PATCH BASED ANALYSIS WITH MACHINE LEARNING TO AID BREAST CANCER RECURRENCE PREDICTION

by

Madison Rose

May, 2024

Director of Thesis: Nic Herndon, PhD

Major Department: Computer Science

Since the introduction of whole slide scanners, machine learning research has become a popular area of interest in digital pathology. Many studies have attempted to use machine learning to aid pathology tasks such as breast cancer diagnosis and metastasis detection. However, one area that has less available research is in applying machine learning to predict patient recurrence risk categories. Since H&E-stained images are routinely collected for diagnostic purposes, creating an image-based recurrence prediction method could help increase accessibility and lower cost for recurrence risk category assessment for breast cancer patients. In this study, patches were extracted from a dataset of 102 whole slide images to train a machine learning model to predict slide level breast cancer Oncotype DX risk category using only H&E-stained images with no additional clinical data or region of interest annotations. Multiple patch size and patch quantity combinations were tested. Patches were extracted from each whole slide image and feature extraction was performed before the features were aggregated together to create a bag of features for each case. These bags were then used to train a logistic regression model. The best scoring model utilized 2,000 patches of size 256 x 256 pixels. This model scored 0.628 ± 0.044 accuracy on 5-fold cross validation across the entire dataset.

PATCH BASED ANALYSIS WITH MACHINE LEARNING TO AID BREAST CANCER RECURRENCE PREDICTION

A Thesis

Presented to The Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Data Science

by

Madison Rose

May, 2024

Director of Thesis: Nic Herndon, PhD

Thesis Committee Members:

Rui Wu, PhD

David Hart, PhD

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Nic Herndon for his incredible support, guidance, and mentorship as my thesis advisor. I also want to thank Dr. Joseph Geradts for providing the idea and data for this study along with answering many pathology related questions. I'd like to thank Dr. David Hart and Dr. Rui Wu for serving on my thesis committee and providing their knowledge and advice. I'd like to thank Jared Ratz with the ECU Brody School of Medicine Department of Pathology for his technical assistance in accessing the data needed for this work. I'd also like to thank Jason Gray with ECU College of Engineering and Technology for his technical support of the computing resources utilized in this work. Lastly, I want to thank my family and friends for their constant support and encouragement throughout this process.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Breast cancer is one of the most common cancers among women worldwide. Although incidence rates have increased in recent years, survival rates are also up [48]. One factor that could be contributing to better survival rates is the use of more personalized treatment plans. One way treatment plans can be personalized is with chemotherapy. Oncologists can decide if a patient would benefit from chemotherapy based on their recurrence score, which is an indicator of how aggressive the cancer is and how likely it is to recur. Most often, the Oncotype DX (ODX) Recurrence Test is used to evaluate recurrence risk. The Oncotype DX Recurrence Test is a 21 gene assay which outputs an integer between 0 and 100, also known as a recurrence score [50]. Generally, scores of 26 or higher are considered to be high-risk and scores below 26 are considered to be low risk, although some studies will further split the low risk group into low and intermediate risk groups. These scores come from the TAILORx study [61]. The ODX test is performed on tissue that is extracted during a breast biopsy. The ODX recurrence score has been shown to be strongly correlated with risk of breast cancer recurrence within 10 years. However, these tests are costly and in high demand which limits patient access [9].

Since whole slide scanners capable of digitizing stained tissue slides from breast biopsies have become more common, there is a strong research interest to use machine learning to aid in digital pathology tasks. Image-based machine learning has been applied to tasks

such as breast cancer diagnosis, tissue segmentation, and metastasis detection with very accurate results [25,35]. However, there is not as much available research focused on using image-based machine learning methods to predict patient recurrence risk.

Some studies have focused on predicting recurrence with non image data including clinical data and electronic health records and have produced good results [4]. There are fewer studies attempting to predict recurrence with images. However, if recurrence could be predicted from images, such as hematoxylin and eosin-stained whole slide images, which are routinely collected for diagnosis, then less laboratory work would be needed [47]. This could also help keep costs lower since no additional information would need to be collected from the patient. Many works that focus on image-based prediction of breast cancer recurrence from H&E-stained images utilize pathologist tumor region of interest annotations which are extremely helpful in determining the area within the whole slide image that is most relevant. Additionally, some studies combine H&E-stained images with clinical data or also investigate recurrence prediction using whole slide images stained with other materials [51].

One study focused on identifying patients who were considered high-risk for early breast cancer recurrence. Early recurrence was defined to be the return of a primary tumor within three years of diagnosis. Instead of using recurrence scores, patients were followed for confirmed recurrence. The dataset used in this case contained 704 images from 202 patients. The dataset was balanced among the classes with 101 of the patients having recurrence within 3 years while the other 101 were non-recurrent. In this paper, VGG16 was used as a feature extractor and was combined with support vector machines to make final predictions. The cross-validation accuracy for predicting recurrence was 62.4%. The model performed better at predicting early-recurrence for low to intermediate grade tumors [47].

Another study looked to predict ODX risk from whole slide histopathology images annotated with tumor region of interests by pathologists. In this work, a novel sampling

method for patches was introduced to select the most discriminative patches from the image tumor regions for use. The framework was analyzed for both H&E-stained slides and Ki67 stained slides. Overall, the Ki67 slides had better results, with a 0.800 accuracy when evaluated with leave two out cross validation. The H&E-stained slides set achieved 0.700 $\pm$ 0.030 accuracy across 5-fold cross validation [51].

Another study looked to predict recurrence risk categories using H&E-stained slides with pathologist annotated regions of interest (ROIs). Once patches were extracted from the ROIs, nuclei detection was performed and then was combined with a model to perform epithelial/stromal separation. Next, feature extraction was performed and a classification model was used to predict patch level recurrence risk. Finally, a patch-based voting method was implemented to predict the patient level risk category. This proposed model scored between 76% and 85% for all tests performed. The authors of this work proposed that these results indicated that it could be possible for recurrence risk to be predicted from H&E-stained images [57].

A related work looked to predict recurrence risk based on a 70 gene signature risk score. Multiple convolutional neural networks were evaluated on the training data and Xception scored the highest, so it was selected for testing. Using Xception, an accuracy of 87% was achieved without region of interest labeling. A patch size of 512 x 512 pixels was originally sampled from the whole slide images but was later resampled to fit the model architecture [41].

