a multi-stained feature matching based WSI registration approach, is specifically targeting the histological image registration. To tackle the remaining image registration

challenge in the TSR procedure, combining the automated tumor segmentation model with an appropriate automated image registration approach, could be considered to build a fully automated TSR pipeline.

5.3 Publications, Contributions, and Declarations:

Chapter 5 is not modified from any published journal articles or working in progress manuscripts. This chapter was written by QG.

The completion of Chapter 5 is fully funded by the University of Minnesota Graduate School Doctoral Dissertation Fellowship for the year of 2022-2023.

QG declared that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this chapter.

QG declared that ChatGPT was used to assist language editing in the writing process. The use of ChatGPT was done with QG's oversight, control, and was carefully reviewed and edited by QG.

Bibliography

1. Louw DF, Sutherland GR, Schulder M. From microscopic to astronomic, the legacy of Carl Zeiss. *Neurosurgery*. 2003;52(3):668-674; discussion 672-674. doi:10.1227/01.neu.0000048477.49196.50
2. Schultz M. Rudolf Virchow. *Emerg Infect Dis*. 2008;14(9):1480-1481. doi:10.3201/eid1409.086672
3. Abels E, Pantanowitz L, Aeffner F, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J Pathol*. 2019;249(3):286-294. doi:10.1002/path.5331
4. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. *Histopathology*. 2012;61(1):1-9. doi:10.1111/j.1365-2559.2011.03814.x
5. Eloy C, Vale J, Curado M, et al. Digital Pathology Workflow Implementation at IPATIMUP. *Diagnostics*. 2021;11(11):2111. doi:10.3390/diagnostics11112111
6. Li X, Li C, Rahaman MM, et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artif Intell Rev*. 2022;55(6):4809-4878. doi:10.1007/s10462-021-10121-0
7. Parwani AV. Next generation diagnostic pathology: use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. *Diagn Pathol*. 2019;14(1):138. doi:10.1186/s13000-019-0921-2
8. Patel A, Balis UGJ, Cheng J, et al. Contemporary Whole Slide Imaging Devices and Their Applications within the Modern Pathology Department: A Selected Hardware Review. *J Pathol Inform*. 2021;12:50. doi:10.4103/jpi.jpi_66_21
9. Aeffner F, Zarella MD, Buchbinder N, et al. Introduction to Digital Image Analysis in Whole-slide Imaging: A White Paper from the Digital Pathology Association. *J Pathol Inform*. 2019;10:9. doi:10.4103/jpi.jpi_82_18
10. Parwani AV, Patel A, Zhou M, et al. An update on computational pathology tools for genitourinary pathology practice: A review paper from the Genitourinary Pathology Society (GUPS). *J Pathol Inform*. 2022;14:100177. doi:10.1016/j.jpi.2022.100177
11. Vu QD, Graham S, Kurc T, et al. Methods for Segmentation and Classification of Digital Microscopy Tissue Images. *Front Bioeng Biotechnol*. 2019;7:53. doi:10.3389/fbioe.2019.00053
12. Cui M, Zhang DY. Artificial intelligence and computational pathology. *Lab Invest*. 2021;101(4):412-422. doi:10.1038/s41374-020-00514-0
13. Davey MG, Hynes SO, Kerin MJ, Miller N, Lowery AJ. Ki-67 as a Prognostic Biomarker in Invasive Breast Cancer. *Cancers*. 2021;13(17):4455. doi:10.3390/cancers13174455
14. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst*. 2011;103(22):1656-1664. doi:10.1093/jnci/djr393
15. Gibson-Corley KN, Olivier AK, Meyerholz DK. Principles for valid histopathologic scoring in research. *Vet Pathol*. 2013;50(6):1007-1015. doi:10.1177/0300985813485099
16. Patel AU, Mohanty SK, Parwani AV. Applications of Digital and Computational Pathology and Artificial Intelligence in Genitourinary Pathology Diagnostics. *Surg Pathol Clin*. 2022;15(4):759-785. doi:10.1016/j.path.2022.08.001
17. Besson S, Leigh R, Linkert M, et al. Bringing Open Data to Whole Slide Imaging. *Digit Pathol 15th Eur Congr ECDP 2019 Warwick UK April 10-13 2019 Proc Eur Congr Digit Pathol 15th 2019 Warwick Engl*. 2019;2019:3-10. doi:10.1007/978-3-030-23937-4_1
18. Khened M, Kori A, Rajkumar H, Krishnamurthi G, Srinivasan B. A generalized deep learning framework for whole-slide image segmentation and analysis. *Sci Rep*. 2021;11(1):11579. doi:10.1038/s41598-021-90444-8

