Transforming Breast Cancer Prognosis with Machine Learning and Digital Pathology

by

Majoie Ngandi

A Signature Honors Project Presented to the

Honors College

East Carolina University

In Partial Fulfillment of the

Requirements for

Graduation with Honors

by

Majoie Ngandi

Greenville, NC

May, 2025

Approved by:

Nic Herndon

Department of Computer Science

Abstract

Breast cancer continues to be a leading contributor to cancer-related mortality among women globally. A key aspect of effective clinical management is the ability to accurately predict the risk of recurrence following initial treatment, which informs therapeutic decisions and helps balance the benefits and risks of intervention. Traditional recurrence prediction relies heavily on expensive genomic assays or time-intensive manual pathology assessments, limiting their scalability and accessibility. With the growing availability of digitized histopathology slides and advancements in machine learning, computational approaches have emerged as promising alternatives. This paper reviews current methodologies in breast cancer recurrence prediction using histopathology images and examines the evolution from region-based models to annotation-free and pixel-level strategies. It highlights both the opportunities and challenges associated with deep learning in this domain and identifies emerging trends in the field. By synthesizing recent research, this work aims to provide a foundational understanding of the field and outline directions for future development in accessible, automated recurrence risk assessment.

1. Introduction

1.1. Background and Significance of Breast Cancer

Breast cancer is the most frequently diagnosed cancer among women globally, representing a major public health concern. In 2022, approximately 2.3 million women were diagnosed with breast cancer worldwide, and the disease was responsible for about 670,000 deaths. This makes it the leading cause of cancer-related mortality among women in many regions (World Health Organization [WHO], 2024). In the United States, breast cancer remains the most commonly diagnosed cancer in women after skin cancer. The American Cancer Society (ACS, 2024) estimated that in 2024, there would be approximately 310,720 new cases of invasive breast cancer and 56,500 cases of ductal carcinoma in situ (DCIS) among U.S. women. The lifetime risk for a woman in the U.S. of developing invasive breast cancer is about 13%, or 1 in 8 (ACS, 2024). Despite significant advances in early detection and treatment, which have contributed to a decline in mortality rates, disparities in access to care and outcomes persist (ACS, 2024; WHO, 2024). Looking ahead, the global burden of breast cancer is expected to grow. Projections indicate that by 2050, there will be approximately 3.2 million new cases and 1.1 million deaths annually, representing increases of 38% and 68%, respectively, from current figures (WHO, 2024). This projected surge underscores the urgent need for enhanced prevention, early diagnosis, and equitable treatment strategies to reduce the impact of breast cancer on women's health worldwide.

1.2. Importance of Recurrence Risk Prediction

One of the most critical components of post-treatment care is the ability to predict the risk of breast cancer recurrence. Recurrence is a leading contributor to mortality, and accurately assessing this risk enables clinicians to customize treatment intensity, ensuring high-risk patients receive the aggressive care they need and low-risk patients are spared unnecessary interventions.

This personalized approach reduces over-treatment and its associated physical, psychological, and financial burdens. Moreover, recurrence predictions have implications for healthcare efficiency, allowing systems to better allocate resources and reduce long-term treatment costs. Ultimately, it improves patient outcomes by aligning therapeutic decisions with individual risk profiles.

1.3 Limitations of Current Approaches (Genomic and Manual Annotation)

Current methods for predicting breast cancer recurrence predominantly rely on genomic assays such as Oncotype DX and MammaPrint. These assays evaluate the expression profiles of multiple genes within tumor tissues to estimate recurrence risk and inform treatment decisions. Although validated and widely used in clinical practice, these genomic tests present several significant limitations.

One major limitation of genomic methods is their high cost, which can exceed $3,000 to $4,000 per test. Such costs pose substantial barriers, particularly in low-resource healthcare settings and developing countries, significantly restricting their accessibility and widespread use. Additionally, genomic assays require specialized laboratory facilities and trained personnel, adding complexity and increasing turnaround time for results, often delaying treatment decisions.

Manual annotation by pathologists, an alternative or adjunct approach often used in digital pathology, also faces critical constraints. Expert pathologist annotations involve careful delineation of cancerous regions and structures in histological images—a highly time-consuming and subjective process. The inherent variability between observers introduces inconsistencies and potential inaccuracies, affecting the reproducibility of predictive assessments.