One common approach when using whole slide images for machine learning is to perform some type of patching of the whole slide image. Due to the large size of whole slide images, they are not suitable for use by machine learning models as is. Therefore, a collection of smaller patches is typically retrieved from each image, and these are used instead since they are smaller and easier to work with. Often studies vary on patching techniques. A wide variety of patches may be used by various studies, including overlapping vs non

overlapping, varying image sizes, varying number of images, and their selection methods. In this work, two different patch sizes are compared as well as multiple quantities of extracted patches per image.

However, using patches instead of the whole slide image presents new challenges. First, there is the possibility that an individual patch's label will not match the slide level label. Secondly, even once patch level labels are obtained, aggregation is needed to generate a level for the entire slide. This is most notable in classification problems as these generally look for a slide or patient level label as opposed to segmentation tasks that predict at the patch/pixel level.

One common approach to address this is by using convolutional neural networks to extract features from image patches first [26]. This can be followed by a number of techniques, but recently multiple instance learning has started to emerge as a popular choice. In multiple instance learning, patch features would be grouped together according to the slide they came from. These features would become instances and their slide group would be considered a bag. In multiple instance learning, instances (and their combination) inside a bag are used to predict the entire bag's label [18].

This study investigates machine learning model performance on predicting breast cancer recurrence based on the associated Oncotype DX recurrence score, given only H&E-stained slides, with no additional clinical data or region of interest annotations. Patch size and quantity selection are also investigated as to their impact on overall model accuracy. The patch sizes 512 x 512 pixels and 256 x 256 pixels were selected for analysis as these are commonly used image sizes [25, 34]. Understanding the impact of patch size and patch quantity on overall model accuracy could assist more researchers when selecting patches for their work. If it were possible to predict the recurrence score with only the H&E-stained image, tumor region of interest annotations would not be needed and this would reduce the time needed for image annotation by pathologists. Ultimately, machine learning could pro-

vide a tool to assist with recurrence risk assessment which could hopefully increase access to more personalized treatment options for breast cancer patients.

Chapter 2

Deep Learning in Digital Breast Pathology

This chapter contains the paper Deep Learning in Digital Breast Pathology from authors Madison Rose, Joseph Geradts, and Nic Herndon. This paper was published through SCITEPRESS and was presented as part of BIOINFORMATICS 2024. The event website can be accessed at https://bioinformatics.scitevents.org/. The paper can be accessed online at the following DOI:10.5220/001257610000365.

2.1 ABSTRACT

The development of scanners capable of whole slide imaging has transformed digital pathology. There have been many benefits to being able to digitize a stained-glass slide from a tissue sample, but perhaps the most impactful one has been the introduction of machine learning in digital pathology. This has the potential to revolutionize the field through increased diagnostic accuracy as well as reduced workload on pathologists. In the last few years, a wide range of machine learning techniques have been applied to various tasks in digital pathology, with deep learning and convolutional neural networks being arguably the most popular choice. Breast cancer, as one of the most common cancers among women worldwide, has been a topic of wide interest since hematoxylin and eosin-stained (H&E)-stained slides can be used for breast cancer diagnosis. This paper summarizes key advancements in digital breast pathology with a focus on whole slide image analysis and provides

insight into popular methods to overcome key challenges in the industry.

2.2 INTRODUCTION

Advancements in whole slide imaging (WSI) have paved the way for digital pathology. This has driven the increasing demand for more research into using machine learning for whole slide image analysis. This paper provides an overview of the main aspects of deep learning in digital breast pathology. Background information is included that can be used to gain an understanding of the field. Key advancements, tools, and insight into popular methods for overcoming key challenges are discussed. While digital pathology is a large field, the focus here will be on analysis of whole slide images through deep learning techniques.

2.2.1 Whole Slide Imaging

Whole slide imaging shows many potential benefits compared to its glass slide counterpart. Digitized slides allow remote users to view slides for secondary or even primary diagnosis. Digitization is also useful for archiving and preserving samples, which is important since physical samples degrade over time. An additional benefit of better archiving of tissue samples is the preservation of rare specimens. Since digitized slides can be accessed remotely, WSI also provides the opportunity to make advancements in standardizing training for pathologists. It is important to consider that while WSI can also be used for diagnosis, there are still factors that can affect diagnostic accuracy from digitized images. While whole slide images are approved for diagnosis, some discrepancies still make glass slide viewing the standard for diagnosis. These issues stem from poor image quality and bad focus. Some specific microscopic details such as mitotic figures that may be needed for analysis can also be difficult to identify on the digitized images, in some cases, due to faint scanning. However, it is important to note that even glass to glass slide studies can show

discrepancies due to observer variability, among other factors [40].

2.2.2 Digital Pathology Tasks

In histopathology image analysis, three main tasks for machine learning have emerged: classification, segmentation, and object detection. Classification involves analyzing an image and giving the image a label, sorting it into a class. There are two types of classification, binary and multi-class classification [20]. In binary classification, there are only two possible labels or classes for an image. In contrast, multi-class labeling has three or more possible labels for a given image. A study by Araujo et al. (2017) displayed both binary and multi-class classification. They used a convolutional neural network for multi-class classification of breast biopsy images into one of four categories: normal, benign, in situ carcinoma, and invasive carcinoma. They also performed binary classification into carcinoma and non-carcinoma. More specific labelling as done in multi-class classification is often very useful in medical diagnosis. However, due to using an increased number of classes, multi-class models often require more complexity than binary models, which can impact their accuracy. For example, in the study mentioned above, the four-class model scored 65% accuracy on test data compared to the binary class model achieving 77%.

Segmentation aims to separate parts of the images - often cancerous cells vs. non-cancerous cells. Object detection focuses on finding landmarks in an image, like individual cells or nuclei. In this paper, segmentation and object detection will briefly be discussed, while classification tasks will be the main focus.

2.3 BACKGROUND

2.3.1 Breast Cancer

In 2023, breast cancer accounted for 31% of all female cancers, making it one of the most common cancers among women. Breast cancer occurrence rates have been steadily increasing since the 2000s by about 0.5% per year. Improvements in treatment have seen the mortality rate for breast cancer decrease despite the increase in incidence [48]. It is well documented that early diagnosis and intervention can greatly improve survival rates in breast cancer patients. Smaller tumors have notably better long-term survival rates than larger tumors [8]. Many techniques are used to screen for and diagnose breast cancer, including mammography and ultrasonography [56]. However, while these are helpful in screening and early detection of breast cancer, a breast biopsy is the only definitive method for diagnosing breast cancer [37]. Tissue samples can provide information about tumor type, grade, and biomarker status. A triple assessment is often used to evaluate patients, consisting of clinical evaluation and imaging in addition to a tissue biopsy [2].