19. Lutnick B, Manthey D, Becker JU, et al. A user-friendly tool for cloud-based whole slide image segmentation with examples from renal histopathology. *Commun Med*. 2022;2(1):1-15. doi:10.1038/s43856-022-00138-z
20. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol*. 2019;20(5):e253-e261. doi:10.1016/S1470-2045(19)30154-8
21. Serag A, Ion-Margineanu A, Qureshi H, et al. Translational AI and Deep Learning in Diagnostic Pathology. *Front Med*. 2019;6:185. doi:10.3389/fmed.2019.00185
22. Tizhoosh HR, Pantanowitz L. Artificial Intelligence and Digital Pathology: Challenges and Opportunities. *J Pathol Inform*. 2018;9:38. doi:10.4103/jpi.jpi_53_18
23. Wahab N, Miligy IM, Dodd K, et al. Semantic annotation for computational pathology: multidisciplinary experience and best practice recommendations. *J Pathol Clin Res*. 2022;8(2):116-128. doi:10.1002/cjp2.256
24. Asif A, Rajpoot K, Snead D, Minhas F, Rajpoot N. Towards Launching AI Algorithms for Cellular Pathology into Clinical & Pharmaceutical Orbits. Published online December 17, 2021. doi:10.48550/arXiv.2112.09496
25. Kim E, Bitterman DS, Kann BH, et al. Hidden in Plain Sight: Clinical Informaticians are the Oncology Subspecialists You Did Not Know You Needed. *Clin Oncol R Coll Radiol G B*. 2022;34(2):135-140. doi:10.1016/j.clon.2021.11.018
26. Berbís MA, McClintock DS, Bychkov A, et al. Computational pathology in 2030: a Delphi study forecasting the role of AI in pathology within the next decade. *EBioMedicine*. 2023;88:104427. doi:10.1016/j.ebiom.2022.104427
27. Steiner DF, Chen PHC, Mermel CH. Closing the translation gap: AI applications in digital pathology. *Biochim Biophys Acta Rev Cancer*. 2021;1875(1):188452. doi:10.1016/j.bbcan.2020.188452
28. Jackson BR, Ye Y, Crawford JM, et al. The Ethics of Artificial Intelligence in Pathology and Laboratory Medicine: Principles and Practice. *Acad Pathol*. 2021;8:2374289521990784. doi:10.1177/2374289521990784
29. Cheng JY, Abel JT, Balis UGJ, McClintock DS, Pantanowitz L. Challenges in the Development, Deployment, and Regulation of Artificial Intelligence in Anatomic Pathology. *Am J Pathol*. 2021;191(10):1684-1692. doi:10.1016/j.ajpath.2020.10.018
30. Försch S, Klauschen F, Hufnagl P, Roth W. Artificial Intelligence in Pathology. *Dtsch Arzteblatt Int*. 2021;118(12):194-204. doi:10.3238/arztebl.m2021.0011
31. Nakagawa K, Moukheiber L, Celi LA, et al. AI in Pathology: What could possibly go wrong? *Semin Diagn Pathol*. 2023;40(2):100-108. doi:10.1053/j.semdp.2023.02.006
32. Oliveira SP, Neto PC, Fraga J, et al. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Sci Rep*. 2021;11(1):14358. doi:10.1038/s41598-021-93746-z
33. Drogt J, Milota M, Vos S, Bredenoord A, Jongsma K. Integrating artificial intelligence in pathology: a qualitative interview study of users' experiences and expectations. *Mod Pathol Off J U S Can Acad Pathol Inc*. 2022;35(11):1540-1550. doi:10.1038/s41379-022-01123-6
34. Shinozuka M, Mansouri B. Synthetic aperture radar and remote sensing technologies for structural health monitoring of civil infrastructure systems. In: Elsevier; 2009:113-151. doi:10.1533/9781845696825.1.114
35. Chen H, Bai F, Wang M, Zhang M, Zhang P, Wu K. The prognostic significance of co-existence ductal carcinoma in situ in invasive ductal breast cancer: a large population-based study and a matched case-control analysis. *Ann Transl Med*. 2019;7(18):484. doi:10.21037/atm.2019.08.16
36. Logullo AF, Godoy AB, Mourão-Neto M, Simpson AJG, Nishimoto IN, Brentani MM. Presence of ductal carcinoma in situ confers an improved prognosis for patients with T1N0M0 invasive breast carcinoma. *Braz J Med Biol Res Rev Bras Pesqui Medicas E Biol*. 2002;35(8):913-919.

doi:10.1590/s0100-879x2002000800008
37. Chen CL, Chen CC, Yu WH, et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nat Commun*. 2021;12(1):1193. doi:10.1038/s41467-021-21467-y
38. Dimitriou N, Arandjelović O, Caie PD. Deep Learning for Whole Slide Image Analysis: An Overview. *Front Med*. 2019;6:264. doi:10.3389/fmed.2019.00264
39. Otálora S, Marini N, Müller H, Atzori M. Combining weakly and strongly supervised learning improves strong supervision in Gleason pattern classification. *BMC Med Imaging*. 2021;21(1):77. doi:10.1186/s12880-021-00609-0
40. Bandi P, Geessink O, Manson Q, et al. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Trans Med Imaging*. 2019;38(2):550-560. doi:10.1109/TMI.2018.2867350
41. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585
42. Ghazvinian Zanjani F, Zinger S, de With PHN. Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces. 2018;10581:105810I. doi:10.1117/12.2293107
43. Kong B, Wang X, Li Z, Song Q, Zhang S. Cancer Metastasis Detection via Spatially Structured Deep Network. In: Niethammer M, Styner M, Aylward S, et al., eds. *Information Processing in Medical Imaging*. Lecture Notes in Computer Science. Springer International Publishing; 2017:236-248. doi:10.1007/978-3-319-59050-9_19
44. Li Y, Ping W. Cancer Metastasis Detection With Neural Conditional Random Field. Published online June 19, 2018. doi:10.48550/arXiv.1806.07064
45. Liu Y, Gadepalli K, Norouzi M, et al. Detecting Cancer Metastases on Gigapixel Pathology Images. Published online March 7, 2017. doi:10.48550/arXiv.1703.02442
46. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep Learning for Identifying Metastatic Breast Cancer. Published online June 18, 2016. doi:10.48550/arXiv.1606.05718
47. Faust K, Bala S, van Ommeren R, et al. Publisher Correction: Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. *Nat Mach Intell*. 2020;2(4):237-237. doi:10.1038/s42256-019-0133-1
48. Faust K, Xie Q, Han D, et al. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. *BMC Bioinformatics*. 2018;19(1):173. doi:10.1186/s12859-018-2184-4
49. Roohi A, Faust K, Djuric U, Diamandis P. Unsupervised Machine Learning in Pathology: The Next Frontier. *Surg Pathol Clin*. 2020;13(2):349-358. doi:10.1016/j.path.2020.01.002
50. Tellez D, Litjens G, van der Laak J, Ciompi F. Neural Image Compression for Gigapixel Histopathology Image Analysis. *IEEE Trans Pattern Anal Mach Intell*. 2021;43(2):567-578. doi:10.1109/TPAMI.2019.2936841
51. Klein C, Zeng Q, Arbaretaz F, et al. Artificial intelligence for solid tumour diagnosis in digital pathology. *Br J Pharmacol*. 2021;178(21):4291-4315. doi:10.1111/bph.15633
52. brief introduction to weakly supervised learning | National Science Review | Oxford Academic. Accessed June 28, 2023. https://academic.oup.com/nsr/article/5/1/44/4093912
53. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301-1309. doi:10.1038/s41591-019-0508-1
54. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555-

