Enabling Microsegmentation: Digital Pathology Corpora for Advanced Model Development

Dmitry Hackel, Maria Bagritsevich, Claudia Dumitrescu, Md. Abdullah Al Mamun, Sadia Purba, Dylan Heathcote, Iyad Obeid and Joseph Picone

The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
{dmitry.hackel, maria.bagritsevich, claudia-anca.dumitrescu, abdullah18, sadia.afrin.purba, dylan.heathcote, iobeid, picone}@temple.edu

Abstract

Pathology remains a foundational discipline in modern medicine, yet traditional microscopy workflows are constrained by manual slide preparation, subjective interpretation, and limited scalability. The advent of digital whole slide imaging has enabled seamless digitization of histological sections, though challenges such as inter-observer variability and workforce shortages persist. Recently, artificial intelligence based decision-support tools – powered by deep-learning algorithms and high-performance computing – have demonstrated remarkable performance in pre-screening, case triage, and morphology segmentation.

Artificial intelligence in diagnostic histopathology hinges on the availability of extensive, richly annotated image repositories. Therefore, in this chapter we introduce two important open source digital pathology corpora: the Temple University Hospital Digital Pathology Breast Tissue Subset (TUBR) and the Fox Chase Cancer Center Digital Pathology Breast Tissue Subset (FCBR). TUBR comprises 3,505 images from 296 patients (1.2 TB of data). FCDP contributes 1,463 oncology-focused slides from 1,397 patients (336 GB of data). Both corpora have been manually annotated and cover a spectrum of tissue states from normal and benign through premalignant intraepithelial lesions and invasive carcinomas.

We also explore baseline performance of some well-established deep learning algorithms on an image classification task based on these corpora. A weighted error of 25.5% was achieved on TUBR using EfficientNet B7, while a weighted error of 25.0% was achieved on FCBR. We examine the challenges involved in training across both datasets to enhance overall system performance and assess the role accurate microsegmentations play in achieving clinically acceptable performance.

Keywords: histopathology, whole slide imaging, artificial intelligence, deep learning

1. Introduction

Pathology constitutes a cornerstone of modern medicine, providing essential support for accurate diagnosis, prognostication, and therapeutic planning across a wide range of tissue-based diseases. Traditionally, examination required a tedious workflow of preparing thin tissue sections on glass slides, followed by manual interpretation of stained tissue morphology under a microscope – an effort heavily reliant on the pathologist's expertise to discern subtle architectural, spatial, and cellular features. In support of enhancing this process, digital whole slide imaging, which emerged in the mid-1990's [1], enabled entire histological slides to be digitized into multi-resolution image pyramids. This allowed seamless, lossless zoom at arbitrary magnifications – unlike optical microscopes, which are limited to discrete objective lenses. This ensured consistent calibration, precise quantitative measurements, and preservation of cellular architecture across the specimen.

Corresponding Author: Joseph Picone, Room 718, College of Engineering, Temple University, 1947 North 12th Street, Philadelphia, Pennsylvania, 19122. Tel: 708-848-2846; Fax: 215-204-5280; Email: picone@temple.edu.

However, despite its potential, whole slide imaging did not mitigate inter-observer variability, a challenge that spans virtually all pathological subspecialties. A recent systematic review of 12 validation studies [2] reported that digital vs. light microscopy inter-observer κ values ranged from 0.45 ("moderate" agreement) to 0.75 ("substantial" agreement) across human pathology cases. More concerning was the fact that only 52% of pathologists report being satisfied with their career choice [3], and the U.S. pathology workforce declined by 17.5 % between 2007 and 2017 [4]. This decrease further worsens job satisfaction as remaining pathologists are being forced to increase workload, a difficult task to implement on a job requiring such precise analysis for each sample. To mitigate the negative impacts of the dwindling workforce, this chapter explores strategies for streamlining sample evaluation by equipping pathologists with advanced artificial intelligence (AI) based slide analysis tools.

Among the drivers of job dissatisfaction, it is worth mentioning that the declining headcounts, an aging workforce, and growing case complexity [5][6], as well as insufficient exposure to pathology during undergraduate and medical education, have been shown to deter up to 40 % of students from considering the specialty [7]. In response, AI-based decision-support tools have begun to fill critical gaps. In 2021, the FDA granted de novo authorization to Paige Prostate, the first AI-based in vitro diagnostic device for prostate cancer detection on digitized slides [8]. By leveraging advances in high-performance computing, gigapixel image storage, and deep-learning algorithms, such systems can pre-screen whole slide images (WSI), triage cases by risk, and highlight regions of interest – thereby reducing pathologist workload and enhancing diagnostic consistency beyond what whole slide imaging alone could achieve [9].

Conventional microscopy is a current standard within pathology. According to a recent study [10], 65% of pathologists remain reluctant to use the digital alternative, considering it unsuitable for routine diagnostic practice. Analog microscopy has a low to medium processing speed due to the need for manual transition of samples, which can lead to increased processing times that vary from a few minutes to hours [11]. In addition to eye strain, the need to be in a laboratory setting, especially when on call, and the increasing cost of manual labor, has fueled the push toward digitizing pathology. Digitized slides can be accessed and analyzed beyond the confines of the laboratory, offering significant advantages for pathologists on call, particularly in time-sensitive scenarios such as organ transplantation, where rapid remote assessment is critical.

The growing adoption of digital pathology within the field has provided a wealth of data that can be reviewed and utilized to further research into machine learning (ML) and deep learning (DL) technology. This technology has the potential to aid a pathologist's workflow. Such assistive resources enable automated slide triage (ordering cases by suspected severity), generation of annotated heatmaps highlighting regions of interest, and quantitative extraction of morphological features. Indeed, many studies report "near-perfect" performance: a recent meta-analysis of 48 AI algorithms across diverse disease types found a pooled sensitivity of 96.3 % and specificity of 93.3 % [12]. While the introduction of DL in digital pathology has demonstrated remarkable performance – achieving mean sensitivities of 95% and specificity of 92% for multiclass tasks – this often stems from evaluation of highly curated datasets with limited diversity, typically from one or two institutions, which raises concerns about bias, lack of generalizability, and reproducibility. As the same review [12] reports, 99% of included studies had at least one domain at high or unclear risk of bias, often due to non-random sampling, inadequate reporting, and domain shifts across institutions or scanners. This makes it difficult to translate AI models into robust clinical tools across different healthcare settings. Hence, in this chapter, we focus on establishing a performance baseline using our recently released datasets from two distinct medical centers, explicitly addressing these limitations through cross-institutional evaluation and transparent reporting practices.

2. Digital Pathology Resources for Machine Learning

To develop robust deep learning models for digital pathology, a well-curated dataset encompassing a variety of normal, benign, and cancerous tissue architectures is essential. At Temple University's Neural Engineering Data Consortium (NEDC), two major open-source digital pathology corpora have been established: the Temple University Hospital (TUH) Digital Pathology Corpus (TUDP) [13][14][15] and the Fox Chase Cancer Center (FCCC) Digital Pathology Corpus (FCDP) [16]. As the names suggest, TUDP was collected at Temple Hospital in Philadelphia – the major public hospital for the city. FCDP was collected at FCCC – a world-recognized leader in cancer research and treatment. TUDP contains over 100,000 high-resolution images, representing a diverse range of tissue types, including both normal and abnormal specimens, supporting a broad spectrum of research applications. In contrast, FCDP, with 14,276 images, is more specifically focused on oncological studies. These corpora have been further refined through extensive evaluation and annotation, particularly the breast pathology subsets, ensuring precise classification of cancerous, non-cancerous, and pre-cancerous structures. A comparison of these datasets to other publicly available datasets is given in Table 1. These enhanced datasets provide a critical foundation for training advanced AI models in digital pathology, facilitating improved diagnostic accuracy and research in cancer detection.

2.1. Slide Preparation

Traditional histological slide preparation [17] requires tissue sections to be thin enough to be mounted while maintaining the structural relationship between cells and extracellular components. This necessitates careful handling to ensure accurate cancer diagnosis. To support tissue integrity during sectioning, specimens are either frozen and cut using a cryostat microtome for rapid diagnosis without chemical interference or infiltrated with a liquid agent such as epoxy resin or histological wax, which subsequently solidifies to encapsulate the sample. Before mounting, tissue specimens must be preserved using a fixative, commonly a formaldehyde suspension for 6-12 hours, to prevent decomposition and eliminate microorganisms. Specimens are then dissected into 4 mm regions and processed with a hardening agent. Once prepared, paraffin sections of 3-5 μm thickness are cut to ensure only a single layer of cells is captured on a glass slide. These sections are then placed in a warm water bath for flattening and ease of handling before being mounted onto individual glass slides.

2.2. Immunostaining

Staining is essential for visibility during imaging [18]. Cells, apart from a few natural pigments like melanin, are mostly colorless. The routine hematoxylin and eosin (H&E) stain is used to provide general cellular structure information, staining nuclei blue and other cellular components in shades of pink. If further diagnostic information is required, special stains targeting specific cellular components are applied. For example, immunohistochemical (IHC) stains utilize antibodies to mark diagnostically relevant proteins, offering deeper insights into cellular processes. Once stained, the slides are covered with glass. Most slides can be achieved for years without damage or discoloration.

Immunostaining plays a critical role in the annotation of digital pathology datasets for machine learning applications, as it is capable of shifting data distribution very easily and making the model miserably fail the generalization test. This is because different staining protocols highlight different cellular and molecular features while failing to reveal others, making cross-dataset transferability challenging. As a result, switching between datasets becomes more complicated due to variation in immunochemistry. In digital pathology, immunohistochemical (IHC) stains utilize antibodies to detect specific cellular markers, such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), which are essential in breast cancer classification [19]. These stains provide a detailed molecular profile of

tissues, aiding in the differentiation of neoplastic from non-neoplastic regions, which is crucial for training deep learning models to recognize cancerous patterns with high specificity.

In contrast, Hematoxylin and Eosin (H&E) staining remains the gold standard for histological examination, offering a broad visualization of cellular and tissue structures. Hematoxylin stains cell nuclei blue-purple, while eosin highlights cytoplasm and extracellular matrix in pink, allowing machine learning models to extract key morphological features that distinguish normal, pre-cancerous, and malignant cells [20]. The integration of both staining techniques enhances model performance by expanding the diversity of input data, enabling algorithms to learn from both molecular and structural tissue characteristics. Moreover, computational pathology tools leveraging deep learning can process large-scale immunostained datasets efficiently, reducing interobserver variability and improving diagnostic consistency [20]. By incorporating a variety of immunostains, machine learning models gain a deeper contextual understanding of tissue pathology, facilitating advancements in automated cancer detection, classification, and prognostic assessment.

2.3. Digitization

Digital pathology involves converting physical glass slide samples into digital images stored using advanced compression techniques. This is achieved using a laser scanner that captures images at a resolution of 0.2 μm/pixel. Similar to a glass slide, a WSI contains one or more specimens along with the background space surrounding the histological sample. All scanning is performed using a Leica Biosystems Aperio AT2 high-volume scanner. The scanning process is described in detail in [13]. The scanner captures a low-resolution image and assigns multiple focus points on the tissue. A green rectangular box marks the scanning frame, and technicians can manually adjust the focus points and frame as needed. This manual review step is crucial in preventing scanning failures caused by inaccurate automatic focus placement.

There are cases where the scanner makes errors in its placement of the focus points and in determining the area of the scanning region. Both cases will cause a failure in the image processing of the specimen, thus causing a failure in scanning. Another error modality occurs when the scanning region is too large, which can cause unnecessary white space to be included in the WSI. In this case, the scanner might produce an image that is larger than necessary. Similarly, there is a risk of producing invalid images if the scanning region only partially encloses the specimen of interest. These events tend to occur with slides that are lightly stained or slides with a significant amount of white space between tissue samples. In such cases, manual adjustment of focus points becomes required

Errors caused by image blur frequently results in resolution issues within some pathology slides. Image blur is a result of the attenuation of high spatial frequencies which occur when the image is compressed or filtered. Blur can also be a result of image acquisition [21]. This makes the pre-scan snapshot phase of the scanning procedure labor intensive. We find about 2% of the slides scanned are likely to experience a scanning failure. However, this number varies according to the quality of the stain applied to the slides. These failed slides were reviewed, readjusted, and rescanned. To date, we have scanned approximately 120,000 slides using these tools.