Once biopsied tissue is collected, it is fixed, processed, sectioned, and stained to color different parts of the cells in the tissue. Hematoxylin and eosin (H&E) staining is considered the gold standard in breast tissue biopsies and has been around for over 100 years [24]. When H&E staining is performed, different cell parts will look distinct based on which type of dye they have an affinity for. Hematoxylin is a basic dye whereas eosin is an acidic dye. Cell structures such as nuclei that have an affinity for hematoxylin appear blue after staining. Structures such as cytoplasm that have an affinity for eosin appear pink after staining. Structures with an affinity for both basic and acidic dyes will appear purple after staining [6, 10].

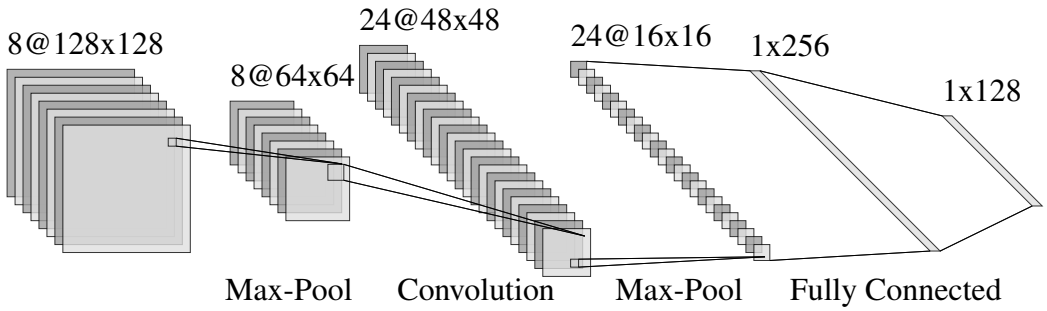


Figure 2.1: An example of a convolutional neural network architecture with convolutional layers, pooling layers, and fully connected layers. As the input moves through the CNN, it continues to be reduced in size. Figure generated with NN-SVG [33].

2.3.2 Whole Slide Image Resolutions

Whole slide image scanners operate by capturing images of tissue sections tile by tile. The whole image is reconstructed at the end. These scans can be performed at multiple magnifications which increases image detail. 20x magnification is a common scan objective and is adequate for typical viewing. However, some types of slides require more detail and need higher levels of magnification such as 40x [59].

2.3.3 Machine Learning Types

There are many learning types in machine learning. The two main types are supervised and unsupervised learning. These types differ based on the types of data they receive.

Supervised learning supplies a machine learning model with input data and its expected output. What this will look like can vary depending on the task being performed. In cases of classification, the output is typically a label. In segmentation tasks, often a mask of pixels is used as the ground truth label [25]. In object detection, bounding boxes in certain parts of the image are typically provided [34]. The model will then try to predict the desired output given only the input. The revolutionary idea behind deep learning models is that they can compare their original predictions with the expected output and internally modify their configurations to see more accurate predictions during the next round of training or

testing. This is done through a method called backpropagation [31]. Convolutional neural networks are a popular type of supervised machine learning model.

In contrast to supervised learning, unsupervised learning uses unlabeled data. Unsupervised learning models will group data based on patterns and similarities but are unable to provide a label [20]. This can be useful in histopathological image analysis for detecting patterns in images that may not be currently recognized by pathologists. Additionally, unsupervised learning can be used for feature extraction on histopathologic images [46].

2.3.4 Deep Learning and CNNs

Deep learning is a subfield of machine learning that focuses on using nodes to form a neural network [19]. These neural networks were originally inspired by the human brain. The nodes in neural networks are also referred to as neurons since they mimic how neurons function in a human brain [38]. Deep learning has become increasingly popular for several reasons. First, these models are successful at a wide variety of tasks such as natural language processing, speech and audio processing, and digital image processing. The way deep learning algorithms extract features decreases the amount of domain knowledge and work needed by researchers [43]. Convolutional neural networks (CNNs) are a type of neural network and are particularly good at image recognition tasks. CNNs take in image pixel values as input and pass these values through a series of layers while performing various operations on the images. Convolutional neural network architecture consists of three main types of layers: convolutional, pooling, and fully connected layers, which can be observed in Figure 2.1. In convolutional layers, two dimensional filters are applied to the image data to extract features such as edges, objects, and colors. An example of a convolution can be seen in Figure 2.2. These features are used to create a feature map. Convolutional layers are often paired with activation functions, such as ReLU (rectified linear units), which improve speed and performance by removing negative values after a convolution has been

performed [28, 60]. The pooling layer does downsampling which reduces the number of parameters used while trying to maintain the features [38, 60]. Lastly, fully connected layers take inputs received by previous layers and connect them to activation units to produce output [60].

2.3.5 CNN History

In 1989, Yann Lecun introduced LeNet-5, a convolutional neural network originally designed to recognize handwriting digits [30]. It became one of the first widely recognized published convolutional neural networks due to its performance with an error rate of 0.95% on the test set of 16x16 pixel handwriting digit images [30, 60]. LeNet-5 was also revolutionary for its use of backpropagation to reconfigure its own internal weights using gradient descent [30]. In 2012, convolutional neural networks surged in popularity after AlexNet won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [28]. Since then, the development of new CNN architectures rapidly expanded with the emergence of VGG16/VGG19 [49], Resnet [21] and Inception [52]. ImageNet is a popular dataset for training and benchmarking convolutional neural networks and contains millions of annotated natural images [14]. Classification and object detection are common computer vision tasks and are included in popular challenges such as ILSVRC [45]. Convolutional neural networks are particularly skilled at computer vision tasks and in turn have been applied to a wide range of medical imaging tasks.