570. doi:10.1038/s41551-020-00682-w
55. Anklin V, Pati P, Jaume G, et al. Learning Whole-Slide Segmentation from Inexact and Incomplete Labels Using Tissue Graphs. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II*. Springer-Verlag; 2021:636-646. doi:10.1007/978-3-030-87196-3_59
56. Cserni B, Bori R, Csörgő E, et al. ONEST (Observers Needed to Evaluate Subjective Tests) suggests four or more observers for a reliable assessment of the consistency of histological grading of invasive breast carcinoma: A reproducibility study with a retrospective view on previous studies. *Pathol Res Pract*. 2022;229:153718. doi:10.1016/j.prp.2021.153718
57. Mun Y, Paik I, Shin SJ, Kwak TY, Chang H. Yet Another Automated Gleason Grading System (YAAGGS) by weakly supervised deep learning. *NPJ Digit Med*. 2021;4:99. doi:10.1038/s41746-021-00469-6
58. Jiang Y, Sui X, Ding Y, Xiao W, Zheng Y, Zhang Y. A semi-supervised learning approach with consistency regularization for tumor histopathological images analysis. *Front Oncol*. 2023;12:1044026. doi:10.3389/fonc.2022.1044026
59. Neto PC, Oliveira SP, Montezuma D, et al. iMIL4PATH: A Semi-Supervised Interpretable Approach for Colorectal Whole-Slide Images. *Cancers*. 2022;14(10):2489. doi:10.3390/cancers14102489
60. Tanha J, van Someren M, Afsarmanesh H. Semi-supervised self-training for decision tree classifiers. *Int J Mach Learn Cybern*. 2017;8(1):355-370. doi:10.1007/s13042-015-0328-7
61. Lee J, Liu C, Kim J, et al. Deep learning for rare disease: A scoping review. *J Biomed Inform*. 2022;135:104227. doi:10.1016/j.jbi.2022.104227
62. Amgad M, Atteya LA, Hussein H, et al. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *GigaScience*. 2022;11:giac037. doi:10.1093/gigascience/giac037
63. Gui J, Chen T, Cao Q, Sun Z, Luo H, Tao D. A Survey of Self-Supervised Learning from Multiple Perspectives: Algorithms, Theory, Applications and Future Trends. Published online January 13, 2023. doi:10.48550/arXiv.2301.05712
64. Yann LeCun - Self Supervised Learning | ICLR 2020 - YouTube. Accessed June 28, 2023. https://www.youtube.com/watch?v=8TTK-Dd0H9U
65. Yann LeCun and Yoshua Bengio: Self-supervised learning is the key to human-level intelligence. VentureBeat. Published May 2, 2020. Accessed June 28, 2023. https://venturebeat.com/ai/yann-lecun-and-yoshua-bengio-self-supervised-learning-is-the-key-to-human-level-intelligence/
66. Zhai X, Oliver A, Kolesnikov A, Beyer L. S4L: Self-Supervised Semi-Supervised Learning. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. ; 2019:1476-1485. doi:10.1109/ICCV.2019.00156
67. Chen C, Lu MY, Williamson DFK, Chen TY, Schaumberg AJ, Mahmood F. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nat Biomed Eng*. 2022;6(12):1420-1434. doi:10.1038/s41551-022-00929-8
68. Fashi PA, Hemati S, Babaie M, Gonzalez R, Tizhoosh HR. A self-supervised contrastive learning approach for whole slide image representation in digital pathology. *J Pathol Inform*. 2022;13:100133. doi:10.1016/j.jpi.2022.100133
69. Koohbanani NA, Unnikrishnan B, Khurram SA, Krishnaswamy P, Rajpoot N. Self-Path: Self-Supervision for Classification of Pathology Images With Limited Annotations. *IEEE Trans Med Imaging*. 2021;40(10):2845-2856. doi:10.1109/TMI.2021.3056023
70. The Probable Error of a Mean on JSTOR. Accessed June 28, 2023. https://www.jstor.org/stable/2331554

71. Sagi O, Rokach L. Ensemble learning: A survey. *WIREs Data Min Knowl Discov*. 2018;8(4):e1249. doi:10.1002/widm.1249
72. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Trans Syst Man Cybern Part C Appl Rev*. 2012;42(4):463-484. doi:10.1109/TSMCC.2011.2161285
73. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140. doi:10.1007/BF00058655
74. Breiman L. Stacked regressions. *Mach Learn*. 1996;24(1):49-64. doi:10.1007/BF00117832
75. Schapire RE. A brief introduction to boosting. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'99. Morgan Kaufmann Publishers Inc.; 1999:1401-1406.
76. Fernández-Carrobles MM, Serrano I, Bueno G, Déniz O. Bagging Tree Classifier and Texture Features for Tumor Identification in Histological Images. *Procedia Comput Sci*. 2016;90:99-106. doi:10.1016/j.procs.2016.07.030
77. Kallipolitis A, Revelos K, Maglogiannis I. Ensembling EfficientNets for the Classification and Interpretation of Histopathology Images. *Algorithms*. 2021;14(10):278. doi:10.3390/a14100278
78. Mohammed M, Mwambi H, Mboya IB, Elbashir MK, Omolo B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. *Sci Rep*. 2021;11(1):15626. doi:10.1038/s41598-021-95128-x
79. Shafieian S, Zulkernine M. Multi-layer stacking ensemble learners for low footprint network intrusion detection. *Complex Intell Syst*. Published online July 5, 2022. doi:10.1007/s40747-022-00809-3
80. Liew XY, Hameed N, Clos J. An investigation of XGBoost-based algorithm for breast cancer classification. *Mach Learn Appl*. 2021;6:100154. doi:10.1016/j.mlwa.2021.100154
81. Bakar WAWA, Zuhairi MA, Man M, Jusoh JA, Josdi NLN. Deep learning algorithm vs XGBoost using Wisconsin breast cancer diagnosis. In: *Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)*. Vol 12506. SPIE; 2022:1732-1740. doi:10.1117/12.2663128
82. Kim YG, Song IH, Cho SY, et al. Diagnostic Assessment of Deep Learning Algorithms for Frozen Tissue Section Analysis in Women with Breast Cancer. *Cancer Res Treat*. 2023;55(2):513-522. doi:10.4143/crt.2022.055
83. Siller M, Stangassinger LM, Kreutzer C, et al. On the acceptance of "fake" histopathology: A study on frozen sections optimized with deep learning. *J Pathol Inform*. 2022;13:100168. doi:10.4103/jpi.jpi_53_21
84. Ratnavelu NDG, Brown AP, Mallett S, et al. Intraoperative frozen section analysis for the diagnosis of early stage ovarian cancer in suspicious pelvic masses. *Cochrane Database Syst Rev*. 2016;3(3):CD010360. doi:10.1002/14651858.CD010360.pub2
85. Qin C, Bai Y, Zeng Z, et al. The Cutting and Floating Method for Paraffin-embedded Tissue for Sectioning. *J Vis Exp JoVE*. 2018;(139):58288. doi:10.3791/58288
86. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *Eur Radiol*. 2015;25(4):932-939. doi:10.1007/s00330-014-3487-0
87. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
88. Cree IA, Tan PH, Travis WD, et al. Counting mitoses: SI(ze) matters! *Mod Pathol Off J U S Can Acad Pathol Inc*. 2021;34(9):1651-1657. doi:10.1038/s41379-021-00825-7
89. Ibrahim A, Lashen AG, Katayama A, et al. Defining the area of mitoses counting in invasive breast cancer using whole slide image. *Mod Pathol Off J U S Can Acad Pathol Inc*. 2022;35(6):739-