The scanned image produced from the Aperio AT2 is stored using a ScanScope Virtual Slide (SVS) format [22]. An SVS file is layered image representation that includes several thumbnails and the original source image. The source image is stored using JPEG compression with a quality factor of 70. (The specific image type in the Aperio ImageScope software is “SVS/JPEG 2”. The parameter “Image Depth” is set to 1 and “Image Channels” is set to 3.) These parameters result in roughly an order of magnitude of compression over lossless compression with minimal image degradation.

A full resolution image, or WSI, is stored as the baseline image using a tile size of 240 x 240 pixels (an image is represented as a series of adjacent tiles, or blocks). The following three layers contain downsampled version of the image at resolutions of 4:1, 16:1 and 32:1. The final layer is a low-resolution thumbnail. Each of these layers is an image encoded using lossy JPEG encoding. An SVS file also contains a low-resolution picture of the slide's label as metadata and stores other information such as the downsample and offset information. The number of layers generated depends on the size of the original image. Smaller images (e.g., the maximum dimension is less than 50K pixels) will generally have only two layers.

Our primary software for viewing SVS files is Aperio ImageScope [23]. This open source software features a wide variety of tools for image editing, adjusting, and annotation. The image adjustment tools include brightness and contrast controls, color balance, and color curve adjustment. These can be applied to all channels or for individual red, green, or blue (RGB) channels. These adjustments only apply to the viewed image and do not modify or overwrite the stored image. The settings applied to the current session can be saved and applied to other scanned images. The default presets for ImageScope are always applied to the scanned images. Then, the image adjustment tools can be utilized to calibrate the image features according to the pathologist's preferences. The stains applied on specimen images cause a specific structure to adopt a distinct color, and this color can be enhanced by adjusting the brightness, contrast, and the color channels of the image via the image adjustment tools. This is a beneficial tool for pathology diagnosis as it can be used to allow specific areas of investigation (such as cancerous tissue) to be more focused or to enhance quality of lightly stained specimens.

As mentioned earlier, the Aperio AT2 scanner features a z-stacking option. The scanner can produce multiple images of a slide tissue that were scanned at different focal depths. This generates a 3D image that allows navigation of the image through different focal depths, which is analogous to the process pathologists use with an analog microscope. ImageScope features a tool that can adjust the focal depth that is similar to using the objective fine and coarse adjustments of a focus slider in a microscope [23]. This feature has yet to be explored in our work but is being used by other hospitals. The z-stacked images are very large in size, often several gigabytes, which poses additional challenges for machine learning research [13].

2.4. Data Organization and Anonymization

A detailed description of the filenaming conventions and organization of the data can be found in [15]. The file naming convention is designed so that every file in the corpus has a unique filename, and simple UNIX commands can be used to locate data.

To ensure compliance with the Health Insurance Portability and Accountability Act (HIPAA), strict protocols are in place to maintain patient anonymity. Protocol No. 24943, approved by TU's Institutional Review Board, governs the handling of all data to prevent any exposure of patient-identifiable information. The deidentification process follows a methodology similar to the one used in the Temple University Hospital EEG Seizure Corpus [24]. Each patient is assigned a unique, randomized 8-character letter sequence.

SVS files initially contain low-resolution tiled images of slide labels, which may include patient initials and specimen IDs. These labels are manually removed before data release. Together, these measures ensure that all released data is completely anonymized while maintaining the integrity required for research.

2.5. Annotation Guidelines

Pathology is the primary method for the confirmation of breast cancer, assessing tumor margins, and evaluating cancer proliferation. Assessing these metrics can be challenging due to the diverse nature of pathology. The variety of different tissue architectures a pathologist needs to assess is vast, which makes our dataset extremely valuable among the existing alternatives, as shown in Table 1. Annotation refers to the manual delineation of both biological and non-biological structures by an annotator. We refer to these as patch-level annotations since they include a polygon defining the region and a label. Our annotation methodology is discussed in detail in [16][25]. Here we summarize some key issues that make our annotations unique and extremely valuable for machine learning.

Annotation labels need to ensure precise margins to minimize extraneous structures such as background fibrous or adipose tissue. Proper margins or delineation of structures of interest are important for models that will be trained to do fine-grained segmentation image analysis, which we refer to as microsegmentation. The ability to delineate differences on a cellular level is incredibly important in pathology. Accurate delineation ensures that the model learns to distinguish the true boundaries of pathological structures, which is critical for reliable feature extraction, tissue classification and overall model performance.

Annotation of an entire slide is extremely time-consuming and is prohibitively expensive. Therefore, images are *partially annotated*, focusing on a few instances of what we consider prominent or significant structures. These structures are either selected based on patient reports, which identify regions of diagnostic interest, or on an annotator's decision of what would be most valuable for ML training. Selective annotation ensures efficient use of time and resources while still capturing clinically relevant tissue architecture.

Areas of interest in pathological slides can vary significantly. Although it would be simple to label tissue as either normal or cancerous, the vast array of different structures poses a real issue for models. In addition, there are many non-cancerous or pre-cancerous structures that would inform a pathologist about making medical determinations. We use a system that includes five labels for relevant pathologies, and four labels for non-cancerous or anomalous tissues. The full list of labels can be found in [15] and is repeated here in Table 2. We briefly describe these labels below.

Normal ductal terminal units and lobules are labeled with the symbol (NORM), which is crucial for training models to allow for the differentiation between normal ducts and pathological ones. One of the many things that makes this type of problem challenging is that a small percentage (less than 1%) of the overall slide area contains a cell or group of cells that represent areas of interest. Further, there can be more than one tissue on each slide, and each of these tissue samples can have multiple labels. Hence, the ability to segment images and localize where events of interest occur in the data is critical. Because of the huge imbalance in the frequency of occurrence of the labels in Table 2, care must be taken during training to avoid models from focusing on normal tissue or white backgrounds.

2.5.1. An Overview of Anatomical and Physiological Considerations

When we discuss how we annotate the data, it is important to understand the basis of the anatomical and physiological components of breast tissue and how cancerous cells develop [26][27]. The mammary gland parenchyma consists primarily of adipose stroma within a connective tissue matrix enriched with interstitial fluid. It contains lobular alveoli responsible for milk synthesis, which is transported through lactiferous ducts that converge and open at the nipple. The terminal ductal lobular units (TDLUs) are a primary interest for pathologists as most primary breast cancers originate in these regions. Lobules, derived from terminal duct lobular units (TDLUs), are hormonally responsive structures composed of glandular epithelial cells encircled by myoepithelial cells, a basement membrane, and fibrocollagenous stroma. These lobules are organized in clusters and play a crucial role in milk secretion.

The epithelial–stromal interfaces within the lobules are composed of an epithelial–myoepithelial layer supported by elastic fibers, fibroblasts, and capillaries, with fiber deposition varying according to hormonal status and age, thereby adding complexity to the interpretation process. The hormonal impacts on the ductal units due to menopause, adolescence, or pregnancy means a far greater variability in lobular appearance. Standard acini and ducts are made up of three layers: the basement membrane, myoepithelial layer, and epithelial lining. These ducts carry milk and are found throughout the TDLUs. The epithelial layer is usually only one thick cell which lines the ducts. However, multiple layers of cells, hyperplasia, is a common occurrence. The degree or rate of growth of these epithelial cells determines if a patient has atypical ductal hyperplasia or ductal carcinoma in situ.

A significant contributor to the complexity of breast tissue analysis, particularly in pathological interpretation, is a dynamic and differing morphological variability of terminal ductal units driven by hormonal influences. Unlike other tissue architecture which may remain uniform, the breast undergoes structural changes depending on factors such as adolescence, menstrual cycling, pregnancy, lactation, and menopause. These states change the proliferative activity and differentiation status of epithelial and stromal components within the terminal duct units. For example, during pregnancy and lactation, lobules become hypertrophic, and exhibit increased acinar formation with prominent secretory activity. This makes the ducts look atypical and could be mistaken as non-neoplastic. Women who are going through menopause have reduced lobular units, atrophic epithelial lining, and increased stromal fibrosis or adipose replacement. These changes could pose a challenge for the development of models that can accurately analyze images for women of all ages and hormone levels.

Morphological variability in pathology is also shaped by genetic and ancestral factors, contributing to disparities in breast cancer characteristics observed within different racial populations. It has been well established that the mutations in BRCA1/BRCA2 genes significantly elevate the risk of developing ductal and invasive carcinoma. Individuals of African-American descent tend to develop higher-grade tumors, more frequently present with basal-like or triple-negative subtypes, and are often diagnosed at a younger age. Race can also play a role in the tissue architecture of the terminal ductal units, meaning that the structure and image analysis of breast tissue specimens can have the same diagnosis, but may look different depending on a multitude of factors. This highlights the importance of data curation and diversity of data when developing such models. It is therefore an ethical concern to design models that uphold fairness, transparency, robustness, and ensure a high degree of explainability and privacy.

2.5.2. Cancerous Structures

The two primary labels utilized for cancerous structures are ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (INDC). Although there is some differentiation between what constitutes ductal carcinoma among pathologists, cross-referencing with patient reports and metadata files ensures consistency and accuracy in labeling. DCIS is defined by abnormal growth of epithelial cells throughout the duct. DCIS presents itself in a multitude of ways. Cribriform formations look like sponges as microcysts from within the tumor inside the duct. Micropapillary DCIS is defined by its microcalcification on mammograms and range in size. Other common formations of DCIS include apocrine, comedo, with comedonecrosis, or papillary DCIS. All forms are characterized by abnormal cells confined within the ductal units of the breast, as cancer tends to proliferate along the path of least resistance. Ductal carcinoma in situ (DCIS) is typically graded on a scale of 1–3, and in some cases 1–5, based on the mitotic activity (the rate at which cells are dividing and creating new cells).

Once DCIS breaches the basement membrane of the ductal epithelium and invades the surrounding stromal tissue, it is classified as INDC. The INDC label makes up 70% to 80% of breast cancer cases [28]. The various presentations of invasive ductal carcinoma patterns include nests, cords, or sheets of small squamous cells. The tumor cells often exhibit moderate to marked nuclear pleomorphism and increased

mitotic activity [29]. Tubule formation is a key histological feature assessed during grading. Tumors with more than 75% tubule formation receive a lower grade, indicating better differentiation and prognosis. However, many INDC occurrences display less tubule formation, correlating with higher grades and more aggressive behavior. The heterogeneity in the histological presentation of INDC poses challenges for AI models in accurately classifying and predicting outcomes, especially when rare subtypes or poorly differentiated forms are underrepresented in training datasets.

Non-neoplastic (NNEO) structures is a broad term that, in a clinical context, is nonspecific and encompasses a wide range of structural types. In TUBR and FCBR, NNEO structures refer to a range of structures that may either present themselves as a benign formation, pre-cancerous, or associated with cancer. The term non-neoplastic can also encompass lesions with precancerous potential, such as atypical ductal hyperplasia, as well as benign structures like cysts that may mimic pathological features.

A NNEO structure refers to the formation of calcium deposits, often detectable on mammograms. These deposits can result from abnormal cellular activity, including cancer proliferation, which increases calcium output that may accumulate and form solid structures [30]. These calcium microcalcifications are likely a byproduct of dysregulated intracellular calcium transport channels due to oncological transformations. The regulatory imbalance, induced by oncological cells, causes an overexpression of calcium pumps such as SPCA2 and PMCA2 [30], creating an environment that allows calcium to react with bone matrix proteins and phosphate transporters to develop a microenvironment conducive to mineralization. This results in microcalcification [31]. The loss of calcium homeostasis and resulting structures could be leveraged to improve model training. We are exploring this in ongoing research and have released a corpus to support this research [32].

NNEO labels include benign proliferative or reactive changes that could mimic cancers, improving model specificity. Inflammation labels (INFL) are given to areas of lymphocyte reactions to either non-neoplastic or cancerous structures. The distinction between these two labels allows for the distinction between an immune response and possible malignant progression.