2.3.6 Modern CNNs

One of the most popular modern CNNs is VGG16. The Visual Geometry Group (VGG) submitted VGG16 to the ILSVRC in 2014. It proposed an increase in CNN depth and was pretrained on the ImageNet dataset. This model varied from its predecessors by using stacks of smaller 3x3 receptive fields instead of larger 11x11 or 7x7 receptive fields. This

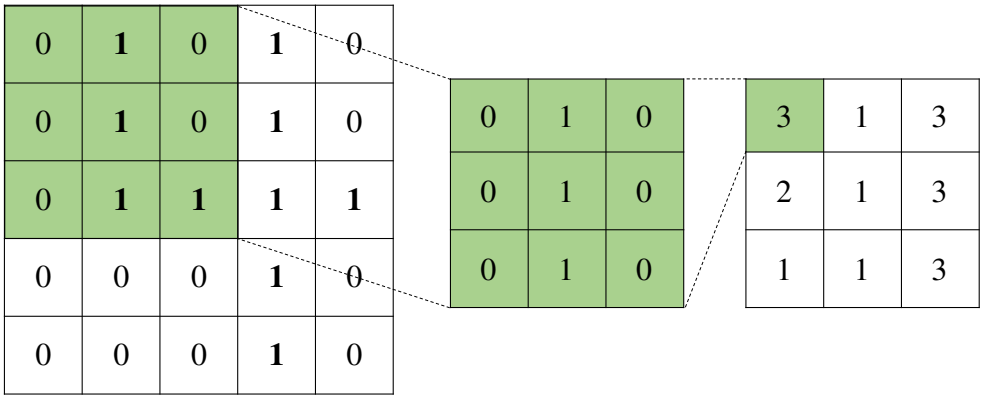


Figure 2.2: An example of the convolution operation. This convolution uses a 3x3 filter (center grid) on an image of size 5x5 (left grid) with a stride of 1 and padding of 0. The result of the convolution is the rightmost grid. The filter used here is an example of a filter that detects vertical lines.

decreased the number of parameters throughout the network. Small convolutions had previously been tried but no other CNNs that used these smaller filters were as deep as VGG16, which boasted sixteen layers as its name suggests. The VGG were able to determine that a larger depth increased the classification accuracy. VGG19 was described in the same paper as VGG16 and follows a similar architecture but with nineteen layers as opposed to sixteen [49].

Another popular modern CNN is Inception (also known as GoogleLeNet). This neural network competed in the ILSVRC 2014 challenge and acheived high performance. Inception implements wider layers as opposed to making the entire network deeper with more layers [52]. This network depth is why Inception is often referred to as a deep convolutional neural network, while other CNNs like VGG16 are considered shallow. This greatly reduces the computational cost of the network in comparison to other modern CNNs such as VGG16 [53]. Like VGG16, Inception uses many smaller filters in place of larger filters

to reduce the number of parameters needed. Inception was also trained on the ImageNet dataset and achieved remarkable error rates while cutting computational costs [52].

Additional modern CNNs that will be discussed in later sections include ResNet and MobileNet. ResNet implements residual functions to reference layer input, which was shown to make optimization easier. Additionally, this allowed for increased network depth with lower complexity and increased accuracy [21]. MobileNet was built to be a lightweight deep neural network by utilizing depthwise separable convolutions [23].

2.3.7 Whole Slide Image Annotations

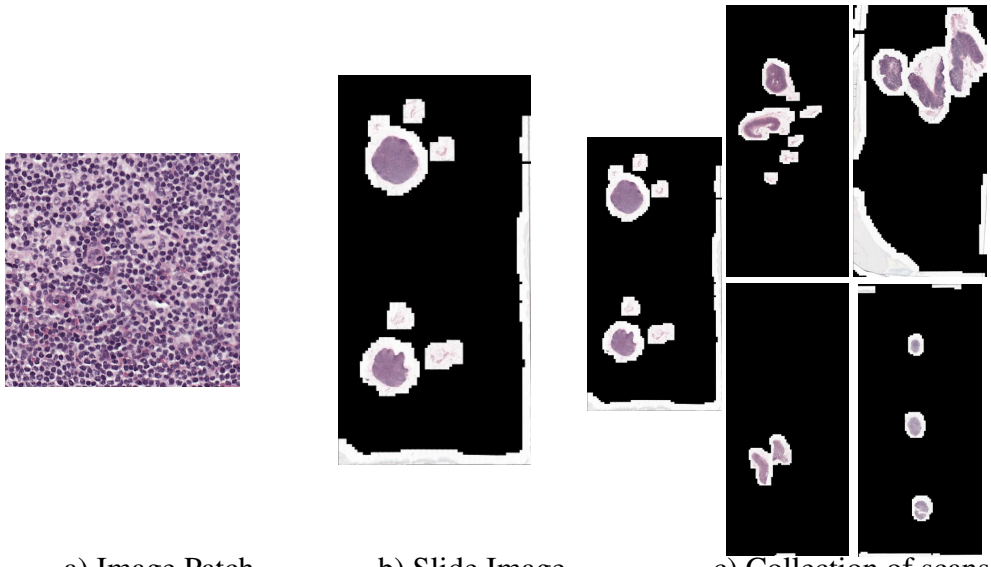
In whole slide imaging, there are three main annotation types. These types are patch level (sometimes called pixel level), slide level, and patient level. These three levels are illustrated in Figure 2.3. Each type of annotation can be useful for different tasks. These annotations can also be organized into a hierarchy of specificity.

Patch level annotation is the most specific level of annotation for whole slide images. Patch level annotations guarantee that when taking patches from WSI, every patch is fully annotated. Examples of patch level annotations include instances where each patch has its own classification label, segmentation mask, or bounding boxes [11]. One example of a segmentation mask would be when a pathologist identifies cancerous regions within a tissue sample and annotates all regions or pixels containing cancerous cells [25]. Patch level annotations are extremely helpful for segmentation tasks as they provide the ability for high supervision. However, these annotations are much more time consuming than slide level annotations and therefore are less commonly available. Often, training will be performed at a lower annotation level such as patch level while expecting a final output at a higher level such as slide level or patient level [15]. Aggregation from lower to higher annotation levels is discussed in Section 4.4.

The next level of annotation is at slide level. Slide level annotations provide one label

per whole slide. For example, a whole slide image with a slide level annotation may be labeled carcinoma vs non-carcinoma [5]. Slide level annotations are much less time consuming to do than pixel level annotations and are therefore more abundant. If looking at individual patches, as is common in WSI analysis, it is possible for a patch to not match its slide level annotation [15, 22]. This is why aggregation is needed when moving from one level during training to another at prediction time.

The least specific level of annotations is patient level. Patient level annotation is similar to slide level annotations in that a single label/class is provided, however, this label is provided to a patient rather than a specific slide. Patients may have multiple images. For example, in the CAMELYON17 dataset, each patient had 5 images [35]. Patient level annotations mean that individual slides will not be provided a label, but rather the patient, which means the label applies to all slides associated with the patient. This is the least specific level of annotation because it is possible to incorrectly label a whole slide [15].



a) Image Patch b) Slide Image c) Collection of scans for a patient

Figure 2.3: Three main annotation types for whole slide images. (a) In patch level annotation, each image patch would have its own classification label or would have a pixel annotation boundary. (b) In slide level annotation, there would be a single classification label for the entire image. (c) In patient level annotation, there would be a single label associated with all five images. Whole slide images come from the CAMELYON17 dataset [35].