748. doi:10.1038/s41379-021-00981-w
90. Joo MW, Kang YK, Yoo CY, Cha SH, Chung YG. Prognostic significance of chemotherapy-induced necrosis in osteosarcoma patients receiving pasteurized autografts. *PLoS ONE*. 2017;12(2):e0172155. doi:10.1371/journal.pone.0172155
91. Al-Janabi S, Huisman A, Willems SM, Van Diest PJ. Digital slide images for primary diagnostics in breast pathology: a feasibility study. *Hum Pathol*. 2012;43(12):2318-2325. doi:10.1016/j.humpath.2012.03.027
92. Bautista PA, Hashimoto N, Yagi Y. Color standardization in whole slide imaging using a color calibration slide. *J Pathol Inform*. 2014;5:4. doi:10.4103/2153-3539.126153
93. Bautista PA, Yagi Y. Improving the visualization and detection of tissue folds in whole slide images through color enhancement. *J Pathol Inform*. 2010;1:25. doi:10.4103/2153-3539.73320
94. Geusebroek JM, Cornelissen F, Smeulders AW, Geerts H. Robust autofocusing in microscopy. *Cytometry*. 2000;39(1):1-9.
95. Graham AR, Bhattacharyya AK, Scott KM, et al. Virtual slide telepathology for an academic teaching hospital surgical pathology quality assurance program. *Hum Pathol*. 2009;40(8):1129-1136. doi:10.1016/j.humpath.2009.04.008
96. Hashimoto N, Bautista PA, Yamaguchi M, Ohyama N, Yagi Y. Referenceless image quality evaluation for whole slide imaging. *J Pathol Inform*. 2012;3:9. doi:10.4103/2153-3539.93891
97. Isse K, Lesniak A, Grama K, Roysam B, Minervini MI, Demetris AJ. Digital transplantation pathology: combining whole slide imaging, multiplex staining and automated image analysis. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2012;12(1):27-37. doi:10.1111/j.1600-6143.2011.03797.x
98. McClintock DS, Lee RE, Gilbertson JR. Using computerized workflow simulations to assess the feasibility of whole slide imaging full adoption in a high-volume histology laboratory. *Anal Cell Pathol Amst*. 2012;35(1):57-64. doi:10.3233/ACP-2011-0034
99. Rocha R, Vassallo J, Soares F, Miller K, Gobbi H. Digital slides: present status of a tool for consultation, teaching, and quality control in pathology. *Pathol Res Pract*. 2009;205(11):735-741. doi:10.1016/j.prp.2009.05.004
100. Stathonikos N, Veta M, Huisman A, van Diest PJ. Going fully digital: Perspective of a Dutch academic pathology lab. *J Pathol Inform*. 2013;4:15. doi:10.4103/2153-3539.114206
101. Xiong W, Tian Q, Lim J. An adaptive enhanced focusing technique for whole slide imaging using contextual information. In: *2008 3rd IEEE Conference on Industrial Electronics and Applications*. ; 2008:1837-1840. doi:10.1109/ICIEA.2008.4582837
102. Wilbur DC, Madi K, Colvin RB, et al. Whole-slide imaging digital pathology as a platform for teleconsultation: a pilot study using paired subspecialist correlations. *Arch Pathol Lab Med*. 2009;133(12):1949-1953. doi:10.5858/133.12.1949
103. Yagi Y, Gilbertson JR. A relationship between slide quality and image quality in whole slide imaging (WSI). *Diagn Pathol*. 2008;3(1):S12. doi:10.1186/1746-1596-3-S1-S12
104. Wang Z, Bi Y, Pan T, et al. Targeting tumor heterogeneity: multiplex-detection-based multiple instance learning for whole slide image classification. *Bioinforma Oxf Engl*. 2023;39(3):btad114. doi:10.1093/bioinformatics/btad114
105. Jansen I, Lucas M, Savci-Heijink CD, et al. Three-dimensional histopathological reconstruction of bladder tumours. *Diagn Pathol*. 2019;14(1):25. doi:10.1186/s13000-019-0803-7
106. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *J Big Data*. 2019;6(1):60. doi:10.1186/s40537-019-0197-0
107. Ruusuvuori P, Valkonen M, Kartasalo K, et al. Spatial analysis of histology in 3D: quantification and visualization of organ and tumor level tissue environment. *Heliyon*. 2022;8(1):e08762. doi:10.1016/j.heliyon.2022.e08762