2.5.3. Non-cancerous and Ambiguous Structures

Artifact (ARTF) refers to the identification of grease-labels or other non-biological artifact found within slides and serves to prevent misclassification. These artifacts, if not properly labeled, could be misinterpreted by the model as meaningful biological signals, potentially leading to false positives.

Null (NULL) labels are given to areas that are considered undistinguishable tissue, either due to a blurry image or poor stain techniques. These regions are intentionally excluded from training to prevent introducing uncertainty or noise to the model.

The suspicious (SUSP) label is given to tissues that are suspected to be precancerous or are incredibly ambiguous. Background labels (BCKG) serve to prevent models from learning patterns in irrelevant regions. Background labels are given to any structure that does not follow in the prior labeling selection and primarily is made up of adipose and fibrous tissue. Since a large percentage of the area of a slide is non-tissue (white), we provide a BCKG label so ML systems can be trained on tissue patterns that are not one of our five significant labels. Without this label, an ML system will tend towards modeling a white image as background and will confuse images of irrelevant tissue as one of the meaningful five classes.

Maintaining balanced label distributions is essential to mitigate bias and improve model robustness. Although randomization and weighted labels can be utilized to mitigate label biases, the importance of ensuring a good label distribution is vital. Class imbalance poses a significant challenge in medical datasets, especially when tasked with identifying uncommon tissue architectures. In digital pathology databases, the

abundance of normal and background tissue is significantly more common than carcinogenic or non-neoplastic structures. If the annotation process is not meticulous, a model would be disproportionately accurate on normal tissue architecture while severely underperform on cancerous structures. This overfitting of dominant labels can lead to models effectively ignoring minority classes. To address this issue, curating these datasets with sufficient representation of all relevant classes has been an important annotation objective, in addition to adding weighted loss functions and randomization during training.

2.6. Typical Examples of Annotated Slides

In Figure 1, we show normal duct and normal lobules, components of a TDLU, surrounded by stromal background tissue. The stroma is the pink-stained tissue found throughout the slides, which functions as supportive tissue around the ducts and lobules. Figure 1 illustrates fibrous stroma surrounding the regions of interest, with the inclusion of a few fat cells which appear as white round shapes. The NORM label is applied to regions where lobules remain circular with no instances of hyperplasia and where the ducts are clearly defined with both layers of the lining intact. The lobules around the duct are composed of small glands that are called acini, which are formed during puberty and represent secretory units of the breast. The acini and the ducts are lined by two different cell layers. The outer layer is made up of myoepithelial cells that are mostly just clear cytoplasm and function to contract and push milk through the breasts. The inner layer is made up of epithelial cells that produce breast milk and are usually where abnormal cell growth develops. The stroma consists of connective tissue, blood vessels and fat cells. The difference between fatty stroma and fibrous stroma can be seen in Figure 2.

NNEO annotations encompass a range of structures that aid in contextualizing the interpretation of specific tissue architecture. The NNEO label typically includes entities such as atypical hyperplasia, intraductal papilloma, calcifications, benign structures, and adenosis. Figure 3 highlights one of the most common NNEO structures: cysts. Cysts present as pale round spaces lined by a single layer of cuboidal epithelium and typically contain proteinaceous fluid. Cysts and cancer often coexist, as tumor growth can obstruct normal ductal drainage pathways and is frequently influenced by hormonal activity. The presence of cysts and other non-neoplastic features provides additional diagnostic clarity, supporting annotation of malignant or benign conditions. It is important to note that NNEO structures are not cancer, but rather benign or pre-malignant findings that assist in accurate classification.

Figure 4 represents an atypical case of DCIS, annotated due to its unique spherical nature resulting from its confinement within the breast ducts. Although DCIS may present itself with multiple structures within the same area, they are often annotated independently from one another. DCIS is identified by malignant epithelial cells that have not breached the basement membrane. Immunochemistry often stains these cells differently compared to the normal tissue. Calcifications are frequently present and may be visible present within the ducts. Ductal carcinoma regions were labeled based on nuclear atypia, preservation of the basement membrane and architectural patterns (e.g., cribriform and calcification).

In Figure 5, we present a terminal duct lobular unit (TDLU) undergoing non-neoplastic changes. The annotation of this structure may include either the entire TDLU, encompassing multiple abnormal ducts simultaneously, or a single duct. This annotation approach is consistent across many labels, contributing to a more diverse and representative dataset for model training.

To aid in this process, it is important to understand the characteristic non-neoplastic changes such as gynecomastia. Also known as gynecomastoid hyperplasia or gynecomastia, this condition is characterized by two cell layer linings instead of one within the epithelial cell wall of the lobule or duct. Those formations are harder to detect since they require high magnification down to the cell layer of the lobule. The cell can appear mushroom-like if the bi-layer juts into the lumen where it can become narrower toward the center of the lumen. Hyperplastic changes arise when the epithelial lining proliferates, encroaching on the lumen

and leaving minimal or no empty space within. An additional feature of hyperplasia is represented by mushroom-like, irregular expansions of the cell. It can be classified as elongated, ovoid-shaped nuclei (dot-like particles) within the cell wall lining. Lymphocytes appearing as small dark dots can be seen, pointing to inflammation of the lobule. These visual markers guided annotators in consistently labeling regions exhibiting hyperplasia and inflammation.

In Figure 6, we show an example of intraductal papilloma, another nonneoplastic growth. Papillary lesions are known to be a heterogenous group histologically and clinically, which represents a significant challenge in making an accurate diagnosis [33]. These lesions are annotated with the NNEO label usually encompassing the entire papilloma or, at a minimum, ensuring that key architectural features are encompassed. This group includes both benign and cancerous formations, as well as lobular neoplasia. Intraductal papillomas are made up of well-kept, vein-like structures that are contained, with little inflammation surrounding the region. Their inner epithelial walls and surrounding myoepithelial tissue are highly proliferated and unbroken. This region may include atypical benign tumors within. Due to its massive and weblike appearance, this region can be mistaken for DCIS. It is important for annotators to pay attention to the classic signs of DCIS when annotating similar regions, such as an unbroken border and or nuclear abnormalities. Lack of those in a visually similar structure is characteristic of intraductal papilloma and other benign papillary lesions.

2.7. Annotation Challenges

Producing high-quality digital pathology data for breast cancer is challenging, particularly in distinguishing non-neoplastic structures from early-stage carcinogenic lesions and ensuring consistency in pathology diagnoses. One of the most complex distinctions is between atypical ductal hyperplasia (ADH) and ductal carcinoma in situ (DCIS), as these lesions share overlapping morphological features, leading to high interobserver variability even among expert pathologists [34]. Studies have shown that agreement rates in diagnosing ADH versus DCIS can be as low as 48%, highlighting the subjective nature of classification [35]. This inconsistency extends to broader breast cancer research, where pathologists often disagree on borderline or pre-malignant conditions, complicating the creation of reliable training datasets for machine learning. Deep learning models require consistent ground truth labels, and if expert pathologists struggle with diagnostic uniformity, AI systems risk learning from misclassified images, reducing their accuracy and real-world applicability [36].

Rare biological structures are underrepresented and pose a significant challenge for models. The lack of sufficient examples of uncommon cases makes it difficult to train models. Both metaplastic breast carcinoma and tubular carcinoma are invasive ductal carcinomas that are representative of less than 2% of all breast cancers. For example, tubular carcinoma can be mistaken by models as normal ductal formation because they are defined by small, angulated tubules with minimal atypia. Metaplastic breast carcinoma is poorly differentiated. It exhibits highly variable morphology, often including spindle cells, squamous components or even bone or cartilage like tissue. The lack of frequency in these presentations of INDC can lead to significant misclassification of models. This variability in morphological presentation of carcinoma further complicates the task of accurately classifying these structures.

Another issue with large datasets like this is the possibility of annotator bias. Due to the partially annotated nature of the data, annotation bias could pose a real issue and be translated into model. This was mitigated by multiple cross-reviews. Furthermore, not all breast cancer diagnoses can be determined solely through digital pathology. While histological analysis of stained tissue samples is fundamental for diagnosing breast cancer, certain molecular and genetic subtypes require additional testing beyond digital imaging. For example, triple-negative breast cancer (TNBC), an aggressive subtype that lacks ER, PR, and HER2 expression, cannot be diagnosed based purely on histopathology [37]. Immunohistochemistry (IHC) and molecular profiling techniques such as fluorescence in situ hybridization (FISH) or next-generation

sequencing (NGS) are necessary to determine receptor status and guide treatment decisions [38]. This limitation underscores the need for multimodal diagnostic approaches, where digital pathology is complemented by molecular testing to provide a comprehensive assessment of breast cancer cases.

3. Database Analysis

Both breast tissue subsets discussed in this chapter are subsets of much larger corpora. A brief comparison of the full corpora from which these were extracted is given in Table 3. Because these two corpora were collected at different hospitals, the data is coded differently, which makes direct comparisons difficult. Hence, one important goal of this work was to normalize the data so they could be used simultaneously for technology development. Each database contains an ample amount of breast tissue data. FCDP is a bit richer in terms of occurrences of cancer because that is a specialty for FCCC. TUDP better reflects general pathology studies at a major urban public hospital.

Annotations are created by undergraduate bioengineering or pre-health students who are trained extensively by senior personnel and board-certified pathologists. These students go through extensive training and supervision by senior annotators before they are allowed to annotate production data. In the early stages of the project, they were trained directly by TUH pathologists. However, once we developed adequate experience in house, a pipeline was developed in which senior annotators would train and supervise junior annotators. A significant amount of data has been reviewed by at least three annotators. Informal annotator agreement experiments have shown over the years that annotators have performed well and compare favorably to experts with a Kappa statistic above 0.8. Clinically trained pathologists have reviewed some of the data as part of their supervision process and provided ample feedback on our annotation process.

In addition, to aid in the annotation process, for TUDP, anonymized patient reports are available for reference. The TUDP reports, which are fairly detailed, were manually anonymized and held on a secure server. Though there is some structure to these reports, they are difficult to parse automatically for specific information due to the complex nature of the language. Hence, we mainly used them to validate challenging cases. Data selection to create various subsets of the database is usually done based on manual review of the reports.

The FCDP data was completely anonymized before we were given access. FCDP includes metadata for each slide that has been aggregated in a spreadsheet [15]. This spreadsheet contains a wealth of information about the tissue samples and diagnoses. The metadata usually do not identify non-neoplastic or abnormal structure but do indicate the existence of ductal carcinoma or invasive ductal carcinoma.

A comparison of the statistics for the annotations for the two breast tissue subsets is given in Table 4. In both cases, we selected all the available data from the master corpora for annotation. The breast tissue subset within these corpora contains a larger number of WSI than most publicly available breast tissue datasets (see Table 1). These subsets contain 4,968 WSI annotated with the nine different labels shown in Table 2. The FCDP and TUDP datasets include more annotated instances than all other breast-focused datasets analyzed, with the exception of the SPIDER subset from HistAI initiative. While many of the alternative datasets rely on binary classification (e.g., cancer vs non-cancer), FCDP and TUDP adopt a more granular annotation framework. While binary classification may suffice for certain machine learning tasks, we argue that a more detailed annotation structure is essential to capture the complexities of tissue architecture and nuance morphological heterogeneity in breast tissue, enabling more robust and biologically informed model development.

3.1. The TUDP Breast Tissue Subset (TUBR)

Temple University Hospital Breast Tissue Subset (TUBR) contains 3,505 high resolution images with various degrees of carcinogenicity, ranging from samples with healthy background, ducts, and lobules to samples with definitive ductal and invasive carcinoma. This subset contains 8,035 non-cancerous, 6,222 carcinogenic, and 2,714 cancerous identified structures. This totals 22,241 labels and an average of 6.3 annotations per slide (Table 4). The most common structures identified, aside from background, are non-neoplastic, normal ductal structures and invasive ductal carcinoma. Due to the general nature of pathologies Temple Hospital investigates, a wide range of non-neoplastic changes and normal structures can be found in this subset. The dataset provides a unique archive of various benign non-cancerous structures such as papilloma, fibroadenoma, ductal and lobular hyperplasia and inflammatory lesions. The dataset is also rich in various samples of normal healthy ducts and lobules, with various degrees of branching and imaging clarity, which can be beneficial as a control for mistakenly marking healthy structures as non-neoplastic or carcinogenic. Each individual slide has at least one background label to equalize the difference in staining saturation and scanning discoloration.