2.4 COMMON CHALLENGES

There are many unique challenges in pathology image analysis when trying to apply deep learning. Solutions to these challenges are the basis of many works in the field.

2.4.1 Image Size

One major issue that must be addressed when trying to apply any type of machine learning technique to histopathology image analysis is the size of whole slide images. Whole slide image scans are extremely large, typically 100,000 x 100,000 pixels each [15]. With images this large, they are not feasible for machine learning use without modifications. For example, CNNs usually perform best with smaller images around 224 x 224 pixels in size [12]. Image compression would certainly be helpful but also has drawbacks including reduced image quality and distortion of important markers. It has been shown that there is a significant performance decrease in benign vs. malignant breast tissue classification once compression levels increase past 32:1 [29]. Even if extreme downsampling were performed, the image would remain too large for use in a convolutional neural network [12]. A common approach to address this issue is to split the image into smaller images that would be more suitable for use by machine learning models [22]. These methods are discussed in Section 4.3.

2.4.2 Data availability

A lack of well-annotated and publicly available training data is a well-known problem in digital histopathology image analysis. Even when images are available, domain knowledge is required to annotate these images to make them suitable for analysis via machine learning methods. Researchers have a few options: use a publicly available dataset, or create their own, using images provided to them by an institution or pathologist. One of

the most popular datasets used in breast histopathology image analysis is the CAMELYON dataset, which is a publicly available dataset of whole slide images along with their associated pathologist annotations. This dataset was collected from Dutch hospitals and contains 1,399 unique whole slide images totaling 2.95 terabytes. Slides were scanned with three different scanners based on which hospital they came from, with the majority of hospitals using the 3DHistech Pannoramic Flash II 250 while the Hamamatsu NanoZoomer-XR C12000-01 scanner and Philips Ultrafast Scanner were both used by one hospital each. All 1,399 WSI were annotated with a slide level label. Additionally, 399 slides from CAMELYON16 and 50 slides from CAMELYON17 were also annotated at the patch level [35].

As of the 2018 publication about the dataset, it had already been accessed by over 1000 users. Along with the dataset came the CAMELYON16 and CAMELYON17 challenges, which encouraged teams to design models to classify breast cancer metastases [35]. Although the main goal of the CAMELYON challenges is breast cancer metastases detection, the dataset is widely used by researchers interested in a variety of breast histopathology image tasks. As of December 2023, the CAMELYON17 challenge website boasts 205 submissions to the leaderboard with 1,943 total participants. The current top 10 submissions on the leaderboard all boast Cohen-Kappa scores of greater than .90 when evaluated by the CAMELYON team.

2.5 COMMON APPROACHES

2.5.1 Transfer Learning/Pretrained Models

One downside to deep learning is the computational complexity and the amount of well annotated data needed. One technique that helps reduce model complexity as well as the amount of domain specific annotated data needed is transfer learning [55]. Transfer learning is another brain inspired technique. It comes from the idea that knowledge in one task

can aid in performing a different, but somewhat related task. The use of transfer learning in convolutional neural networks is often used to pretrain a CNN with large amounts of publicly available and well annotated data, such as the ImageNet dataset. Later, the CNN can be finetuned with domain specific data [26]. Ultimately, this reduces the amount of domain specific data needed since the original weights will already be pretrained. Training time is also reduced when using pretraining methods since some portion of the training is already complete [19]. This is particularly useful in fields such as digital pathology where there may be a lack of widely available annotated data.

In a study comparing transfer learning methods in medical imaging, Kim et al. (2022) defined four types of transfer learning based on how the training is handled after the original pretraining. The feature extractor method freezes the convolutional layers and only retrains model weights in the fully connected layers. The feature extractor hybrid also freezes the convolutional layers but replaces the fully connected layers with another machine learning model, such as a support vector machine (SVM). The fine-tuning method unfreezes a few of the convolutional layers to be retrained. Finally, fine tuning from scratch completely retrains the model on the new data. After analysis of 121 publications focused on using transfer learning on convolutional neural networks with medical images, they recommended the feature extractor approach and then incrementally fine tuning the layers. Fine tuning from scratch appeared to be a prevalent method but did not show significant improvements in model accuracy despite being much more computationally expensive than other transfer learning methods [26].

2.5.2 Common Models

There are many pretrained convolutional neural networks available for use. Some models such as Inception have become commonly used because of their good performance. One review of medical imaging using CNNs found that most works use multiple models.

However, Inception was the leading model when only one model was used [26].

While tumor detection is a common task in breast digital pathology, it is not the only task that interests researchers. One study attempted to predict early recurrence from histopathological images. Early recurrence was defined as the return of a primary tumor within three years of the original diagnosis. VGG16 pretrained on ImageNet was used in conjunction with support vector machines (SVM). This approach observed a 70.3% accuracy (67.7% sensitivity) using within-patient validation [47]. Another study focused on predicting breast cancer recurrence from whole slide images used six pretrained models were used including VGG16, ResNet50, ResNet101, Inception_ResNet, EfficientB5 and Xception. Two fully connected layers were added to help reduce the computational load. Here, Xception was found to have the highest accuracy on the training data (91%) and was used for further testing where it achieved an accuracy of 87% [41].

2.5.3 Image Patches

Due to the enormous size of whole slide images, one common solution is to use patches of a whole slide image rather than the entire image itself. However, this adds another variable, what is the optimal patch size? There isnât a clear-cut answer and researchers select different patch sizes based on their specific needs. However, some patch sizes are more often used and are selected as default values. When selecting patch size, it is important to consider several factors. Finding the optimal patch size is important because it plays a role in how long training takes and can also impact model accuracy. Patch size often depends on the overall goal of a work. For example, works looking to perform slide level classification more often use larger patch sizes such as 512 x 512 and 1024 x 1024 pixels [25, 32, 42]. This allows for more information to be captured by each patch used in training and gives a better overall view of the tissue elements and cell architecture. However, other studies such as those focused on object detection and labeling individual cells and nuclei, may

decide to use smaller patch sizes. Additionally, researchers need to decide whether they will use overlapping or non-overlapping patches. An example of the differences between overlapping and non-overlapping patches can be found in Figure 2.4.

In the CAMELYON challenge there are a wide range of approaches taken and this extends to selected patch size. The top submission on the leaderboard for CAMELYON17 uses a patch size of 704 x 704 pixels [32]. However, the other submission in the top 5 all use either 512 x 512 pixel patches, 1024 x 1024 pixel patches or some combination of the two [25, 32, 42].