Given these constraints, there is a compelling need for alternative methods capable of overcoming cost, scalability, and reliability challenges. Novel approaches leveraging computational techniques, such as machine learning and advanced image analysis, offer promising pathways toward cost-effective, rapid, and scalable recurrence risk prediction, potentially reducing reliance on expensive genomic tests and subjective manual annotations.

1.4. Potential of Digital Pathology and Machine Learning

Digital pathology refers to the practice of converting traditional glass histopathology slides into high-resolution digital images that can be viewed, analyzed, managed, and interpreted using computer-based tools. By transforming tissue slides into digital formats, pathologists can leverage computational methods to enhance diagnostic accuracy, reproducibility, and efficiency. These digital images, particularly those stained with hematoxylin and eosin (H&E), contain rich morphological and cellular information critical to disease characterization and prognosis.

Machine learning models, especially Convolutional Neural Networks (CNNs), can analyze raw pixel data to detect patterns without requiring explicit feature engineering. These networks are adept at identifying visual cues—from cell morphology to tissue organization—that might escape human detection. By doing so, they offer scalable and cost-effective solutions that can democratize access to high-quality recurrence assessment.

1.5 Objectives and Scope of the Literature Review

The primary objective of this literature review is to comprehensively summarize existing approaches in predicting breast cancer recurrence risk, with a specific emphasis on methodologies leveraging digital pathology images and machine learning. This review systematically explores current practices—including genomic assays and manual pathology assessments—as well as emerging computational techniques, particularly deep learning models.

Specifically, this paper aims to summarize and critically assess the current state-of-the-art methods in recurrence prediction, identify methodological gaps within existing research, particularly concerning cost, scalability, dependence on manual annotations, and computational feasibility, and propose potential future research directions, highlighting innovative approaches such as pixel-level analysis, unsupervised or semi-supervised learning, multi-modal data integration, and scalable machine learning frameworks.

By clearly defining these objectives, this literature review seeks to establish a robust foundation for further research aimed at improving the accuracy, accessibility, and clinical applicability of breast cancer recurrence prediction methodologies.

2. Breast Cancer Recurrence Prediction: Current Practices

2.1. Clinical and Genomic Assays (Oncotype DX, MammaPrint)

Genomic assays like Oncotype DX and MammaPrint have become widely adopted tools for estimating the risk of breast cancer recurrence, particularly in early-stage, hormone receptor-positive cases. These assays function by analyzing the expression levels of specific gene panels within tumor tissues. Oncotype DX evaluates 21 genes to generate a Recurrence Score (RS) ranging from 0 to 100, while MammaPrint analyzes a 70-gene signature to classify patients into

high-risk or low-risk categories. Both tests are grounded in the understanding that molecular activity within the tumor can reveal its potential for future aggressiveness, independent of visual appearance under a microscope.

These assays have undergone extensive clinical validation in large-scale trials such as TAILORx (for Oncotype DX) and MINDACT (for MammaPrint), which confirmed their utility in guiding decisions about chemotherapy. However, their use is not without limitations. The tests are expensive, require specialized processing facilities, and often have turnaround times that delay treatment planning. Additionally, their predictive power may be limited in certain populations, and they offer no insights into morphological patterns or tissue-level context. As a result, their implementation is often restricted to well-resourced healthcare systems, leaving gaps in global accessibility.

2.2. Role of Histopathology in Breast Cancer Prognosis

One of the most common histopathology techniques used is Hematoxylin and Eosin (H&E) staining, a process that highlights cellular structures and tissue architecture. Hematoxylin stains nuclei a deep blue or purple, while eosin counterstains the extracellular matrix and cytoplasm in varying shades of pink. These slides are then examined by pathologists who assess features such as tumor grade, mitotic activity, and invasion patterns—important indicators of how aggressive a tumor is likely to be.

Morphological characteristics derived from H&E slides carry prognostic value. For instance, high nuclear pleomorphism, increased mitotic count, and disrupted tissue architecture often correlate with poor outcomes. Tumor cellularity, lymphocytic infiltration, and the ratio of epithelial to stromal content are additional features that inform clinical judgment. Although this form of evaluation is relatively low-cost and widely used, it heavily relies on the subjective interpretation of experienced pathologists and lacks standardization across institutions. Nevertheless, the inherent richness of visual and structural information in H&E-stained images presents an untapped opportunity for computational methods to extract objective, high-resolution insights into tumor behavior.