We can see that a fairly small percentage of the overall slide area has been annotated (1.96%). Fully annotating each slide is simply not feasible or cost-effective, so our annotation team focused on events that would be most meaningful for ML technology development. In Table 5, we show the average dimensions of an annotated region sorted by label. DCIS and NNEO are the largest, and INFL are the smallest. Inflammation often accompanies the carcinogenic structure because it is an immune response from the body. Therefore, many times we see it as an "additional layer" surrounding abnormality as the body tries to respond to it. In addition, lymphocytes, the cell which we label as INFL, are composed of very small cells. The "structure" of inflammation is not really a structure at all but rather a clump of lymphocytes. Their small nature and their proximity to the structure of interest results in smaller-sized regions.

3.2. FCCC Breast Tissue Subset

Fox Chase Cancer Center Breast Tissue Subset (FCBR) contains 12,164 non-cancerous, 1,967 carcinogenic, and 5,954 cancerous identified structures. This totals 20,086 labels and an average of 13.7 annotations per slide, as shown in in Table 4. Each individual slide has at least one background label to allow ML to equalize the difference in staining saturation and scanning discoloration. When compared to TUBR, FCBR is more heavily weighted towards malignant pathology, offering a greater proportion of invasive ductal carcinoma (INDC). This makes it valuable for training models and validating models focused on identification of cancerous structures.

The FCBR Corpus includes extensive metadata about each slide [15]. In Table 6, we show a histogram of the ICDO-10 codes [39] for both the tissue site and the tumor site. The breast tissue subset was the subset of FCDP that c50.* codes for both. A few entries did not contain tumor site codes and were assigned the value "cxx.x" as a placeholder. The most common combination of values for these were c50.9 ("Malignant neoplasm of breast of unspecified site") and c50.4 ("Malignant neoplasm of upper-outer quadrant of breast"). Over 90% of the samples are contained in the top 10 code pairs.

FCBR metadata also contains a wealth of metadata about the diagnosis. Diagnosis from a single pathology slide is, of course, a challenging problem (and the reason we are investing research in these resources). In Table 7, we show a histogram of several important metadata fields that can be used to interpret the slides and guide ML. The first column contains a clustering of the slides into four values [15]: high grade (hg), intermediate grade (ig), low grade (lg), and unknown (UNK). These judgements were made using all the available metadata and manual annotations. The next three columns represent selected metadata fields that were used to support these judgements. Note that only the top 10 most frequent combinations are shown. The top 10 for "hg" covers 54% of the corpus while the top 10 for "ig" covers 29% of the corpus. Together

these two categories cover over 80% of the corpus. FCBR is much richer with respect to indications of cancer than TUBR.

These datasets are one-of-a kind, meticulously curated, and multi-reviewed data corpora, distinguished by their accompanying unique annotations. Each annotation has undergone validation based on patient reports, and has been cross-reviewed. This thorough review enhances their reliability for training deep models in digital pathology, particularly for breast tissue analysis. The combination of extensive image diversity and precise makes the corpora invaluable resources for advancing research in cancer detection and image analysis. Given that the dataset originates from two entirely different sources but share the same format, they offer a practical framework for evaluating a model's capacity for generalization across separate data domains.

4. Baseline Experiments

There has been rapid growth in interest in automated methods for interpretation of digital pathology images in recent years. Though many current models have limited capabilities in terms of the type of diagnosis they can provide (e.g., only making whole slide classifications), the range of analysis techniques available has grown significantly over the years. A generation of models first debuted with only two prediction classes (cancerous and non-cancerous). However, models have since evolved to detect a broader range of pathologies, including multiple cancer types, precancerous lesions, and various non-cancerous conditions (Table 1). Advances in AI-driven digital pathology enable these models to identify complex biomarkers and subtle histopathological features, facilitating more precise and personalized treatment options. Furthermore, the AI systems can integrate molecular and genomic data, providing a deeper understanding of disease mechanisms, prognosis, and potential therapeutic responses, thus transforming the landscape of diagnostic pathology.

Accurate identification of the labeled regions described in Table 2 requires analysis of high resolution images at full resolution. This is a problem we refer to as microsegmentation. Since only 2% of the image area is annotated, the problem is further complicated by many of the usual challenges in dealing with imbalance in data. Also, like many healthcare related applications, false alarm rates are very important, since pathologists often need to report on an image even if there are only one or two areas indicated the potential for cancer. To avoid over-reporting, false alarm rates need to be vanishingly small, and that is often very difficult to achieve for modern ML systems. In this section, we provide some baseline experiments using common-off-the-shelf (COTS) technology to demonstrate the challenges faced in automatic interpretation of digital pathology images.

The data was split into three subsets (/train, /dev and /eval) following standard practices to make the sets mutually exclusive from a patient perspective, and to achieve approximately a 60% train, 20% dev and 20% eval split. We also tried to balance the number of labels occurring in each subset, though in practice there are limits to what can be achieved given the number of labeled regions available in the corpus. We prefer a 60/20/20 split so there are more instances of the less frequently occurring labels in the evaluation set. This gives a much more realistic measure of performance in practice.

Evaluation of performance for this type of data is challenge and often an underappreciated problem. Since the vast majority of an image is BCKG, common popular measures such as the DICE score [40] that score every pixel are not accurate since they over-emphasize BCKG. Instead, we must use weighted approaches [41] that focus only annotated areas and assess the accuracy of the detected boundaries of an annotated patch. Factoring in the accuracy of a segmentation is extremely critical in this type of application.

However, since the images are partially annotated, we must align hypotheses with reference annotations and evaluate the extent to which the partially annotated data is correctly segmented and identified. We

cannot really assess the accuracy of hypotheses that correspond to unannotated regions since we don't know ground truth. Therefore, we have developed a modified version of the DICE measure that compares the degree of overlap between two comparable patches in the reference and hypothesized annotations. This scoring system is available as part of our open source digital pathology tools [42]. We refer to this modified DICE score as MDICE. It is one of two metrics we will use to analyze performance.

We will also analyze performance using a modified F1 (MF1) score. In this case, since there is a predominance of images that do not contain indications of cancerous morphologies, we will use an F1 score averaged across 5 classes of interest (DCIS, INDC, NNEO, INFL and NORM), and one class indicating background (BCKG). We compute the F1 score for each of these two groups for an overall judgment about the image based on the annotated label, weight the averaged F1 score for the first group by 90% and the second group by 10%, and sum them. We ignore NULL, SUSP and ARTF. This modified metric, in our experience, gives a much better estimate of the clinical relevance of a hypothesis.

In this section we will present results using several popular COTS approaches. Though we have generated results for a much wider range of algorithms discussed in [43][44], here we focus on three relatively robust and popular algorithms:

- *ResNet18*: a deep convolutional neural network designed to enable training of very deep architectures by addressing the vanishing gradient problem through the use of residual connections [45][46][47].
- *EB0/EB7*: two variants from the EfficientNet family of convolutional neural networks developed to optimize both accuracy and efficiency (speed, memory, and power usage) [48].
- *ViT-16/ViT-32*: an alternative to convolutional neural networks that analyzes images using 16x16 or 32x32 pixel patches, serializes patches into vectors, and uses a transformer-based architecture for classification [49].

These algorithms were selected based on our experiences that they give good performance across a wide range of signal and image processing applications. The first two are based on convolutional network networks (CNNs), which have proven to be extremely powerful for signal and image applications. The third algorithm is based on a transformer architecture. Transformers, which use self-attention to model context, are widely considered to be a disruptive force in ML due to their ability to encode long-term sequential and spatial dependences [44]. We have adapted these to a variety of tasks involving EEG, digital pathology and cardiology applications.

4.1. Window Size and Sub-Image Selection

Our first set of experiments were designed to optimize the process of selecting an analysis window. The goal with these experiments was to understand how much spatial context was needed to accurately classify a patch. We evaluated two approaches to selecting an image to represent an annotated region. In the first approach, we drew a rectangular bounding box around the image and then experimented with different resampling approaches to make the resulting box a uniform size and/or square (a common shape in image processing applications). We refer to this as the "BBOX" method. Since annotated regions are irregularly shaped, this method has the potential to include significant amounts of tissue that do not share the same label with the reference label. This can result in divergence of a model during training.

The second approach, which we ultimately preferred because of its simplicity and overall good performance, was to compute the center of mass of an annotated region, and then select a window of data centered around the center of mass of the region. We refer to this as the "COM" method. Since annotated regions can be elongated and irregularly shaped, this approach attempted to identify the most significant portion of the region. We included a guard band that allowed the area of the window to be expanded by a percentage (e.g., "g25" means we expand the area by 25%). This has the potential to account for imperfections in the annotations but also can have an adverse effect of including too much out-of-class data

adjacent to the patch of interest. Our hypothesis is that a good ML algorithm should be able to learn to differentiate between correctly labeled data and mislabeled data when it sees enough training data.

The results of several experiments with the COM method are given in Table 8 for TUBR using the MF1 metric. We see that a window of 1024x1024 pixels with no guard band (g00) gives best performance. Results for FCBR are comparable. The results in Table 8 somewhat validate the data in Table 5 in which we see that the majority of the annotated regions for labels of interest are in the 1024x1024 range in size. This also somewhat validates the accuracy of the manual annotations since they seem to contain enough information to reliably classify the window. In fact, experiments with greedy algorithms such as Random Forests have confirmed this in that performance on the training data when overtraining can be quite good. Hence, for the remainder or our experiments, we focused on the com_w1024x1024_g00 data.

Using the com_w1024x1024_g00 data, we evaluated the three popular algorithms mentioned above. The results are shown in Table 9, which also uses the MF1 metric. In this case, the TUBR experiments involved training and evaluating on TUBR. Similarly, the FCBR results were generated by training and evaluating solely on FCBR. We see that EB7 performs slightly better than the other two algorithms. We also see that performance on the dev set generally correlates well with performance on the blind evaluation set. However, there is some evidence of overtraining since the error rates on the training data (closed-set testing) are surprisingly low. On one hand, since the datasets are reasonably large, this suggests there is ample information in these large windows for good classification performance. On the other hand, it suggests overall open-set testing performance could be improved if generalization was improved. Since ViT is a transformer based architecture, we expect it to outperform the other algorithms given enough training data. Therefore, its slightly inferior performance might be an indication the database is not sufficiently large.

4.2. Performance on a Sequential Decoding Task Using a Frame-Based Analysis

Whole-slide pathology images demand models that capture both fine-grained cell morphology and large-scale tissue architecture. Traditional CNNs, which are used in ResNet18 and EB7, excel at capturing local texture patterns, but must rely on very deep stacks (or large filters) to model larger amounts of context. This is inefficient for these extremely high resolution pathology images. In contrast, vision transformers use self-attention to flexibly integrate information across the entire image, encoding long-range dependencies that CNNs inherently miss. Indeed, systematic comparisons report that transformer-based models often outperform CNNs in cancer segmentation tasks precisely because they encode global context [50]. For example, virtually all pure-transformer architectures exceeded CNNs and ResNet architectures in histopathology segmentation, thanks to their image-wide attention [51]. This ability to “see” the whole tissue at once is critical for breast cancer detection, where tumor regions may only be distinguishable by their relationship to distant anatomical landmarks or by subtle patterns that span many cells.

These limitations motivated our exploration of CNNs better suited for such regimes. EfficientNet offers a principled scaling of network depth, width, and resolution to balance representational power and computational demand. While larger variants like EB7, which use approximately 66M parameters, achieve excellent performance, their memory and FLOPS requirements render them impractical for gigapixel whole slide images (WSIs) without hardware acceleration or significant image downsampling. In contrast, EB0, which uses approximately 5.3M parameters, delivers competitive accuracy with significantly reduced computational overhead – an advantage confirmed in our experiments in Table 8 and Table 9, where EB0 performance was competitive with EB7. These findings suggest that in data-scarce histopathology pipelines, smaller, well-scaled CNNs can generalize more reliably than their deeper, more resource-intensive counterpart.