A study focusing on segmenting whole slide images from the CAMELYON dataset tried two ensembles with different patch sizes, 256 x 256 non-overlapping patches and 1024 x 1024 overlapping patches. The ensemble that used 1024 x 1024 overlapping patches performed slightly better than the ensemble with 256 x 256 non-overlapping patches. This study is currently a top 5 score on the CAMELYON17 leaderboard [25].

One work focused-on object detection of signet ring cells. Images from 10 different organs were used, with breast among them. Out of 127 images, each had 3 patches of size 2000 x 2000 selected for annotation with a bounding box. A total of 12,381 signet ring cells were annotated. However, due to overcrowding, some signet ring cells were not able to be annotated. This work is also a top 5 scorer on the CAMELYON17 leaderboard [34].

Another work used randomly selected 1000 x 1000 pixel sized patches for the task of tumor region recognition. The patches had to be downsampled four times to 224 x 224 to satisfy the requirements of their selected model, MobileNetV2 [24].

There are many instances where a larger patch size, such as 512 x 512, is initially chosen and then cropped or resized to a smaller size like 128 x 128 to make the image better match the selected modelâs input size [41]. One work applied this, originally selecting patch sizes of 512 x 512 before randomly cropping to 448 x 448. The 448 x 448 pixel patches then went through dimension reduction to achieve a size of 224 x 224 for training with the

EfficientNet framework. This study found an improvement in results in both slide level classification and segmentation tasks with randomly cropped patches [12].

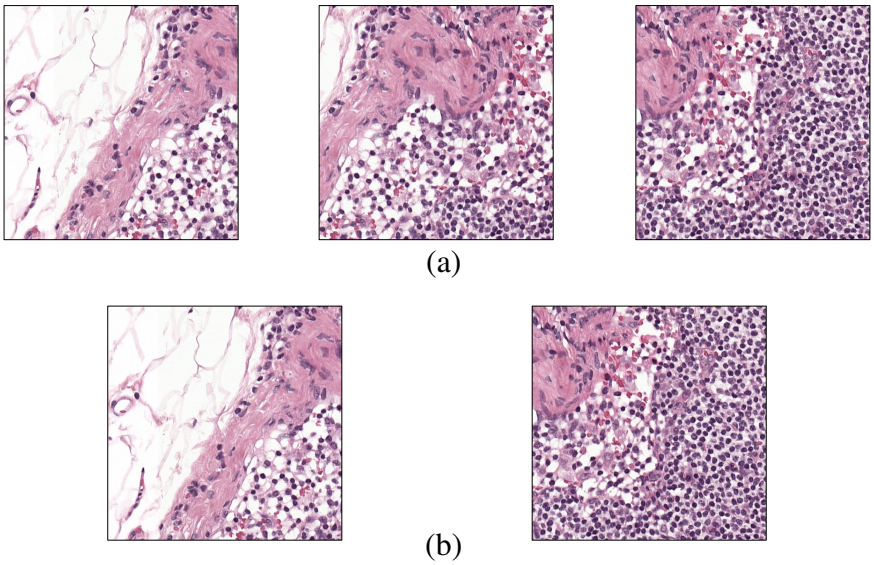


Figure 2.4: An example of overlapping (a) and non-overlapping (b) patches. These patches cover the same image region, but with overlapping, three patches are needed to cover the same area as two non-overlapping patches. The first and last patches of (a) match the patches of (b), but the middle patch of (a) is a combination of the patches from (b). Note: these patches were generated from whole slide images in the CAMELYON17 dataset [35].

2.5.4 Annotation Aggregation

Training is often performed at a lower annotation level such as patch level while expecting a final output at a higher level such as slide level or patient level. In these cases, aggregation is needed to combine the results from many patches to achieve the output for the higher level [15]. One study converted from patch level to slide level predictions for tumor and tumor bed detections. In this case, if one or more patches were determined to be positive or a tumor bed was detected, the entire WSI would be labeled as tumor positive [12]. Another study proposed a diffusion model for aggregating from patch level to slide level [22].

One study used patch level classification to extract features for slide level classification.

86.67% accuracy was achieved using Inception for patch level classification. An overall accuracy of 90.43% was achieved for the slide level classification of normal, benign, in situ carcinoma, and invasive carcinoma [36].

2.5.5 Thresholding

When working with patches in an image, there are hundreds of thousands of possible patches to be selected depending on the patch size selected. Patch selection methods vary between studies, with some studies performing random patch selection while others incorporate algorithms to select âbestâ patches [22]. However, one thing they all have in common is avoiding irrelevant patches with no cells and only background material. In most whole slide images there are large background areas that are irrelevant for image analysis [54]. With any patch selection technique, preprocessing is typically performed to eliminate irrelevant background patches. Often, thresholding is used to separate the image background from the relevant material. Thresholding is a technique that maps all image pixels into one of two groups. This technique is best used when there is high variance between an imageâs background and foreground. One popular method of thresholding in whole slide imaging is the Otsu thresholding technique. The Otsu threshold is determined by finding the maximum inter-class variance [39, 58]. While the Otsu threshold is popular in whole slide image segmentation of background and foreground, there are instances where it is not as effective. For instance, in Khened et al. (2021), the Otsu threshold could not be used to segment the CAMELYON dataset due to black regions within the WSI. Instead, the black pixels were changed to white first and then a median blur filter of size 7x7 kernel was used prior to performing the Otsu thresholding. [25].

While the Otsu threshold is popular, it is not the only thresholding method used. One study used a custom threshold to segment areas without nuclei from the image as regions that lack nuclei are not relevant in tumor identification. Their thresholding removed any

regions that met the following criteria: hue between 0.5 and 0.65, saturation greater than 0.1, and value between 0.5 and 0.9. These bounds were derived from experimentation with whole slide images, and patches with at least 25% foreground were included in the study [12]. Neural networks have also been applied to the segmentation task of tissue sample from its background with success [3].

2.5.6 Staining Techniques

Another common issue is variability in slide images. Although hematoxylin and eosin (H&E) staining is the most commonly used staining technique, it does have some drawbacks. This staining technique does not label the nuclei and cytoplasm in cells exclusively. Sometimes other staining techniques are used such as fluorescent staining, which is more common in tissue morphology clinical research. Whole slide image datasets using H&E stained slides that are publicly available are already scarce, so these alternative staining methods have limited annotated data to be used for machine learning. One study attempted to bridge H&E stained images with fluorescent stained images. Due to color variations, cross analysis can be difficult. Through methods involving color normalization techniques for preprocessing and nuclei extraction, they were able to create a model that had 89.6% accuracy in identifying tumor regions in H&E images and 80.5% accuracy in identifying those same regions in fluorescent stained slides. Further work into cross analysis between staining methods will increase the amount of available data for all types of stained whole slide image analysis [24].