3. Machine Learning in Digital Pathology

3.1. Overview of Machine Learning Techniques in Histopathology

Before applying machine learning models to histopathology images, several preprocessing steps are essential to ensure consistency and computational feasibility. Large whole-slide images (WSIs) are typically divided into smaller patches—a process called tiling—to reduce memory usage and localize analysis. These patches may then undergo color normalization to adjust for staining variability across samples and scanners, which is crucial for generalizing models across

datasets. Preprocessing often also includes background removal and noise reduction to isolate biologically relevant tissue from irrelevant slide regions.

Following preprocessing, segmentation techniques are employed to distinguish structures of interest—such as nuclei, glands, or stromal areas—from surrounding tissue. Common approaches include intensity thresholding, which separates foreground from background based on pixel intensity, and connected component labeling, which identifies contiguous pixel regions of similar values. Once segmented, regions can be analyzed using feature extraction methods. These may include texture descriptors (e.g., Local Binary Patterns), shape features (e.g., circularity, aspect ratio), and color metrics. These extracted features are then input into classic classifiers such as Support Vector Machines (SVM), Random Forests, or k-Nearest Neighbors (k-NN). Performance is typically assessed using validation techniques like k-fold cross-validation, with metrics such as accuracy, precision, and the Receiver Operating Characteristic (ROC) curve used to evaluate predictive quality.

3.2. Deep Learning Approaches: Focus on CNNs

While traditional machine learning methods rely on handcrafted features, Convolutional Neural Networks (CNNs) automate this process by learning hierarchical patterns directly from image data. A CNN is composed of several types of layers: convolutional layers that detect spatial features (e.g., edges, textures), pooling layers that downsample feature maps to reduce dimensionality, and fully connected layers that perform classification. This layered architecture enables CNNs to progressively build high-level representations of image content, capturing both local and global structures within tissue samples, as shown in Figure 1.

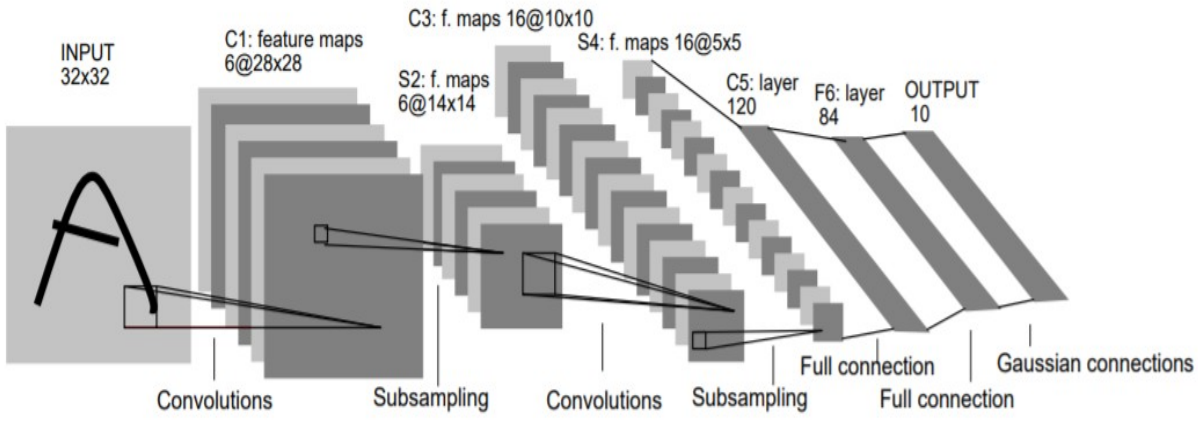

Figure 1: Illustration of a traditional Convolutional Neural Network (CNN) architecture used for image classification. The input image is processed through layers of convolutions, subsampling (pooling), and fully connected layers to extract hierarchical features. Adapted from Kumar Ajitesh(2023)

CNNs are especially powerful in histopathology due to their ability to model complex spatial relationships and subtle visual differences that might be overlooked by manual inspection or simpler models. They have been successfully applied to tasks such as tumor classification, metastasis detection, and nuclei segmentation. Popular CNN architectures used in digital pathology include VGG16, known for its simplicity and depth; ResNet, which introduces skip connections to prevent vanishing gradients; and EfficientNet, which balances accuracy and computational efficiency. These models, often pre-trained on large datasets, can be fine-tuned for medical images to accelerate training and improve performance in scenarios with limited annotated data.