In our next set of baseline experiments, we took the models from the previous experiment and integrated them into a frame-level classification scheme. Previously, the systems were only exposed to perfectly

registrated images centered around the manually derived annotation. This is often what we call an oracle experiment in which we ask the question how good can performance be if the segmentation is perfect. But, as is well known in sequential decoding tasks like speech recognition and EEG interpretation, segmentation is the hardest part of the problem. For this reason, state of the art algorithms often do simultaneous segmentation and classification, so both aspects of the algorithm can be jointly optimized. Hence, in this section, we explore performance when these three algorithms are inserted into a system that iterates over an image frame by frame, uses a 1024x1024 analysis window, and classifies each frame.

Of course, this adds a level of complexity to the task, because these frame-level decisions must be aggregated into patches, and assessed against the manual annotations by comparing the similarities of these patches. Converting frame-level hypotheses to patch-level hypotheses is a well-studied problem and normally involves using a heuristic algorithm to decide which frames should be aggregated, and which frames should be discarded. Typical approaches to this problem include smoothing filters [52], majority filters [53], and morphological techniques based on edges and boundary detection [54]. Ideally, this is a problem best solved using another deep learning system as a postprocessor, but we often lack adequate data to train such systems. We have implemented a variant of these approaches known as flood-fill that uses morphological operations to aggregate similarly labeled frames [55]. We use this in conjunction with MDICE to evaluate the overall performance of a segmentation and classification algorithm.

In Table 10, we provide results for frame-level decoders using a 256x256 frame and a 1024x1024 window, and using a pretrained model from Table 9. For this analysis we use a traditional micro F1 score. We see that EB0 is the most promising approach. Note that these results are not yet optimal because the model being used was only trained on the perfectly registrated images used in Table 9. The model has not been exposed to situations where an analysis window only partially overlaps with an annotated region. Hence, there is still a mismatch between training and decoding. We believe this contributed to the slight degradation in performance for EB7 compared to EB0.

5. Joint Segmentation and Classification

The emergence of powerful statistical methods such as hidden Markov Models (HHMs) in the 1990s [56] demonstrated the value of jointly optimizing segmentation and classification by doing an exhaustive search within the decoding process. These systems share encoders for both localization and diagnostic labeling, facilitating shared feature learning, implicit attention, and reduced error propagation from separate preprocessors. Their ability to localize information in the data made them extremely valuable as diagnostic and discovery tools. Though systems such as SWIN [57][58] and U-NET [59][60] have emerged that emulate this approach in a deep learning context, adapting these systems to high resolution pathology images has proven to be a challenge. However, our unit tests with these algorithms on smaller manageable datasets have shown that their performance is comparable to the algorithms previously discussed.

As an alternative, we have implemented frame-based versions of ResNet, EfficientNet and ViT that are trained on large amounts of pathology data. To avoid data imbalance, we have selected approximately 100,000 windows from each corpus that best represent the diversity of the data and balance the ratio of labels to background. We evaluated these systems on WSIs and show the results in Table 11. We again see that a simpler system, EB0, outperforms the other systems at significantly reduced complexity.

In Table 12, we present a complexity analysis of our three leading systems. The parameter counts inversely correlate with performance – the system with the least number of parameters, EB0, performs best. The computational complexity was compared using the same machine (AMD EPYC™ 7413 Processor 24-core 2.65GHz 128MB Cache) and a single Nvidia A40 GPU. All jobs were run using a single GPU so we could generate a fair comparison. As expected, EB0 is quite fast and produces competitive performance, and the ViT systems use the most resources but lag slightly in performance.

6. Cross-Modal Training Results

The makeup of training data strongly shapes the model architecture and training strategy. Consider two example datasets: a large, heterogeneous collection like TUDP (~100k images of diverse normal and pathological tissues) versus a focused set like FCDP (~14k images emphasizing oncological cases, including a breast cancer subset with detailed annotations). These differences lead to very different learning problems. For instance, FCDP's breast subset contains a high proportion of malignant tissue, whereas TUDP's slides include many normal biopsies and benign findings. A model trained on TUDP would see many more normal patterns, risking that it learns to under-detect cancer (a majority-class bias). By contrast, training on FCDP might make the model sensitive to cancerous features but possibly prone to false positives when deployed on general data

The origin and diversity of images also influence generalization. Models trained on the narrow domain of FCDP's cancer cases might perform very well on similar breast images but poorly on new hospitals or on other tissue types. In contrast, a model trained on TUDP's wide-ranging tissues might be more robust across subtypes. Analogously, multi-center studies (e.g. separate lung or breast histology cohorts) often find that stain differences or scanner artifacts shift the distribution [61]. For example, in a breast histology benchmark the target dataset (BreakHis) contained thousands of patches from one source and had a very different class balance than the small ICIAR source set. In practice, one must watch how differing dataset composition affects the model's sensitivity and specificity. A model trained on high-prevalence cancer data might achieve high sensitivity but at the cost of specificity when applied to general screening images. Conversely, one trained mostly on normal tissue might under-predict disease and suffer low recall. Careful calibration, perhaps through threshold tuning or meta-learning of prevalence, is necessary to maintain an appropriate balance between false positives and false negatives given each dataset's statistics.

In Table 13, we present a series of experiments in which we evaluate mismatched training conditions. In the first two groups of experiments, each system is trained on one corpus and evaluated on both. In the third set of experiments, the systems are trained on the pooled data. We see that best performance is obtained when the training and evaluation data are drawn from the same distribution, and pooling the data results in a small improvement in performance in both cases. This is a lesson learned all too often in ML – learning how to use more data is an important part of the process of improving performance by adding new data. These experiments suggest that more research is needed to understand how best to leverage the pooled dataset into better generalization.

7. Conclusions

The integration of newly annotated breast tissue data from FCDP and TUDP strengthens the robustness and generalizability of machine learning models in digital pathology. These datasets enable training and validation on diverse tissue structures, disease stages, and pathological features. A dual-dataset approach fosters the development of resilient AI models capable of handling image variability, annotation styles, and disease presentations, ultimately enhancing diagnostic precision and adaptability in pathology.

In this chapter we have introduced the science behind the annotation process so that ML researchers can better appreciate the challenges with this data. Improving the performance of an ML system and making that system clinically relevant often requires developing a good understanding of the basic science. We have also established baseline performance of some well-known algorithms. Even though performance has improved dramatically in recent years, it is clear that we have a way to go before the technology will be clinically acceptable.

Going forward, we plan to develop transformer based architectures that encode long-term spatial context. There are numerous computational challenges associated with this since the data must be streamed into the

trainer because it is impossible to fit all the data in computer memory in one time. Hence, we will need to rethink the architecture and workflows so that we can train on the entire corpus using context lengths of 100,000 steps. The amount of memory required to hold sufficient context for each frame is extremely large, so we need to explore architectures that efficiently encode context.

For more information about the resources presented here, please refer to our project web site: *www.nedcdata.org*.

Acknowledgements

This material is based on work supported by several organizations over the years including the National Science Foundation (grants nos. CNS-1726188 and 1925494), the Temple University Catalytic Collaborative Funding Initiative and most recently by the Pennsylvania Breast Cancer Coalition Breast and Cervical Cancer Research Initiative. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations.

References

[1] Bacus, J. V., & Bacus, J. W. (1995). Method and apparatus for automated assay of biological specimens (United States Patent US-5473706-A). url: *https://patents.google.com/patent/US5473706A*.

[2] Kusta, O., Rift, C. V., Risør, T., Santoni-Rugiu, E., & Brodersen, J. B. (2022). Lost in digitization – A systematic review about the diagnostic test accuracy of digital pathology solutions. *Journal of Pathology Informatics*, 13, 100136. doi: *10.1016/j.jpi.2022.100136*.

[3] Xu, H., & Remick, D. G. (2016). Pathology: A Satisfying Medical Profession. *Academic Pathology*, 3, 2374289516661559. doi: *10.1177/2374289516661559*.

[4] Metter, D. M., Colgan, T. J., Leung, S. T., Timmons, C. F., & Park, J. Y. (2019). Trends in the US and Canadian Pathologist Workforces From 2007 to 2017. *JAMA Network Open*, 2(5), Article 5. doi: *10.1001/jamanetworkopen.2019.4337*.

[5] Walsh, E., & Orsi, N. M. (2024). The current troubled state of the global pathology workforce: A concise review. *Diagnostic Pathology*, *19*(1), Article 1. doi: *10.1186/s13000-024-01590-2*.

[6] Jhala, N. (2017). Digital Pathology: Advancing Frontiers. *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. *https://doi.org/10.1109/SPMB.2017.8257013*.

[7] Holloman, A. M., Berg, M. P., Bryant, B., Dixon, L. R., George, M. R., Karp, J. K., Knollmann-Ritschel, B. Ec., Prieto, V. G., Timmons, C. F., Childs, J. M., Lofgreen, A., Johnson, K., & McCloskey, C. B. (2023). Experiential exposure as the key to recruiting medical students into pathology. *Academic Pathology*, 10(2), Article 2. doi: *10.1016/j.acpath.2023.100074*.

[8] *The Paige Prostate Suite: Assistive Artificial Intelligence for Prostate Cancer Diagnosis: Emerging Health Technologies* (EH0123; p. 23). (2024). Canadian Agency for Drugs and Technologies in Health. url: *https://www.ncbi.nlm.nih.gov/books/NBK608438/*.

[9] Matthews, G. A., McGenity, C., Bansal, D., & Treanor, D. (2024). Public evidence on AI products for digital pathology. *Npj Digital Medicine*, 7(1), Article 1. doi: *10.1038/s41746-024-01294-3*.

[10] Bellis, M., Metias, S., Naugler, C., Pollett, A., Jothy, S., & Yousef, G. M. (2013). Digital Pathology: Attitudes and practices in the Canadian pathology community. *Journal of Pathology Informatics*, 4(1), Article 1. doi: *10.4103/2153-3539.108540*.

[11] Clarke, E., Doherty, D., Randell, R., Grek, J., Thomas, R., Ruddle, R. A., & Treanor, D. (2023). Faster than light (microscopy): Superiority of digital pathology over microscopy for assessment of immunohistochemistry. *Journal of Clinical Pathology*, *76*(5), 333. doi: *10.1136/jclinpath-2021-207961*.

[12] McGenity, C., Clarke, E. L., Jennings, C., Matthews, G., Cartlidge, C., Freduah-Agyemang, H., Stocken, D. D., & Treanor, D. (2024). Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy. *Npj Digital Medicine*, *7*(1), Article 1. doi: *https://doi.org/10.1038/s41746-024-01106-8*.

[13] Shawki, N., Shadhin, M. G. M., Elseify, T., Jakielaszek, L., Farkas, T., Persidsky, Y., Jhala, N., Obeid, I., & Picone, J. (2020). The Temple University Digital Pathology Corpus. In I. Obeid, I. Selesnick, & J. Picone (Eds.), *Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications* (1st ed., pp. 67–104). doi: *10.1007/978-3-030-36844-9*.

[14] Doshna, B., Wevodau, Z., Jhala, N., Akhtar, I., Obeid, I., & Picone, J. (2021). The Temple University Digital Pathology Corpus: The Breast Tissue Subset. In I. Obeid, I. Selesnick, & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1–3). IEEE. doi: *https://doi.org/10.1109/SPMB52430.2021.9672275*.

[15] Shalamzari, S. S., Bagritsevich, M., Melles, Anne-Mai, Obeid, I., Picone, J., Connolly, D., Wu, C., Brown, B., James, J., Gong, Y., & Wu, H. (2023). Big Data Resources for Digital Pathology. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 1–19. doi: *10.1109/SPMB59478.2023.10372721*.

[16] Bagritsevich, M., Hackel, D., Obeid, I., & Picone, J. (2024). Annotation of the Fox Chase Cancer Center Digital Pathology Corpus. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 1–4. doi: *10.1109/SPMB62441.2024.10842255*.