108. Jiang J, Prodduturi N, Chen D, et al. Image-to-image translation for automatic ink removal in whole slide images. *J Med Imaging Bellingham Wash*. 2020;7(5):057502. doi:10.1117/1.JMI.7.5.057502
109. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703-715. doi:10.1038/s41571-019-0252-y
110. News P. Three Ways Digital Pathology Optimizes Multidisciplinary Team Meetings – Pathology News. Published March 6, 2023. Accessed June 28, 2023. https://www.pathologynews.com/digital-pathology/three-ways-digital-pathology-optimizes-multidisciplinary-team-meetings/
111. Patel AU, Shaker N, Mohanty S, et al. Cultivating Clinical Clarity through Computer Vision: A Current Perspective on Whole Slide Imaging and Artificial Intelligence. *Diagn Basel Switz*. 2022;12(8):1778. doi:10.3390/diagnostics12081778
112. Henricks WH, Wilkerson ML, Castellani WJ, Whitsitt MS, Sinard JH. Pathologists as stewards of laboratory information. *Arch Pathol Lab Med*. 2015;139(3):332-337. doi:10.5858/arpa.2013-0714-SO
113. Paige Receives First Ever FDA Approval for AI Product in Digital Pathology | Business Wire. Accessed June 28, 2023. https://www.businesswire.com/news/home/20210922005369/en/Paige-Receives-First-Ever-FDA-Approval-for-AI-Product-in-Digital-Pathology.
114. Raciti P, Sue J, Ceballos R, et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod Pathol Off J U S Can Acad Pathol Inc*. 2020;33(10):2058-2066. doi:10.1038/s41379-020-0551-y
115. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. *CA Cancer J Clin*. 2017;67(1):7-30. doi:10.3322/caac.21387
116. Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. *Nat Rev Dis Primer*. 2019;5(1):1-31. doi:10.1038/s41572-019-0111-2
117. Nounou MI, ElAmrawy F, Ahmed N, Abdelraouf K, Goda S, Syed-Sha-Qhattal H. Breast Cancer: Conventional Diagnosis and Treatment Modalities and Recent Patents and Technologies. *Breast Cancer Basic Clin Res*. 2015;9(Suppl 2):17-34. doi:10.4137/BCBCR.S29420
118. Davidson NE, Rimm DL. Expertise vs evidence in assessment of breast biopsies: an atypical science. *JAMA*. 2015;313(11):1109-1110. doi:10.1001/jama.2015.1945
119. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep Learning for Computer Vision: A Brief Review. *Comput Intell Neurosci*. 2018;2018:7068349. doi:10.1155/2018/7068349
120. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88. doi:10.1016/j.media.2017.07.005
121. Barisoni L, Lafata KJ, Hewitt SM, Madabhushi A, Balis UGJ. Digital pathology and computational image analysis in nephropathology. *Nat Rev Nephrol*. 2020;16(11):669-685. doi:10.1038/s41581-020-0321-6
122. Deng S, Zhang X, Yan W, et al. Deep learning in digital pathology image analysis: a survey. *Front Med*. 2020;14(4):470-487. doi:10.1007/s11684-020-0782-9
123. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci Rep*. 2020;10(1):1504. doi:10.1038/s41598-020-58467-9
124. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3(1):9. doi:10.1186/s40537-016-0043-6
125. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ; 2016:2818-2826. doi:10.1109/CVPR.2016.308
126. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. Published online January 28, 2018. doi:10.48550/arXiv.1608.06993

127. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Published online December 10, 2015. doi:10.48550/arXiv.1512.03385
128. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Published online April 10, 2015. doi:10.48550/arXiv.1409.1556
129. Fei-Fei L. Knowledge transfer in learning to recognize visual objects classes. In: ; 2006. Accessed May 19, 2023. https://www.semanticscholar.org/paper/Knowledge-transfer-in-learning-to-recognize-visual-Fei-Fei/35a198cc4d38bd2db60cda96ea4cb7b12369fd3c
130. Koch GR. Siamese Neural Networks for One-Shot Image Recognition. In: ; 2015. Accessed May 19, 2023. https://www.semanticscholar.org/paper/Siamese-Neural-Networks-for-One-Shot-Image-Koch/f216444d4f2959b4520c61d20003fa30a199670a
131. Jia X. Image recognition method based on deep learning. In: *2017 29th Chinese Control And Decision Conference (CCDC)*. ; 2017:4730-4735. doi:10.1109/CCDC.2017.7979332
132. Bergstra J, Yamins D, Cox DD. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. JMLR.org; 2013:I-115-I-123.
133. *Pattern Recognition and Machine Learning*. Accessed May 19, 2023. https://link.springer.com/book/9780387310732
134. Zhou S, Song W. Deep learning-based roadway crack classification using laser-scanned range images: A comparative study on hyperparameter selection. *Autom Constr*. 2020;114:103171. doi:10.1016/j.autcon.2020.103171
135. Wei S, Shanglian Z. Laser-scanned range image dataset from asphalt and concrete roadways for DCNN-based crack classification. Published online February 17, 2020. Accessed May 19, 2023. https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published//PRJ-2681
136. Heidari M, Mirniaharikandehei S, Khuzani AZ, Danala G, Qiu Y, Zheng B. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. *Int J Med Inf*. 2020;144:104284. doi:10.1016/j.ijmedinf.2020.104284
137. Aresta G, Araújo T, Kwok S, et al. BACH: Grand challenge on breast cancer histology images. *Med Image Anal*. 2019;56:122-139. doi:10.1016/j.media.2019.05.010
138. Marami B, Prastawa M, Chan M, Donovan M, Fernandez G, Zeineh J. Ensemble Network for Region Identification in Breast Histopathology Slides. In: Campilho A, Karray F, ter Haar Romeny B, eds. *Image Analysis and Recognition*. Lecture Notes in Computer Science. Springer International Publishing; 2018:861-868. doi:10.1007/978-3-319-93000-8_98
139. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. AAAI Press; 2017:4278-4284.
140. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(null):2825-2830.
141. Abadi M, Barham P, Chen J, et al. TensorFlow: A system for large-scale machine learning. Published online May 31, 2016. doi:10.48550/arXiv.1605.08695
142. Genomic Classification of Cutaneous Melanoma. *Cell*. 2015;161(7):1681-1696. doi:10.1016/j.cell.2015.05.044
143. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ; 2009:248-255. doi:10.1109/CVPR.2009.5206848
144. Cortes C, Mohri M, Rostamizadeh A. L2 regularization for learning kernels. In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09. AUAI Press; 2009:109-116.

145. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. Published online December 5, 2017. doi:10.48550/arXiv.1706.03762
146. Carbonneau MA, Cheplygina V, Granger E, Gagnon G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit*. 2018;77:329-353. doi:10.1016/j.patcog.2017.10.009
147. Pennie ML, Soon SL, Risser JB, Veledar E, Culler SD, Chen SC. Melanoma outcomes for Medicare patients: association of stage and survival with detection by a dermatologist vs a nondermatologist. *Arch Dermatol*. 2007;143(4):488-494. doi:10.1001/archderm.143.4.488
148. Ferreira R, Moon B, Humphries J, et al. The Virtual Microscope. *Proc Conf Am Med Inform Assoc AMIA Fall Symp*. Published online 1997:449-453.
149. Method and apparatus for creating a virtual microscope slide - Patent WO-9839728-A1 - PubChem. Accessed July 3, 2023. https://pubchem.ncbi.nlm.nih.gov/patent/WO-9839728-A1
150. Campbell WS, Hinrichs SH, Lele SM, et al. Whole slide imaging diagnostic concordance with light microscopy for breast needle biopsies. *Hum Pathol*. 2014;45(8):1713-1721. doi:10.1016/j.humpath.2014.04.007
151. Goacher E, Randell R, Williams B, Treanor D. The Diagnostic Concordance of Whole Slide Imaging and Light Microscopy: A Systematic Review. *Arch Pathol Lab Med*. 2017;141(1):151-161. doi:10.5858/arpa.2016-0025-RA
152. Williams BJ, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor D. Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. *Histopathology*. 2018;72(4):662-671. doi:10.1111/his.13403
153. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J Pathol Inform*. 2016;7(1):29. doi:10.4103/2153-3539.186902
154. Chang HY, Jung CK, Woo JI, et al. Artificial Intelligence in Pathology. *J Pathol Transl Med*. 2019;53(1):1-12. doi:10.4132/jptm.2018.12.16
155. Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. *Comput Med Imaging Graph Off J Comput Med Imaging Soc*. 2011;35(7-8):515-530. doi:10.1016/j.compmedimag.2011.02.006
156. Hart SN, Flotte W, Norgan AP, et al. Classification of Melanocytic Lesions in Selected and Whole-Slide Images via Convolutional Neural Networks. *J Pathol Inform*. 2019;10:5. doi:10.4103/jpi.jpi_32_18
157. Hekler A, Utikal JS, Enk AH, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. *Eur J Cancer Oxf Engl 1990*. 2019;115:79-83. doi:10.1016/j.ejca.2019.04.021
158. Elder DE, Bastian BC, Cree IA, Massi D, Scolyer RA. The 2018 World Health Organization Classification of Cutaneous, Mucosal, and Uveal Melanoma: Detailed Analysis of 9 Distinct Subtypes Defined by Their Evolutionary Pathway. *Arch Pathol Lab Med*. 2020;144(4):500-522. doi:10.5858/arpa.2019-0561-RA
159. Elder DE. Precursors to melanoma and their mimics: nevi of special sites. *Mod Pathol Off J U S Can Acad Pathol Inc*. 2006;19 Suppl 2:S4-20. doi:10.1038/modpathol.3800515
160. Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In: Niethammer M, Styner M, Aylward S, et al., eds. *Information Processing in Medical Imaging*. Lecture Notes in Computer Science. Springer International Publishing; 2017:146-157. doi:10.1007/978-3-319-59050-9_12
161. Zenati H, Foo CS, Lecouat B, Manek G, Chandrasekhar VR. Efficient GAN-Based Anomaly

Detection. Published online May 1, 2019. doi:10.48550/arXiv.1802.06222
162. Donahue J, Krähenbühl P, Darrell T. Adversarial Feature Learning. Published online April 3, 2017. doi:10.48550/arXiv.1605.09782
163. Akcay S, Atapour-Abarghouei A, Breckon TP. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. Published online November 13, 2018. doi:10.48550/arXiv.1805.06725
164. Di Mattia F, Galeone P, De Simoni M, Ghelfi E. A Survey on GANs for Anomaly Detection. Published online September 14, 2021. doi:10.48550/arXiv.1906.11632
165. Berg A, Ahlberg J, Felsberg M. Unsupervised Learning of Anomaly Detection from Contaminated Image Data using Simultaneous Encoder Training. Published online November 20, 2019. doi:10.48550/arXiv.1905.11034
166. Karras T, Aila T, Laine S, Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. Published online February 26, 2018. doi:10.48550/arXiv.1710.10196
167. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J*. 1991;37(2):233-243. doi:10.1002/aic.690370209
168. Huszar F, Theis L, Shi W, Cunningham A. Lossy Image Compression with Compressive Autoencoders. Published online May 1, 2020. doi:10.17863/CAM.51995
169. Lazarou C. Autoencoding Generative Adversarial Networks. Published online April 11, 2020. doi:10.48550/arXiv.2004.05472
170. Synced. The Staggering Cost of Training SOTA AI Models. SyncedReview. Published June 27, 2019. Accessed July 3, 2023. https://medium.com/syncedreview/the-staggering-cost-of-training-sota-ai-models-e329e80fa82
171. Pathak D, Krähenbühl P, Donahue J, Darrell T, Efros AA. Context Encoders: Feature Learning by Inpainting. In: IEEE Computer Society; 2016:2536-2544. doi:10.1109/CVPR.2016.278
172. Yang C, Lu X, Lin Z, Shechtman E, Wang O, Li H. High-Resolution Image Inpainting Using Multi-scale Neural Patch Synthesis. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ; 2017:4076-4084. doi:10.1109/CVPR.2017.434
173. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: Open source software for digital pathology image analysis. *Sci Rep*. 2017;7:16878. doi:10.1038/s41598-017-17204-5
174. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of wasserstein GANs. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc.; 2017:5769-5779.
175. The Shapely User Manual — Shapely 2.0.1 documentation. Accessed May 30, 2023. https://shapely.readthedocs.io/en/latest/manual.html
176. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial Autoencoders. Published online May 24, 2016. doi:10.48550/arXiv.1511.05644
177. Schlegl T, Seeböck P, Waldstein SM, Langs G, Schmidt-Erfurth U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med Image Anal*. 2019;54:30-44. doi:10.1016/j.media.2019.01.010
178. Colorectal cancer: Epidemiology, risk factors, and protective factors. Accessed May 30, 2023. https://www.medilib.ir/uptodate/show/2606
179. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi:10.3322/caac.21492
180. Dulskas A, Gaizauskas V, Kildusiene I, Samalavicius NE, Smailyte G. Improvement of Survival over Time for Colorectal Cancer Patients: A Population-Based Study. *J Clin Med*. 2020;9(12):4038. doi:10.3390/jcm9124038
181. Permission to Cite or Use NCCN Content. NCCN. Accessed May 31, 2023. https://www.nccn.org/guidelines/permission-to-cite-or-use-nccn-content