4. Review of Recent Studies in Breast Cancer Recurrence Prediction from Histopathology

4.1. Patch-based and Region-of-Interest Approaches

One of the leading models in patch-based analysis is BCR-Net, proposed by Su et al. (2023), which employs a multiple instance learning framework to predict recurrence from histopathology images. The model uses a patch-sampling strategy to focus on high-density regions and aggregates predictions across patches to classify entire slides. Figure 2 illustrates the architecture of BCR-Net. This model achieved AUCs of 0.775 and 0.811 on H&E- and Ki67-stained slides, respectively, demonstrating solid performance. However, it relies on region-level annotations provided by expert pathologists, which limits scalability and introduces human bias.

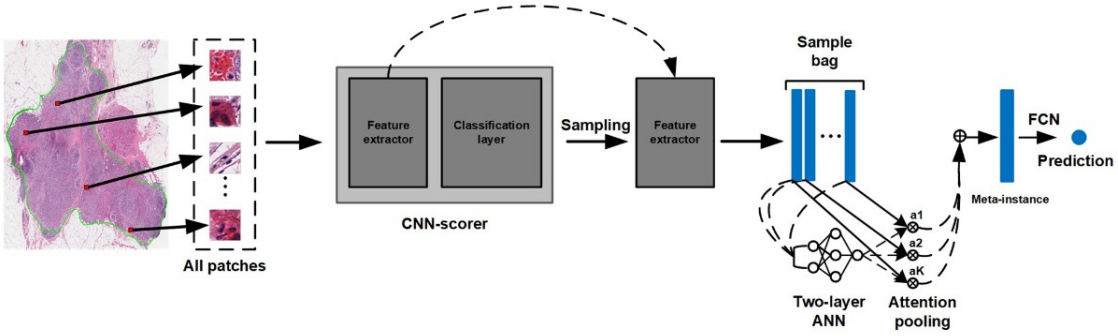

Figure 2: Architecture of the BCR-Net model for breast cancer recurrence prediction. The framework extracts patches from annotated tumor regions in whole-slide images and ranks them using a CNN-based scoring module. Top-ranked patches are sampled and reprocessed through a shared feature extractor.

Shi et al. (2023) explored a hybrid model combining VGG16—a deep CNN used as a feature extractor—with a Support Vector Machine (SVM) classifier. Their method focused on predicting early recurrence using features derived from annotated tumor regions. Although they achieved modest accuracy (62.4%), the approach reinforced the viability of combining traditional machine

learning with deep feature extraction. In a different line of work, Whitney et al. (2018) implemented nuclear histomorphometry by segmenting epithelial and stromal regions and using patch-based classifiers to predict Oncotype DX categories. Their model achieved 76–85% accuracy and demonstrated that interpretable, structure-aware features could offer strong predictive power when paired with appropriate classifiers.

4.2. Annotation-Free Models and Whole-Slide Image Analysis

Phan et al. (2021) introduced a deep learning pipeline based on the Xception CNN, trained on 512×512 pixel patches without the need for manual annotations. Their model achieved an impressive 87% accuracy by relying on large-scale patch training and optimized architecture, showcasing that annotation-free deep learning models can perform competitively if the dataset is well-curated and sufficiently diverse.

Rose (2024) proposed another annotation-free approach using ResNet-18 for feature extraction, followed by logistic regression for classification. The model was trained on 2,000 patches per slide and reached an average cross-validation accuracy of 62.8%. While not as performant as other models, it demonstrated that whole-slide learning without labeled Region of Interest(ROI) remains a promising direction, especially for low-resource environments.

These studies highlight the growing trend toward automation and scalability by removing the need for labor-intensive manual annotations while still preserving reasonable predictive performance.

4.3. Challenges and Limitations Identified in Literature

Despite promising results, several challenges persist in current methodologies. A common limitation across studies is the use of small or institution-specific datasets, which often lack diversity and reduce model generalizability. This limitation is further compounded by a reliance on manually annotated regions, which restricts scalability and introduces variability due to differing annotation standards and pathologist interpretation.

Additionally, many models are adapted from architectures originally trained on natural images (e.g., ImageNet), which poses issues when applied to medical data. These pre-trained models often struggle to capture the fine-grained, domain-specific features present in histopathology slides, leading to performance degradation without significant retraining or fine-tuning. Bridging the gap between natural image models and medical image complexity remains a major technical and methodological hurdle in advancing recurrence prediction tools.