[17] Rolls, G. (2019). An Introduction to Specimen Preparation. In *Leica Biosystems* (p. 1). url: *https://www.leicabiosystems.com/pathologyleaders/an-introduction-to-specimen-preparation*.

[18] Anderson, J. (2019). An Introduction to Routine and Special Staining. In *Leica Biosystems* (p. 1). url: *https://www.leicabiosystems.com/pathologyleaders/an-introduction-to-routine-and-special-staining*.

[19] Niyas, S., Bygari, R., Naik, R., Viswanath, B., Ugwekar, D., Mathew, T., Kavya, J., Kini, J. R., & Rajan, J. (2023). Automated Molecular Subtyping of Breast Carcinoma Using Deep Learning Techniques. *IEEE Journal of Translational Engineering in Health and Medicine*, 11, 161–169. doi: *10.1109/JTEHM.2023.3241613*.

[20] Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33, 170–175. doi: *10.1016/j.media.2016.06.037*.

[21] Wu, H., Phan, J. H., Bhatia, A. K., Cundiff, C. A., Shehata, B. M., & Wang, M. D. (2015). Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 727–730. doi: *10.1109/EMBC.2015.7318465*.

[22] Leica Biosystems. (2008). *ScanScope SVS file format specifications*. url: *https://www.leicabiosystems.com/products/whole-slide-imaging/scanscope-svs-file-format/*.

[23] Goode, A., Gilbert, B., Harkes, J., Jukic, D., & Satyanarayanan, M. (2013). OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, 4(1), Article 1. doi: *10.4103/2153-3539.119005*.

[24] Obeid, I., & Picone, J. (2016). The Temple University Hospital EEG Data Corpus. In M. A. Lebedev (Ed.), *Augmentation of Brain Function: Facts, Fiction and Controversy. Volume I: Brain-Machine Interfaces* (1st ed., Vol. 10, pp. 394–398). Frontiers Media S.A. doi: *10.3389/fnins.2016.00196*.

[25] Simons, J., Wevodau, Z., Doshna, B., Obeid, I., & Picone, J. (2021). *The Temple University Hospital DPATH Corpus: Annotation Guidelines* (p. 18). Temple University. url: *https://isip.piconepress.com/publications/reports/2021/tuh_dpath/annotations/*.

[26] Hoda, S. A., Rosen, P. P., Brogi, E., & Koerner, F. C. (2020). *Rosen's Breast Pathology*. Wolters Kluwer Health. url: *https://shop.lww.com/Rosen-s-Breast-Pathology/p/9781496398918*.

[27] Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological Image Analysis: A Review. *IEEE Reviews in Biomedical Engineering*, *2*, 147–171. doi: *10.1109/RBME.2009.2034865*.

[28] Raghavendra, A. S., Bassett, R., Damodaran, S., Barcenas, C. H., Mouabbi, J. A., Layman, R., & Tripathy, D. (2025). Clinical Characteristics and Survival Outcomes of Metastatic Invasive Lobular and Ductal Carcinoma. *JAMA Network Open*, *8*(4), e251888. doi: *10.1001/jamanetworkopen.2025.1888*.

[29] Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 16(Pt 2), 411–418. doi: *10.1007/978-3-642-40763-5_51*.

[30] Condrescu, M., Opuni, K., Hantash, B. M., & Reeves, J. P. (2002). Cellular regulation of sodium-calcium exchange. *Annals of the New York Academy of Sciences*, 976, 214–223. doi: *10.1111/J.1749-6632.2002.TB04744.X*.

[31] Cross, B. M., Breitwieser, G. E., Reinhardt, T. A., & Rao, R. (2014). Cellular calcium dynamics in lactation and breast cancer: From physiology to pathology. *American Journal of Physiology-Cell Physiology*, 306(6), Article 6. doi: *10.1152/ajpcell.00330.2013*.

[32] Hackel, D., Bagritsevich, M., Dumitrescu, C., Al Mamun, Md. A., Purba, S. A., Heathcote, D., Obeid, I., & Picone, J. (2026). Enabling Microsegmentation: Digital Pathology Corpora for Advanced Model Development. In *Signal Processing in Medicine and Biology: Applications of Artificial Intelligence in Medicine and Biology* (Vol. 1, p. 50). Springer. url: *https://isip.piconepress.com/publications/book_sections/2026/springer/dpath/*. (in review).

[33] Sapino, A., & Kulka, J. (Eds.). *Breast Pathology* (1st ed. 2020). Cham : Springer International Publishing : Imprint: Springer, 2020. url: *https://link.springer.com/referencework/10.1007/978-3-319-62539-3*.

[34] Tozbikian, G., Brogi, E., Vallejo, C. E., Giri, D., Murray, M., Catalano, J., Olcese, C., Van Zee, K. J., & Wen, H. Y. (2017). Atypical Ductal Hyperplasia Bordering on Ductal Carcinoma In Situ. *International Journal of Surgical Pathology*, 25(2), 100–107. doi: *10.1177/1066896916662154*.

[35] Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N. A., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., O'Malley, F. P., & Weaver, D. L. (2015). Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *Journal of the American Medical Association*, 313(11), 1122–1132. doi: *10.1001/JAMA.2015.1405*.

[36] Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical

image analysis. *Artificial Intelligence in Medicine*, 150, 102830. doi: *10.1016/j.artmed.2024.102830*.

[37] Yin, L., Duan, J.-J., Bian, X.-W., & Yu, S. (2020). Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research*, 22(1), 61. doi: *10.1186/s13058-020-01296-5*.

[38] Nong, L., Zhang, Z., Xiong, Y., Zheng, Y., Li, X., Li, D., He, Q., & Li, T. (2019). Comparison of next-generation sequencing and immunohistochemistry analysis for targeted therapy-related genomic status in lung cancer patients. *Journal of Thoracic Disease*, 11(12), 4992–5003. doi: *10.21037/jtd.2019.12.25*.

[39] "ICD-10-CM, Official Guidelines for Coding and Reporting", Centers for Medicare & Medicaid Services (CMS), January 01, 2020, url: *https://www.hhs.gov/guidance/sites/default/files/hhs-guidance-documents/ICD-10-CM_Guidelines-FY2020_final.pdf*.

[40] Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., & Li, J. (2020). Dice Loss for Data-imbalanced NLP Tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 465–476. doi: *10.18653/v1/2020.acl-main.45*.

[41] Jadon, S. (2020). A survey of loss functions for semantic segmentation. *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 1–7. doi: *10.1109/CIBCB48159.2020.9277638*.

[42] Bagritsevich, M., Picone, J., & Obeid, I. (2024). *The TUH Digital Pathology Corpus*. url: *https://isip.piconepress.com/projects/nedc/html/tuh_dpath/*.

[43] Thai, B., McNicholas, S., Shalamzari, S. S., Meng, P., & Picone, J. (2023). Towards a More Extensible Machine Learning Demonstration Tool. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 1–4. doi: *10.1109/SPMB59478.2023.10372731*.

[44] Thundiyil, S. C., & Picone, J. (2025). Time Series Analysis from Classical Methods to Transformer-Based Approaches: A Review. In *Signal Processing in Medicine and Biology: Applications of Artificial Intelligence in Medicine and Biology* (Vol. 1, p. 56). Springer. url: *https://isip.piconepress.com/publications/book_sections/2024/springer/transformers/*. (in publication).

[45] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. doi: *10.1109/CVPR.2016.90*.

[46] Khalkhali, V., Shawki, N., Shah, V., Golmohammadi, M., Obeid, I., & Picone, J. (2021). Low Latency Real-Time Seizure Detection Using Transfer Deep Learning. In I. Obeid, I. Selesnick, & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1–7). IEEE. doi: *10.1109/SPMB52430.2021.9672285*.

[47] Alexandrov, D., & Picone, J. (2024). The Impact of ECG Channel Reduction on Multi-Label Cardiac Diagnosis. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 6. doi: *10.1109/SPMB62441.2024.10842233*.

[48] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the International Conference on Machine Learning (ICML)* (Vol. 97, pp. 6105–6114). PMLR. url: *http://proceedings.mlr.press/v97/tan19a.html*.

[49] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth

16x16 Words: Transformers for Image Recognition at Scale. *Proceedings of the International Conference on Learning Representations (ICLR)*, 1–21. url: *https://iclr.cc/virtual/2021/oral/3458*.

[50] Atabansi, C. C., Nie, J., Liu, H., Song, Q., Yan, L., & Zhou, X. (2023). A survey of Transformer applications for histopathological image analysis: New developments and future directions. *BioMedical Engineering OnLine*, *22*(1), 96. doi: *10.1186/s12938-023-01157-0*.

[51] Cam Nguyen, Zuhayr Asad, Ruining Deng, & Yuankai Huo. (2022). Evaluating transformer-based semantic segmentation networks for pathological image segmentation. *Proceedings of the SPIE* 12032, Medical Imaging 2022: Image Processing, 120323N. doi: *10.1117/12.2611177*.

[52] Gonzalez, R., & Woods, R. (2017). *Digital Image Processing*. Pearson Deutschland. url: *https://elibrary.pearson.de/book/99.150005/9781292223070*.

[53] Janz, A., Jakimow, B., Thiel, F., Goswami, A., van der Linden, S., & Hostert, P. (2025, June 15). *Generic Filter (Majority)*. EnMAP-Box 3 Documentation. url: *https://enmap-box.readthedocs.io/en/latest/usr_section/usr_cookbook/generic_filter.html*.

[54] ArcGIS Pro 3.5. (2025, June 15). *Smoothing zone edges with Boundary Clean and Majority Filter*. ArcGIS Pro. url: *https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/smoothing-zone-edges-with-boundary-clean-and-majority-filter.htm*.

[55] Nghiem, Y., Berman, L., & Bulik, A. (2024). Machine Learning in Digital Pathology. *Senior Design I, College of Engineering, Temple University*, 1–37. url: *https://isip.piconepress.com/publications/presentations_misc/2024/senior_design/mladp*.

[56] Picone, J. (1990). Continuous Speech Recognition Using Hidden Markov Models. *IEEE Acoustics, Speech, and Signal Processing Society (ASSP)*, *7*(3), 26–41. Doi: *10.1109/53.54527*.

[57] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. doi: *10.1109/ICCV48922.2021.00986*.

[58] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., & Guo, B. (2022). Swin Transformer V2: Scaling Up Capacity and Resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11999–12009. doi: *10.1109/CVPR52688.2022.01170*.

[59] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 234–241). Springer International Publishing. doi: *10.1007/978-3-319-24574-4_28*.

[60] Cui, J., Guo, H., Wang, H., Chen, F., Shu, L., & Li, L. C. (2020). Fully-automatic segmentation of coronary artery using growing algorithm. *Journal of X-Ray Science and Technology*, *28*(6), 1171–1186. doi: *10.3233/XST-200707*.

[61] Asadi-Aghbolaghi, M., Darbandsari, A., Zhang, A., Contreras-Sanz, A., Boschman, J., Ahmadvand, P., Köbel, M., Farnell, D., Huntsman, D. G., Churg, A., Black, P. C., Wang, G., Gilks, C. B., Farahani, H., & Bashashati, A. (2024). Learning generalizable AI models for multi-center histopathology image classification. *NPJ Precision Oncology*, *8*(1), 151. doi: *10.1038/s41698-024-00652-4*.