182. Uson PLS, Riegert-Johnson D, Boardman L, et al. Germline Cancer Susceptibility Gene Testing in Unselected Patients With Colorectal Adenocarcinoma: A Multicenter Prospective Study. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. 2022;20(3):e508-e528. doi:10.1016/j.cgh.2021.04.013
183. Moretz C, Byfield SD, Hatchell KE, et al. Comparison of Germline Genetic Testing Before and After a Medical Policy Covering Universal Testing Among Patients With Colorectal Cancer. *JAMA Netw Open*. 2022;5(10):e2238167. doi:10.1001/jamanetworkopen.2022.38167
184. Zhuang Y, Wang H, Jiang D, et al. Multi gene mutation signatures in colorectal cancer patients: predict for the diagnosis, pathological classification, staging and prognosis. *BMC Cancer*. 2021;21(1):380. doi:10.1186/s12885-021-08108-9
185. Ballester V, Cruz-Correa M. How and When to Consider Genetic Testing for Colon Cancer? *Gastroenterology*. 2018;155(4):955-959. doi:10.1053/j.gastro.2018.08.031
186. Smits LJH, Vink-Börger E, van Lijnschoten G, et al. Diagnostic variability in the histopathological assessment of advanced colorectal adenomas and early colorectal cancer in a screening population. *Histopathology*. 2022;80(5):790-798. doi:10.1111/his.14601
187. Sari CT, Gunduz-Demir C. Unsupervised Feature Extraction via Deep Learning for Histopathological Classification of Colon Tissue Images. *IEEE Trans Med Imaging*. 2019;38(5):1139-1149. doi:10.1109/TMI.2018.2879369
188. Boyd J, Liashuha M, Deutsch E, Paragios N, Christodoulidis S, Vakalopoulou M. Self-Supervised Representation Learning using Visual Field Expansion on Digital Pathology. Published online September 7, 2021. doi:10.48550/arXiv.2109.03299
189. Wang KS, Yu G, Xu C, et al. Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. *BMC Med*. 2021;19:76. doi:10.1186/s12916-021-01942-5
190. Deep learning-based histopathological segmentation for whole slide images of colorectal cancer in a compressed domain | Scientific Reports. Accessed May 30, 2023. https://www.nature.com/articles/s41598-021-01905-z
191. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks. Published online June 10, 2014. doi:10.48550/arXiv.1406.2661
192. Goldstein M, Uchida S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*. 2016;11(4):e0152173. doi:10.1371/journal.pone.0152173
193. Lee Y, Kang P. AnoViT: Unsupervised Anomaly Detection and Localization with Vision Transformer-based Encoder-Decoder. Published online March 21, 2022. doi:10.48550/arXiv.2203.10808
194. Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. Published online January 7, 2016. doi:10.48550/arXiv.1511.06434
195. Gillard R, Meroueh C, Gu Q, et al. Using Progressive Context Encoders for Anomaly Detection in Digital Pathology Images. Published online July 4, 2021:2021.07.02.450957. doi:10.1101/2021.07.02.450957
196. Jam J, Kendrick C, Drouard V, Walker K, Hsu GS, Yap MH. Symmetric Skip Connection Wasserstein GAN for High-Resolution Facial Image Inpainting. Published online September 12, 2020. doi:10.48550/arXiv.2001.03725
197. Otsu N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1):62-66. doi:10.1109/TSMC.1979.4310076
198. GeoPandas 0.13.0 — GeoPandas 0.13.0+0.gaa5abc3.dirty documentation. Accessed May 30, 2023. https://geopandas.org/en/stable/
199. Aperio GT 450 - Automated, High Capacity Digital Pathology Scanner. Accessed May 30, 2023. https://www.leicabiosystems.com/us/digital-pathology/scan/aperio-gt-450/

200. Wang S, Zhu Y, Yu L, et al. RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification. *Med Image Anal*. 2019;58:101549. doi:10.1016/j.media.2019.101549
201. Wang X, Chen H, Gan C, et al. Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis. *IEEE Trans Cybern*. 2020;50(9):3950-3962. doi:10.1109/TCYB.2019.2935141
202. Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. *Mach Learn Appl*. 2022;7:100198. doi:10.1016/j.mlwa.2021.100198
203. Ding M, Qu A, Zhong H, Liang H. A Transformer-based Network for Pathology Image Classification. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. ; 2021:2028-2034. doi:10.1109/BIBM52615.2021.9669476
204. Hu W, Li X, Li C, et al. A state-of-the-art survey of artificial neural networks for Whole-slide Image analysis: From popular Convolutional Neural Networks to potential visual transformers. *Comput Biol Med*. 2023;161:107034. doi:10.1016/j.compbiomed.2023.107034
205. Gutman DA, Cobb J, Somanna D, et al. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc JAMIA*. 2013;20(6):1091-1098. doi:10.1136/amiajnl-2012-001469
206. Litjens G, Bandi P, Ehteshami Bejnordi B, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*. 2018;7(6):giy065. doi:10.1093/gigascience/giy065
207. Veta M, Heng YJ, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Med Image Anal*. 2019;54:111-121. doi:10.1016/j.media.2019.02.012
208. Babaie M, Kalra S, Sriram A, et al. Classification and Retrieval of Digital Pathology Scans: A New Dataset. Published online May 21, 2017. doi:10.48550/arXiv.1705.07522
209. Richardson E, Weiss Y. On GANs and GMMs. Published online November 3, 2018. doi:10.48550/arXiv.1805.12462
210. Weights & Biases. W&B. Accessed July 6, 2023. https://wandb.ai/authors/DCGAN-ndb-test/reports/Measuring-Mode-Collapse-in-GANs-Using-Weights-Biases--VmlldzoxNzg5MDk
211. Common Problems | Machine Learning. Google for Developers. Accessed July 6, 2023. https://developers.google.com/machine-learning/gan/problems
212. Loss Functions | Machine Learning. Google for Developers. Accessed July 6, 2023. https://developers.google.com/machine-learning/gan/loss
213. Frogner C, Zhang C, Mobahi H, Araya-Polo M, Poggio T. Learning with a Wasserstein Loss. Published online December 29, 2015. doi:10.48550/arXiv.1506.05439
214. Wang D, Qin X, Song F, Cheng L. Stabilizing Training of Generative Adversarial Nets via Langevin Stein Variational Gradient Descent. *IEEE Trans Neural Netw Learn Syst*. 2022;33(7):2768-2780. doi:10.1109/TNNLS.2020.3045082
215. Bang D, Shim H. MGGAN: Solving Mode Collapse Using Manifold-Guided Training. In: IEEE Computer Society; 2021:2347-2356. doi:10.1109/ICCVW54120.2021.00266
216. Bhagyashree, Kushwaha V, Nandi GC. Study of Prevention of Mode Collapse in Generative Adversarial Network (GAN). In: *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*. ; 2020:1-6. doi:10.1109/CICT51604.2020.9312049
217. Metz L, Poole B, Pfau D, Sohl-Dickstein J. Unrolled Generative Adversarial Networks. Published online May 12, 2017. doi:10.48550/arXiv.1611.02163
218. PhD JRS. What are Stable Diffusion Models and Why are they a Step Forward for Image Generation? Medium. Published September 20, 2022. Accessed July 6, 2023. https://towardsdatascience.com/what-are-stable-diffusion-models-and-why-are-they-a-step-forward-for-image-generation-aa1182801d46
219. Dhariwal P, Nichol A. Diffusion Models Beat GANs on Image Synthesis. Published online June