2.5.7 Tools for Whole Slide Image Analysis

Several approaches have produced free and open-source software to aid others conducting research in this area. One available tool is DigiPathAI. This is a generalized deep learning-based framework for histopathology tissue analysis. When creating DigiPathAI,

four main problems were addressed - the large size of WSI images, minimal training samples, stain variability and extraction of clinically relevant features. Four datasets were used in training the model including CAMELYON16 and CAMELYON17 along with DigestPath (colon) and PAIP (liver). DigiPathAI used an ensemble of 3 fully convolutional networks - Deeplabv3, Inception-ResNet and DenseNet. A divide and conquer approach was taken for the WSI image size problem. Patches of the image were selected and segmented. Once all patches were segmented, they stitched together the segments to generate the whole slide image segmentation. The researchers used data augmentation to combat a lower number of training samples as well as to generalize across different staining and scanning protocols. This included horizontal/vertical flip, rotations, and Gaussian blurring and color augmentation [25].

MIA (Microscopic Image Analyzer) is another open-source tool developed for deep learning on microscopic images. MIA provides a graphical user interface for using deep learning tools for classification, segmentation, and object detection of microscopic images. By providing the graphical user interface, programming skills are not required to work with MIA. MIA simply requires training data, although the user needs to be able to select a model and hyperparameters. MIA also provides image labeling tools for annotating datasets [27].

The creators of the CAMELYON dataset also have created an open-source tool for visualizing and interacting with the CAMELYON dataset. This tool is called ASAP (Automated Slide Analysis Platform) and works on Linux and Windows operating systems. ASAP offers tools for both viewing and annotation [35].

2.5.8 Comparing Machine Learning Approaches to Pathologist Analysis

In 2017, a study put pathologists and coding teams up to the CAMELYON16 challenge. They split pathologists into two groups and provided them with the same WSI images for

two tasks - metastases identification through pixel level annotation and slide level labeling of metastases. 129 WSI were provided for annotation. The first group of pathologists was given a time constraint of two hours while the second group had no time constraint. The group without time constraint took approximately 30 hours to assess all 129 images. The challenge was open to coding teams and 32 total algorithms were submitted across 23 teams. Of the 32 algorithms, 25 were based on deep convolutional nueral networks, showing their popularity for whole slide imaging tasks. GoogleLeNet team scored 0.994 AUC on the image classification task. In comparison, the median AUC for pathologists without time constraint was 0.966 and 0.81 for pathologists with time constraint. This showed that the model outperformed pathologists with time constraint. This is more realistic since pathologists have many cases to analyze and a limited amount of time. The algorithm was comparable to the results achieved by human pathologists with unlimited time to view and classify whole slide imaging [7]. Importantly, while not perfect, CNNs can be used to predict the phenotype of breast cancers, potentially reducing the need for expensive biomarker assays [13, 51]. Deep learning algorithms also have the potential to predict patient outcome, which is hard to achieve with pathologic evaluation of a breast cancer tissue sample [17, 47].

2.6 CONCLUSION

Since the advent of whole slide imaging, research in digital pathology has surged. Computer-aided diagnosis and medical image analysis have become a focus for researchers, especially in digital pathology. While there are still many challenges when working with whole slide images, current research shows promise for finding the solutions to overcome these challenges. Deep learning in digital pathology has the potential to become a powerful tool for pathologists and assist them with the high demand of the field, which could ultimately lead

to better care for breast cancer patients. This paper provides background information about breast cancer, whole slide images, and deep learning along with key challenges and the techniques employed by researchers in the field to overcome these challenges. The implementation of deep learning shows potential for incredible benefits that can both propel digital pathology forward as well as help patients.

Chapter 3

Methodology

While whole slide image scanners have popularized machine learning research in digital breast pathology, there are still some research avenues that need more attention, such as recurrence prediction. The approach in this work uses only H&E-stained images to train a machine learning model to predict breast cancer recurrence. No tumor ROI annotations or clinical data were included in the dataset. A total of 424 cases were collected from ECU Health. For each case, tissue was biopsied, sectioned, and then stained with hematoxylin and eosin. All slides were prepared in the Department of Pathology at ECU Health. Each tissue section was scanned using Philips IntelliSite Pathology 760001 Ultra-Fast Scanner 1.8. All images were uploaded to Philips Digital Pathology Solution. This allowed for each image to be viewed at different magnifications. All images were saved using a scan factor of 40x (magnification) and were saved using the TIFF format. Each case was assigned a recurrence score between 0 and 100. The cases were sorted into recurrence risk categories based on this score. Recurrence scores of 25 or less were deemed low-risk while cases with recurrence scores of 26 or greater were sorted into the high-risk category. Due to the lower natural occurrence of high-risk scores, only 51 of these cases were high-risk. All available high-risk cases were used. A handpicked group of 51 low risk cases was selected for a total of 102 cases. These cases were handpicked rather than randomly selected to ensure the data was consistent among a variety of factors for the low and high-risk groups. The

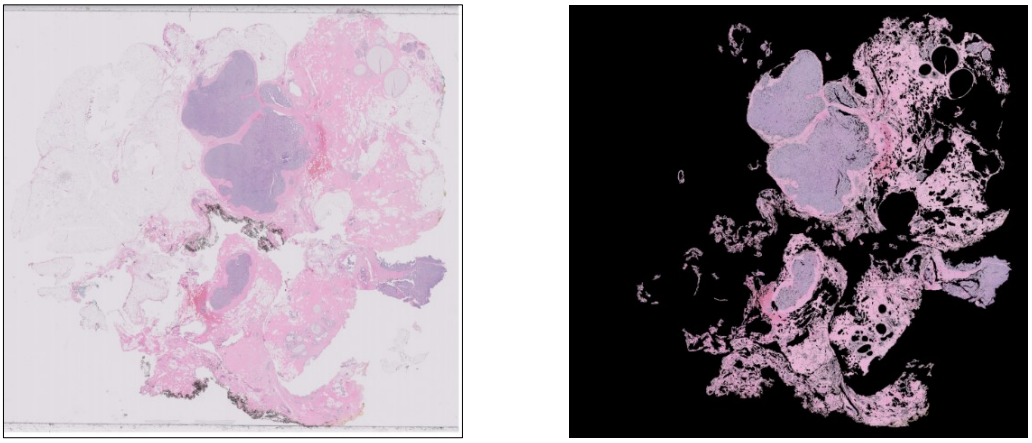
Tumor Statistic	Low	High
Tumor Grade		
1	7	4
2	22	22
3	22	25
No. of Lymph Nodes		
0	28	35
1	10	9
2	4	4
Other	9	3