List of Tables

Table 1. An overview of publicly available breast histological H&E corpora

Dataset	Size (GB)	Subjects	Annotations	No. Samples	Data Origin	Pixel size (µm/pixel)
ACROBAT	1164	1152	Invasive Cancer, Non-Malignant, Artefacts, Ductal Carcinoma, Lobular Carcinoma, Normal	4212 WSIs	Karolinska Institutet (SE)	0.91
BACH / ICIAR 2018	13	-	Normal, Benign, InSitu, Invasive	400 slides 30 WSIs	Ipatimup and INEB (PT)	0.50
BCNB	33	1058	Coordinates of Tumor	1058 WSIs	Beijing University of Posts and Telecommunications (CN)	-
BRACS	1100	189	Benign, Ductal Hyperplasia (2), Flat Epithelial Ductal Carcinoma, Invasive Carcinoma, Normal	547 WSIs	Institute for High Performance Computing and Networking (IT)	0.25
BreaKHist	2	82	4 Benign, 4 Malignant subtypes	7909 slides	P&D Lab, Federal University of Parana (Brazil)	multiple
CAMELYON16	1160	399	Tumor, Normal	400 WSI	Radboud University Medical Center, UMC Utrecht (NL)	0.24
CAMELYON17	2950	200	Negative, ITC, Micro-metastasis, Macro-metastasis	1399 WSIs	Radboud University Medical Center, UMC Utrecht (NL)	0.24
CPTAC-BRCA	113	134	Cancer, Normal	642 WSIs	Clinical Proteomic Tumor Analysis Consortium (USA)	0.25 0.50
GTEx-breast	80	894	Pathology (unlocated)	894 WSIs	Broad Institute of MIT and Harvard (USA)	0.50
HER2 / Warwick	20	86	HER2 Score	172 WSIs	University of Warwick, University of Nottingham, and AIDPATH consortium (UK)	0.23
HEROHE	820	360	Binary	500 WSIs	Institute of Research and Innovation in Health from Porto (PT)	0.24
SPIDER	85		18 morphologies	1925 slides	HistAI initiative	0.40
TCGA-BRCA	1640	1098	Nuclei segmentations	3111 WSIs	The Cancer Genome Atlas (USA)	0.25
TIGER	169	370	Tumor, Stroma, Lymphocytes, Necrosis	370 WSIs	Radboud University Medical Center of Nijmegen (NL), Jules Bordet Institut (BE) and TCGA-BRCA	0.50
TUBR	**1226**	**296**	**Artifact, Background, Ductal Carcinoma, Invasive Ductal Carcinoma, Inflammation, Nonneoplastic, Normal, Null, Suspicious**	**3505**	**Temple University Hospital (USA)**	**0.20**
FCBR	**336**	**1397**		**1463**	**Fox Chase Cancer Center (USA)**	**0.20**

Table 2. Labels for the TUH and FCCC breast tissue corpora

Label	Description / Features
Normal (NORM)	normal ducts and lobules
Ductal Carcinoma in Situ (DCIS)	ductal carcinoma in situ, and lobular carcinoma in situ
Invasive Ductal Carcinoma (INDC)	invasive ductal carcinoma, invasive lobular carcinoma, and invasive mammary carcinoma
Non-Neoplastic (NNEO)	fibrosis, hyperplasia, intraductal papilloma, adenosis, ectasia, etc.
Inflammation (INFL)	areas of inflammation
Artifact (ARTF)	grease pen marks, stitches, foreign bodies, etc.
Indistinguishable (NULL)	indistinguishable tissue, normally due to issues with the cut/stain
Suspected (SUSP)	regions that are at risk of developing into cancerous regions
Background (BCKG)	stroma, no ducts or lobules

Table 3. A comparison of TUDP and FCDP

Attribute	TUDP (v1.0.0)	FCDP (v1.0.0)
No. Files	99,123	14,276
Amount of Data	21.123 GB	3.497 GB
Average File Size	456 MB	237 MB
Approximate Composition (top four categories)	C64 (kidney): 16% C61 (prostate): 13% C34 (lung): 10% C50 (breast): 10%	urinary/prostate: 37% breast: 21% gastro: 18% gyneco: 8%

Table 4. A comparison of the annotations for TUBR and FCBR

Attribute	TUBR (v5.0.0)	FCBR (v3.0.1)
No. Files:		
/train	1,652	765
/dev	932	373
/eval	921	325
TOTAL	3,505	1,463
Amount of Data (Gbytes)	1.226	3.497
No. Labels	22,283	20,085
Avg. No. Labels Per Slide	6.36	13.73
Annotated Area Per Slide (%)	1.96%	2.53%
No. Labels:		
norm	4,797 [21.53%]	325 [1.62%]
dcis	1,048 [4.70%]	1,209 [6.02%]
indc	1,512 [6.79%]	10,955 [54.54%]
nneo	7,162 [32.14%]	728 [3.62%]
infl	970 [4.35%]	1,215 [6.05%]
artf	1,118 [5.02%]	914 [4.55%]
null	618 [2.77%]	1,090 [5.43%]
susp	115 [0.52%]	24 [0.12%]
bckg	4,953 [22.18%]	3,625 [18.05%]
Area of Labels:		
64 x 64	2 [0.01%]	250 [1.24%]
128 x 128	130 [0.58%]	1,001 [4.98%]
256 x 256	1,132 [5.08%]	2,677 [13.33%]
512 x 512	4,346 [19.50%]	5,051 [25.15%]
1,024 x 1,024	6,555 [29.42%]	5,651 [28.14%]
2,048 x 2,048	4,892 [21.95%]	3,764 [18.74%]
4,096 x 4,096	3,372 [15.13%]	1,363 [6.79%]
8,192 x 8,192	1,433 [6.43%]	304 [1.51%]
16,384 x 16,384	386 [1.73%]	23 [0.11%]
32,768 x 32,768	35 [0.16%]	1 [0.00%]
65,536 x 65,536	0 [0.00%]	0 [0.00%]

Table 5. A comparison of annotation dimensions

Label	TUBR	FCBR
dcis:		
64 x 64	0 [0.00%]	3 [0.25%]
128 x 128	0 [0.00%]	91 [7.53%]
256 x 256	5 [0.48%]	147 [12.16%]
512 x 512	137 [13.07%]	311 [25.72%]
1,024 x 1,024	970 [4.35%]	427 [35.32%]
2,048 x 2,048	346 [33.02%]	210 [17.37%]
4,096 x 4,096	91 [8.68%]	19 [1.57%]
8,192 x 8,192	7 [0.67%]	1 [0.08%]
16,384 x 16,384	0 [0.00%]	0 [0.00%]
32,768 x 32,768	0 [0.00%]	0 [0.00%]
65,536 x 65,536	0 [0.00%]	0 [0.00%]
indc:		
64 x 64	0 [0.00%]	4 [0.04%]
128 x 128	4 [0.26%]	132 [1.20%]
256 x 256	47 [3.11%]	1,031 [9.41%]
512 x 512	179 [11.84%]	3,093 [28.23%]
1,024 x 1,024	450 [29.76%]	3,545 [32.26%]
2,048 x 2,048	551 [36.44%]	2,181 [19.91%]
4,096 x 4,096	254 [16.80%]	790 [7.21%]
8,192 x 8,192	26 [1.72%]	169 [1.54%]
16,384 x 16,384	1 [0.07%]	10 [0.09%]
32,768 x 32,768	0 [0.00%]	0 [0.00%]
65,536 x 65,536	0 [0.00%]	0 [0.00%]
nneo:		
64 x 64	0 [0.00%]	0 [0.00%]
128 x 128	2 [0.03%]	8 [1.10%]
256 x 256	126 [1.76%]	133 [18.27%]
512 x 512	1,398 [19.52%]	266 [36.54%]
1,024 x 1,024	3,010 [42.03%]	223 [30.63%]
2,048 x 2,048	1,933 [26.99%]	82 [11.26%]
4,096 x 4,096	597 [8.34%]	15 [2.06%]
8,192 x 8,192	84 [1.17%]	1 [0.14%]
16,384 x 16,384	12 [0.17%]	0 [0.00%]
32,768 x 32,768	0 [0.00%]	0 [0.00%]
65,536 x 65,536	0 [0.00%]	0 [0.00%]
infl:		
64 x 64	0 [0.00%]	66 [0.00%]
128 x 128	74 [7.63%]	310 [0.26%]
256 x 256	225 [23.20%]	453 [3.11%]
512 x 512	337 [34.74%]	282 [11.84%]
1,024 x 1,024	277 [28.56%]	9 [29.76%]
2,048 x 2,048	50 [5.15%]	1 [36.44%]
4,096 x 4,096	7 [0.72%]	0 [16.80%]
8,192 x 8,192	0 [0.00%]	0 [1.72%]
16,384 x 16,384	0 [0.00%]	0 [0.07%]
32,768 x 32,768	0 [0.00%]	0 [0.00%]
65,536 x 65,536	0 [0.00%]	0 [0.00%]

Table 6. A histogram of tissue site and tumor site codes for FCBR

Tissue Site (ICDO Code)	Tumor Site (ICDO Code)	Count
c50.9	c50.4	334
c50.9	c50.8	278
c50.9	c50.9	158
c50.4	c50.4	155
c50.8	c50.8	92
c50.9	c50.2	86
c50.9	c50.5	70
c50.9	c50.1	62
c50.2	c50.2	55
c50.9	c50.3	50
c50.5	c50.5	36
c50.3	c50.3	26
c50.1	c50.1	25
c50.9	cxx.x	20
c50.0	c50.0	3
c50.9	c50.6	2
c50.9	c50.0	2
c50.1	c50.4	2
c50.9	c63.9	1
c50.8	c50.4	1
c50.8	cxx.x	1
c50.6	c50.6	1
c50.5	c50.4	1
c50.4	cxx.x	1
c50.3	c50.2	1

Table 7. The top 10 entries for derived grades for FCBR

Derived Grade	Behavior	ICDO (Histology)	Cancer Status	Count
hg	MALIG-PRIMARY	8500/3	No Evidence of this cancer	399
hg	MALIG-PRIMARY	8500/3	Evidence of this cancer	170
hg	MALIG-PRIMARY	8500/3	Unknown, indeterminate whether this cancer present	91
hg	MALIG-PRIMARY	8522/3	No Evidence of this cancer	34
hg	MALIG-PRIMARY	8520/3	No Evidence of this cancer	31
hg	MALIG-PRIMARY	8523/3	No Evidence of this cancer	19
hg	MALIG-PRIMARY	8520/3	Evidence of this cancer	18
hg	MALIG-PRIMARY	8522/3	Evidence of this cancer	14
hg	MALIG-PRIMARY	8520/3	Unknown, indeterminate whether this cancer present	8
hg	CA IN SITU	8501/2	No Evidence of this cancer	7
ig	MALIG-PRIMARY	8500/3	No Evidence of this cancer	220
ig	MALIG-PRIMARY	8520/3	No Evidence of this cancer	65
ig	MALIG-PRIMARY	8522/3	No Evidence of this cancer	35
ig	MALIG-PRIMARY	8500/3	Evidence of this cancer	33
ig	MALIG-PRIMARY	8500/3	Unknown, indeterminate whether this cancer present	30
ig	MALIG-PRIMARY	8522/3	Evidence of this cancer	14
ig	MALIG-PRIMARY	8520/3	Evidence of this cancer	13
ig	CA IN SITU	8523/2	No Evidence of this cancer	8
ig	MALIG-PRIMARY	8524/3	No Evidence of this cancer	5
ig	MALIG-PRIMARY	8523/3	No Evidence of this cancer	5
lg	MALIG-PRIMARY	8500/3	No Evidence of this cancer	23
lg	MALIG-PRIMARY	8480/3	No Evidence of this cancer	5
lg	MALIG-PRIMARY	8500/3	Unknown, indeterminate whether this cancer present	4
lg	MALIG-PRIMARY	8523/3	No Evidence of this cancer	3
lg	MALIG-PRIMARY	8520/3	No Evidence of this cancer	3
lg	MALIG-PRIMARY	8522/3	No Evidence of this cancer	2
lg	MALIG-PRIMARY	9020/3	Evidence of this cancer	1
lg	MALIG-PRIMARY	8523/3	Unknown, indeterminate whether this cancer present	1
lg	MALIG-PRIMARY	8522/3	Evidence of this cancer	1
lg	MALIG-PRIMARY	8522/3	Unknown, indeterminate whether this cancer present	1
UNK	UNKNOWN			29

Table 8. Patch-level classification results (MF1) as a function of the window dimensions for TUBR