1, 2021. doi:10.48550/arXiv.2105.05233
220. Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models. Published online December 16, 2020. doi:10.48550/arXiv.2006.11239
221. List of skin conditions. In: *Wikipedia*. ; 2023. Accessed July 6, 2023. https://en.wikipedia.org/w/index.php?title=List_of_skin_conditions&oldid=1157888468
222. Rook's Textbook of Dermatology, 4 Volume Set, 9th Edition | Wiley. Wiley.com. Accessed July 6, 2023. https://www.wiley.com/en-us/Rook%27s+Textbook+of+Dermatology%2C+4+Volume+Set%2C+9th+Edition-p-9781118441190
223. Branch NSC and O. Skin Diseases. National Institute of Arthritis and Musculoskeletal and Skin Diseases. Published April 20, 2017. Accessed July 6, 2023. https://www.niams.nih.gov/health-topics/skin-diseases
224. Skuhala T, Trkulja V, Rimac M, Dragobratović A, Desnica B. Analysis of Types of Skin Lesions and Diseases in Everyday Infectious Disease Practice-How Experienced Are We? *Life Basel Switz*. 2022;12(7):978. doi:10.3390/life12070978
225. Sirinukunwattana K, Alham NK, Verrill C, Rittscher J. Improving Whole Slide Segmentation Through Visual Context - A Systematic Study. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Lecture Notes in Computer Science. Springer International Publishing; 2018:192-200. doi:10.1007/978-3-030-00934-2_22
226. Guo Z, Liu H, Ni H, et al. A Fast and Refined Cancer Regions Segmentation Framework in Whole-slide Breast Pathological Images. *Sci Rep*. 2019;9(1):882. doi:10.1038/s41598-018-37492-9
227. Bándi P, van de Loo R, Intezar M, et al. Comparison of different methods for tissue segmentation in histopathological whole-slide images. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. ; 2017:591-595. doi:10.1109/ISBI.2017.7950590
228. Cui Y, Zhang G, Liu Z, Xiong Z, Hu J. A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images. *Med Biol Eng Comput*. 2019;57(9):2027-2043. doi:10.1007/s11517-019-02008-8
229. [PDF] Automated Localization of Breast Ductal Carcinoma in Situ in Whole Slide Images | Semantic Scholar. Accessed July 7, 2023. https://www.semanticscholar.org/paper/Automated-Localization-of-Breast-Ductal-Carcinoma-Seth/cf5ca179c98592e893199b6d3083b55e9f8ea269
230. Seth N, Akbar S, Nofech-Mozes S, Salama S, Martel AL. Automated Segmentation of DCIS in Whole Slide Images. In: Reyes-Aldasoro CC, Janowczyk A, Veta M, Bankhead P, Sirinukunwattana K, eds. *Digital Pathology*. Lecture Notes in Computer Science. Springer International Publishing; 2019:67-74. doi:10.1007/978-3-030-23937-4_8
231. Feng Y, Hafiane A, Laurent H. A deep learning based multiscale approach to segment cancer area in liver whole slide image. Published online July 25, 2020. doi:10.48550/arXiv.2007.12935
232. Xu Y, Jia Z, Wang LB, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics*. 2017;18(1):281. doi:10.1186/s12859-017-1685-x
233. Mehta S, Mercan E, Bartlett J, Weaver D, Elmore J, Shapiro L. Learning to Segment Breast Biopsy Whole Slide Images. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. ; 2018:663-672. doi:10.1109/WACV.2018.00078
234. Dong N, Kampffmeyer M, Liang X, Wang Z, Dai W, Xing E. Reinforced Auto-Zoom Net: Towards Accurate and Fast Breast Cancer Segmentation in Whole-Slide Images. In: Stoyanov D, Taylor Z, Carneiro G, et al., eds. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Lecture Notes in Computer Science. Springer International Publishing; 2018:317-325. doi:10.1007/978-3-030-00889-5_36
235. Lerousseau M, Classe M, Battistella E, et al. Weakly Supervised Pan-Cancer Segmentation Tool.

In: de Bruijne M, Cattin PC, Cotin S, et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Lecture Notes in Computer Science. Springer International Publishing; 2021:248-256. doi:10.1007/978-3-030-87237-3_24

236. Lerousseau M, Vakalopoulou M, Classe M, et al. Weakly Supervised Multiple Instance Learning Histopathological Tumor Segmentation. In: Martel AL, Abolmaesumi P, Stoyanov D, et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Lecture Notes in Computer Science. Springer International Publishing; 2020:470-479. doi:10.1007/978-3-030-59722-1_45
237. Encyclopedia of Bioinformatics and Computational Biology. ScienceDirect. Accessed July 7, 2023. http://www.sciencedirect.com:5070/referencework/9780128114322/encyclopedia-of-bioinformatics-and-computational-biology
238. Lowe DG. Distinctive Image Features from Scale-Invariant Keypoints. *Int J Comput Vis*. 2004;60(2):91-110. doi:10.1023/B:VISI.0000029664.99615.94
239. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-Up Robust Features (SURF). *Comput Vis Image Underst*. 2008;110(3):346-359. doi:10.1016/j.cviu.2007.09.014
240. Rublee E, Rabaud V, Konolige K, Bradski G. ORB: An efficient alternative to SIFT or SURF. In: *2011 International Conference on Computer Vision*. ; 2011:2564-2571. doi:10.1109/ICCV.2011.6126544
241. Alcantarilla PF, Bartoli A, Davison AJ. KAZE Features. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, eds. *Computer Vision – ECCV 2012*. Lecture Notes in Computer Science. Springer; 2012:214-227. doi:10.1007/978-3-642-33783-3_16
242. Li Y, Liu L, Wang L, Li D, Zhang M. Fast SIFT algorithm based on Sobel edge detector. In: *2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*. ; 2012:1820-1823. doi:10.1109/CECNet.2012.6201824
243. Hoque MdZ, Keskinarkaus A, Nyberg P, Mattila T, Seppänen T. Whole slide image registration via multi-stained feature matching. *Comput Biol Med*. 2022;144:105301. doi:10.1016/j.compbiomed.2022.105301