5. Emerging Trends and Potential Directions

5.1. Pixel-level Analysis: Opportunities and Challenges

Pixel-level analysis approaches histopathological images at the most granular level by treating each pixel as a potential carrier of diagnostic information. Unlike patch-based methods that treat tiles as independent units, pixel-level strategies aim to preserve detailed spatial relationships and capture subtle textural or morphological variations that may be diluted in aggregated patches. This fine-grained resolution could enable the discovery of novel histological patterns linked to recurrence risk, offering greater sensitivity to intratumoral heterogeneity and microstructural features.

Although this level of detail introduces significant computational challenges—as processing millions of pixels per slide requires substantial memory and computing power, often necessitating high-performance infrastructure or cloud-based solutions—it remains a promising direction. The potential to uncover subtle, fine-grained histological patterns that may be overlooked in patch-based models makes pixel-level analysis a valuable method to pursue. With ongoing advancements in computational efficiency and model optimization, this approach may become increasingly feasible for real-world diagnostic applications. As shown in Figure 3, pixel-level analysis workflows involve detailed scanning and recognition of minute features across histopathology images that might otherwise be overlooked.

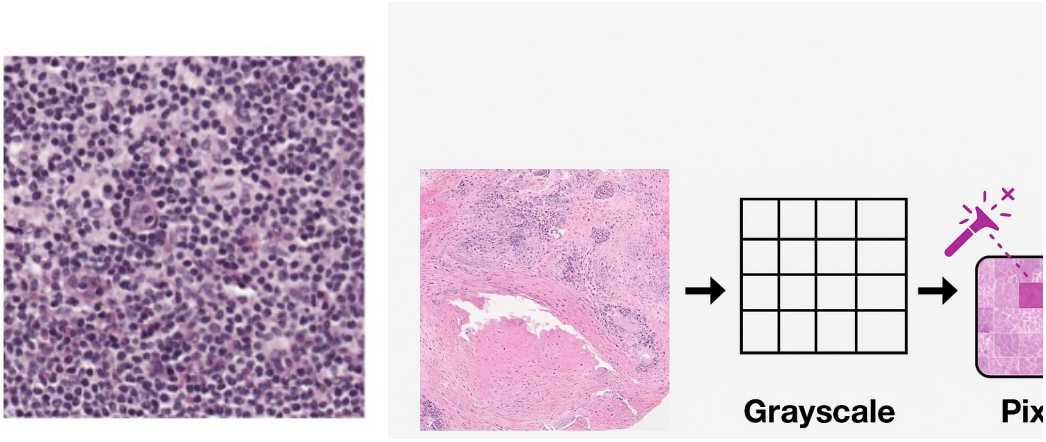

Figure 3 (a) Representative H&E-stained patch from the CAMELYON17 dataset (Litjens et al., 2017), illustrating the cellular detail captured in histopathology slides. (b) Conceptual illustration of a pixel-level analysis pipeline: whole-slide image is tiled, converted to grayscale, and scanned at the pixel level for fine-grained pattern recognition. Generated using ChatGPT-4 for explanatory purposes.

5.2. Integration of Multi-modal Data (Genomics + Imaging)

Integrating multi-modal data—such as combining genomic profiles with histopathology images—is a growing and promising direction in recurrence prediction. This approach leverages the spatial and morphological detail captured in digital pathology alongside the molecular insights provided by genomic data, enabling a more holistic understanding of tumor biology. When used together, these modalities can enhance predictive power and support biologically informed, personalized prognostic assessments.

While standalone genomic assays remain costly, multimodal approaches that integrate genomic features with histopathology images aim to reduce costs by extracting key molecular information computationally, without requiring full clinical assays for every patient.

Recent frameworks, such as those discussed by Boehm et al. (2022), emphasize the value of multimodal data integration in precision oncology, highlighting how clinical, pathological, radiological, and genomic data can be combined to improve patient stratification and treatment planning. In the context of breast cancer, models that jointly analyze H&E-stained slides and transcriptomic data have shown improved accuracy compared to unimodal approaches. As tools and datasets for multi-modal learning continue to evolve, this fusion of visual and molecular information represents a transformative advancement in developing more nuanced, scalable, and clinically actionable recurrence prediction systems. Figure 4 provides a conceptual overview of multimodal data integration and illustrates how distinct data streams are unified into predictive models.