Arch	Dataset	Train	Dev	Eval
eb0	com_w0128x0128_g00	22.42	39.91	43.01
eb0	com_w0128x0128_g25	13.89	35.42	39.87
eb0	com_w0128x0128_g50	9.09	35.31	40.45
eb0	com_w0256x0256_g00	4.13	30.45	34.56
eb0	com_w0256x0256_g25	15.51	25.87	32.03
eb0	com_w0256x0256_g50	8.16	28.69	35.02
eb0	com_w0512x0512_g00	4.95	25.53	31.43
eb0	com_w0512x0512_g25	15.77	28.22	33.94
eb0	com_w0512x0512_g50	7.87	26.31	32.68
eb0	com_w1024x1024_g00	8.12	22.43	28.21
eb0	com_w1024x1024_g25	13.15	26.92	30.11
eb0	com_w1024x1024_g50	13.03	25.33	32.56
eb7	com_w0128x0128_g00	8.98	34.97	38.32
eb7	com_w0128x0128_g25	13.41	34.20	39.59
eb7	com_w0128x0128_g50	11.42	31.48	36.46
eb7	com_w0256x0256_g00	8.28	29.36	32.52
eb7	com_w0256x0256_g25	8.57	25.24	30.03
eb7	com_w0256x0256_g50	6.81	28.85	33.56
eb7	com_w0512x0512_g00	8.44	22.06	27.77
eb7	com_w0512x0512_g25	5.71	22.23	26.57
eb7	com_w0512x0512_g50	9.03	21.82	29.15
eb7	com_w1024x1024_g00	4.46	21.86	25.95
eb7	com_w1024x1024_g25	3.63	23.66	27.36
eb7	com_w1024x1024_g50	3.48	20.58	25.31

Table 9. A comparison (MF1) of selected popular algorithms on window-based classification

	TUBR			FCBR		
Arch	Train	Dev	Eval	Train	Dev	Eval
ResNet18	15.03	25.51	34.01	21.46	33.73	31.91
EB0	8.12	22.43	28.22	21.09	27.81	29.40
EB7	0.01	22.06	25.39	3.67	26.09	24.85
ViT-16	0.17	20.66	30.08	0.08	27.87	27.56
ViT-32	5.28	24.94	29.54	10.55	26.28	27.33

Table 10. A comparison of decoder performance (MDICE) using pretrained models

	TUBR			FCBR		
Arch	Train	Dev	Eval	Train	Dev	Eval
ResNet18	45.21	41.98	39.13	51.97	45.98	48.50
EB0	69.83	56.60	51.72	62.32	54.49	55.02
EB7	67.17	55.67	50.72	57.98	47.00	49.39
ViT-16	47.75	39.00	35.23	50.26	41.78	41.58
ViT-32	57.30	48.57	43.64	52.24	43.64	44.65

Table 11. The performance of optimized systems (MDICE)

	TUBR			FCBR		
Arch	Train	Dev	Eval	Train	Dev	Eval
ResNet18	40.87	37.42	33.28	44.40	44.34	42.11
EB0	62.32	55.31	51.10	66.83	49.49	46.02
EB7	56.03	47.10	43.81	31.53	31.31	32.41
ViT-16	62.68	48.32	44.72	26.11	24.44	25.07
ViT-32	57.12	46.34	41.39	29.58	26.11	28.06

Table 12. Complexity analysis for three popular ML algorithms

Arch	No. Parameters	Training Time	Decoding Time
ResNet18	11.69M	2,356 (1.0x)	1,597 (1.0x)
EB0	5.30M	3,900 (1.7x)	1,520 (1.0x)
EB7	66.35M	4,883 (2.1x)	1,735 (1.1x)
ViT-16	86.60M	3,713 (1.6x)	1,726 (1.1x)
ViT-32	88.20M	3,869 (1.6x)	1,623 (1.0x)

Table 13. The impact of mismatched training conditions (MDICE)

Arch	Training Condition	TUBR			FCBR		
		Train	Dev	Eval	Train	Dev	Eval
ResNet18	TUBR	40.87	37.43	33.28	31.52	29.05	31.66
EB7	TUBR	56.03	47.10	43.81	31.53	31.31	32.41
ViT-32	TUBR	57.12	46.34	41.39	30.48	28.49	31.19
ResNet18	FCBR	47.42	48.03	43.85	44.40	44.34	42.11
EB7	FCBR	30.76	32.08	33.66	73.31	53.22	55.82
ViT-32	FCBR	17.28	18.38	17.43	29.58	26.11	28.06
ResNet18	TUBR + FCBR	37.99	35.12	32.18	32.72	30.05	32.87
EB7	TUBR + FCBR	75.94	64.08	62.47	55.11	41.08	40.85
ViT-32	TUBR + FCBR	47.97	38.93	35.75	37.62	31.52	33.88

List of Figures

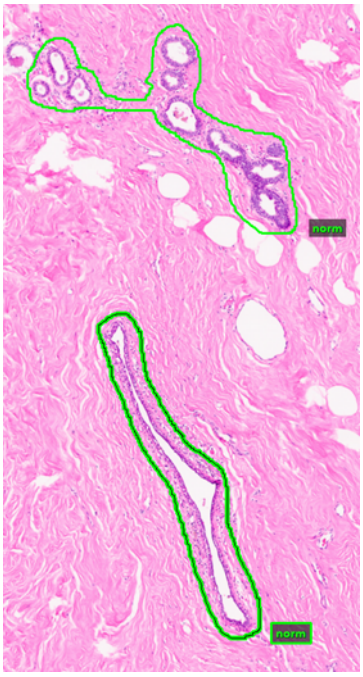

Figure 1. A normal terminal duct unit surrounded by healthy fibrous stroma

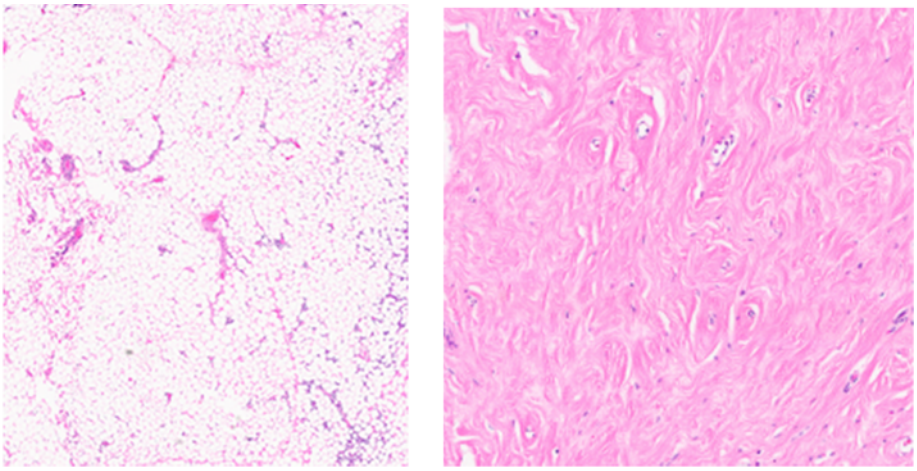

Figure 2. Examples of fatty stroma (left) and fibrous stroma (right)

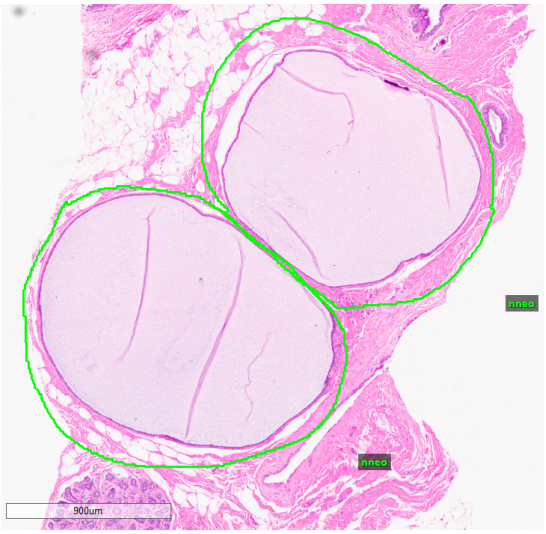

Figure 3. Microcysts annotated as NNEO

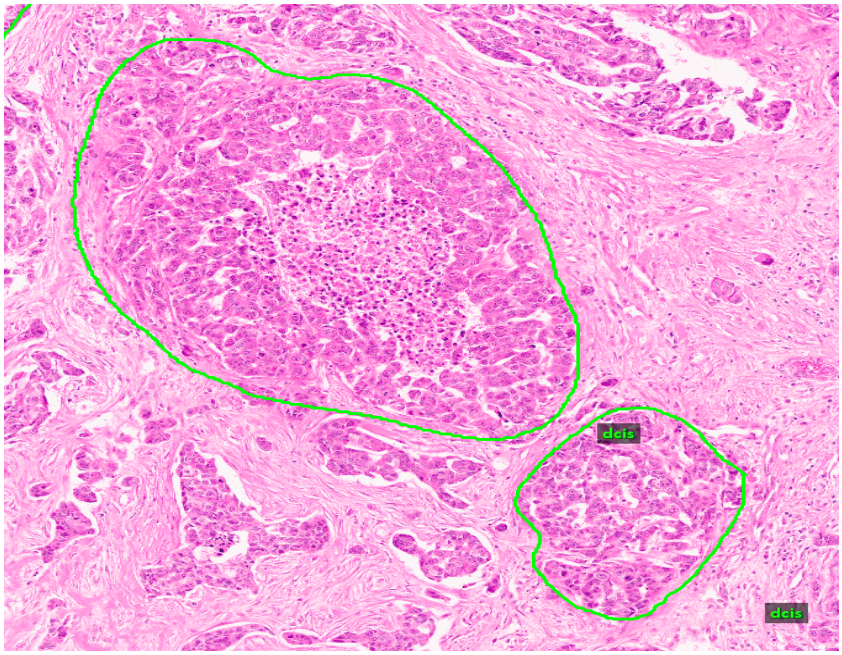

Figure 4. A typical DCIS presentation of malignant epithelial cells confined to the ductal-lobular system

Figure 5. A TDLU undergoing non-neoplastic changes (gynecomastoid hyperplasia and benign proliferation)

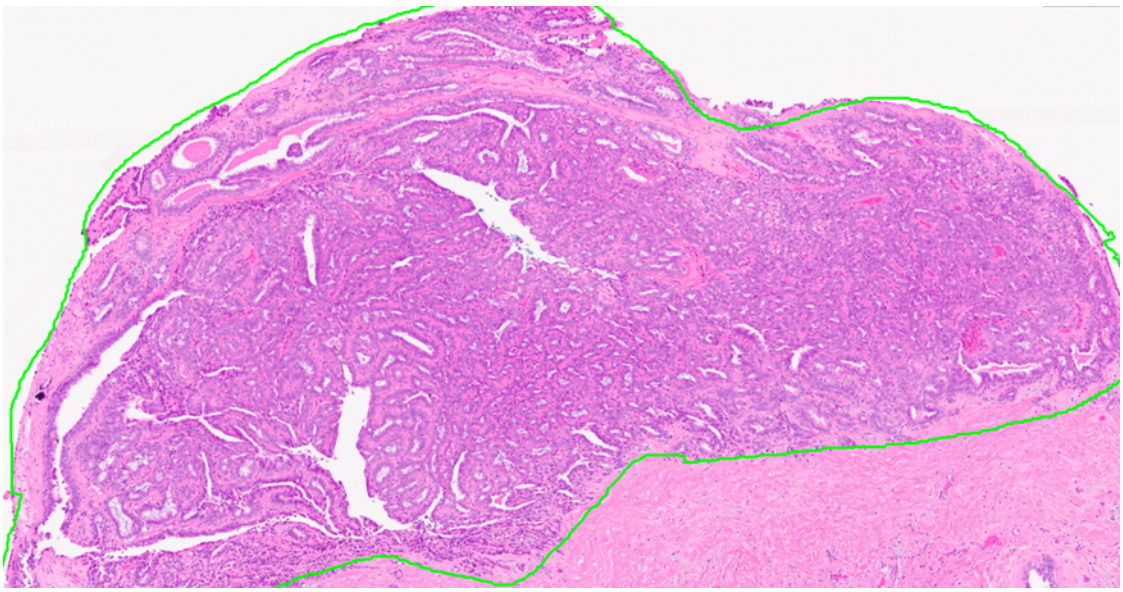

Figure 6. A benign intraductal papilloma is annotated with an NNEO label that encompasses the entire papilloma.

Don’t delete this line