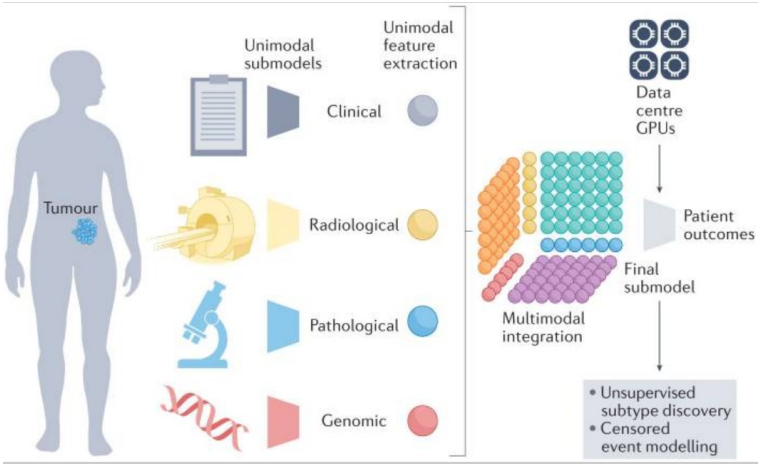

Figure 4: Conceptual framework for multimodal data integration in oncology, illustrating how clinical, radiological, pathological (including histopathology), and genomic data can be processed through unimodal submodels and fused into a unified predictive model for patient outcomes. Adapted from Boehm et al. (2022), Nature Reviews Cancer, 22(2), 114–126.

5.3. Computational Efficiency and Scalability

To be adopted widely in clinical settings, recurrence prediction tools must be computationally efficient and scalable. Many deep learning models, particularly those used in pixel- or patch-level analysis, demand significant resources for training and inference. Without optimization, these tools may be too slow or costly to integrate into real-time diagnostic pipelines.

Advancements in cloud computing, GPU acceleration, and distributed processing have made it increasingly feasible to deploy large-scale models efficiently. Platforms like Google Cloud AI and Amazon SageMaker allow for scalable model training, while containerization tools like Docker facilitate deployment across systems. By leveraging such technologies, developers can reduce latency, support batch processing of large pathology datasets, and create solutions that are both powerful and clinically practical.

6. Conclusion

This paper explores the evolving landscape of breast cancer recurrence prediction through the lens of digital pathology and machine learning. As traditional methods face limitations in cost, accessibility, and scalability, emerging computational techniques offer promising alternatives. From patch-based models to annotation-free strategies and pixel-level analysis, the field is moving toward more automated, interpretable, and inclusive diagnostic frameworks.

Key trends signal a shift toward more adaptable and clinically viable solutions. While challenges remain in generalizability, computational intensity, and real-world deployment, the convergence of histological imaging and artificial intelligence is paving the way for more precise, affordable, and accessible recurrence prediction tools. Continued research at the intersection of pathology, data science, and clinical practice will be critical to translating these advancements into meaningful improvements in cancer care globally.

References

Ajitesh, K. (2023, December 4). Different types of CNN architectures explained with examples. VitalFlux. https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/. Accessed: 2025-04-22.

American Cancer Society. (2024, July 11). Breast cancer facts & figures 2024 (PDF). https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2024/breast-cancer-facts-and-figures-2024.pdf. Accessed: 2025-04-29.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., & Campilho, A. (2017). Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS ONE*, 12(6), e0177544.

Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J., & Shah, S.(2022). Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer 22*, 114–126.

Camelyon17 - grand challenge. https://camelyon17.grand-challenge.org/Home/. Accessed: 2025-04-22.

Ciga, O., Xu, T., & Martel, A. L. (2022). Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications, 7,* 100198.

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. *ArXiv.*

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88.

Phan, T., Jin, Y., Lee, B., & Nguyen, H. (2021). Predicting breast cancer recurrence risk using Xception and deep CNNs on H&E-stained WSIs. *Computerized Medical Imaging and Graphics*, 89, 101878.

Rose, M. (2024). Patch-Based Analysis with Machine Learning to Aid Breast Cancer Recurrence Prediction (Master's thesis, East Carolina University).

Shi, Y., Wang, Z., & Xu, Y. (2023). Early recurrence prediction in breast cancer using VGG16-based features and support vector machines. *Medical Image Analysis*, 82, 102635.

Su, H., Wei, L., He, Y., Qian, Y., & Liu, Q. (2023). BCR-Net: A Deep Learning Framework to Predict Breast Cancer Recurrence from Histopathology Images. *Journal of Digital Imaging*, 36, 123–134.

Whitney, H. M., Giger, M. L., Li, H., & Zhou, Y. (2018). Predicting Oncotype DX recurrence scores using nuclear features from H&E images. *Medical Physics*, *45*(3), 1221–1230.

World Health Organization. (2024, March 13). Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed 2025-04-29.