Demographic	Low	High
Age Range		
≤ 30	0	1
31-40	2	2
41-50	6	6
51-60	17	17
61-70	15	16
71-80	10	8
81-90	1	1
Race		
Black	25	25
Hispanic	2	2
White	24	24

Table 3.1: Dataset Statistics. Here, two tables break down the dataset used based on demographics and tumor statistics. Information on patient age range, race, tumor grade, and number of lymph nodes involved are included.

data in both categories was balanced on patient age group, tumor grade, race, and number of lymph nodes impacted. The exact breakdowns for these categories across the low and high-risk groups can be seen in Table 3.1. These were kept consistent across both the low and high-risk groups to reduce the chance the model would be influenced by factors other than recurrence.

Due to the large size of whole slide images, image patches were used in place of the whole slide image for each case. Multiple patch sizes and patch quantities were extracted so their impact on the model accuracy could be analyzed. After cases were selected to be included in the dataset, they were analyzed for patch quality based on the total amount of tissue available in each patch. The code to analyze patch quality was modified from code that is publicly available at https://github.com/deroneriksson/python-wsi-preprocessing [16]. Code from this repository was updated to fix some bugs and errors, as well as to customize file paths and analysis variables before being used. Two CSV files were generated for each image using the code, one for each patch size (256 x 256 pixels



a) Original RGB WSI b) WSI after filters

Figure 3.1: Original RGB and filtered images. Image (a) is a compressed and downsized version of a whole slide image used in the dataset. Image (b) shows the image after all filters have been applied. Any pixels not mapped to 0 (black) are now considered tissue and impact each tile's score.

and 512 x 512 pixels). New code was written to extract the top tiles from each image using the CSV data.

Each slide was filtered to remove background, errant pen markings, and other irrelevant information on the slide. Any image data determined to not be tissue was mapped to a black value using masking to get a clear picture of the tissue within the image. An example of a whole slide image from the dataset before and after filtering can be seen in Figure 3.1. Each possible patch was scored based on a number of factors including tissue percentage, tissue quantity factor, color factor, and saturation and value factor [16].

The tissue quantity per patch was determined by what percentage of pixels were not 0 after the filter masks were applied. High and low tissue thresholds were set to 75% and 10% respectively. Patches with a tissue percentage at least equal to the high threshold were marked as "High" tissue quantity. Patches with tissue percentages between the high and

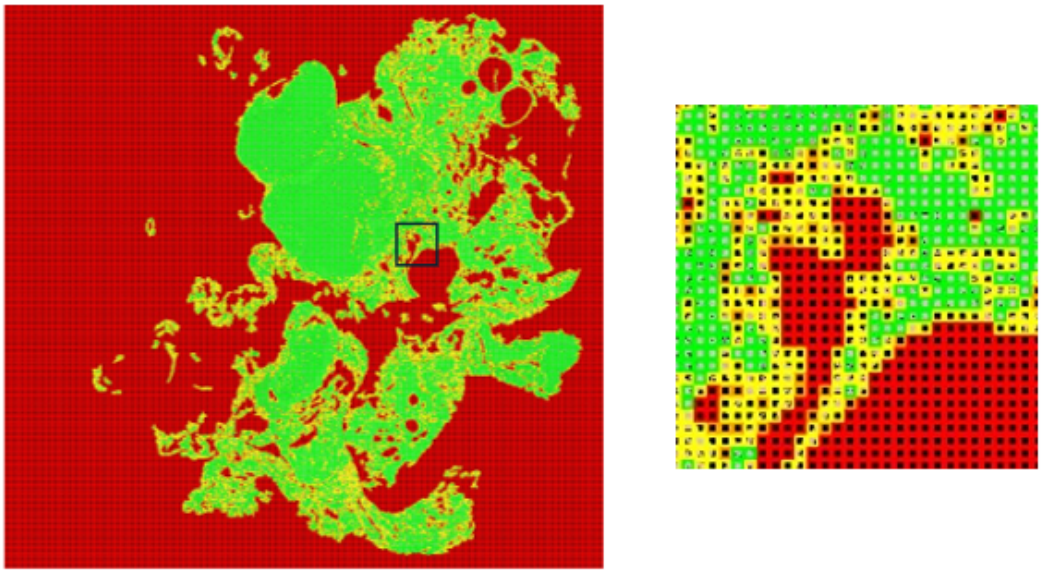
low thresholds were considered to have "Medium" tissue quantity. Patches with less tissue percentage than the low threshold but more than 0% were considered "Low", and patches with 0% were considered "None". Each patch received a quantity factor based on the amount of available tissue. The values for high, medium, low, and none tissue quantities are as follows, respectively: 1.0, 0.2, 0.1, 0.0. A color factor was also determined for each patch. The color factor favors patches that contain more purple (hematoxylin) values. Lastly, the saturation and value factor is used to reduce the scores of patches with small standard deviations of HSV saturation and HSV values. This can help eliminate blurred patches as significant patches often have broad standard deviations [16].

Once the tissue quantity factor, color factor, and saturation and value factor were all computed, they were multiplied together to get a combined factor to be used in the final score calculation. The score calculation can be seen in equation 3.1. Scores were normalized to values between 0 and 1 before use, as can be seen in equation 3.2. These equations and factor value calculations come from the GitHub previously cited [16].

$$score = (tissuepercent)^2 * ln(1 + combinedfactor)/1000.0 \tag{3.1}$$

$$scorenormalized = 1 - (10.0/10.0 + score) \tag{3.2}$$

Patches were considered high quality if they had a score of 75% or better. An example of an image patch scoring can be seen in Figure 3.2. For both patch sizes, the top 1,000 and top 2,000 scoring patches were extracted for analysis. Later, a set of 3,000 tiles of size 256 x 256 pixels were extracted from each image. If an image had less than 3,000 patches that were considered high quality, then first all high quality patches were extracted. Then, these patches were oversampled until a total of 3,000 patches were extracted for each whole slide image. Data on the total number of patches and high quality patch statistics can be found in Table 3.2.



a) Whole Slide Patch Summary b) Section Patch Summary

Figure 3.2: Patch overlay. This shows the same whole slide image as shown in Figure 3.1. Here, the patch summary overlay is placed on top of the image. The smaller squares each represent a patch and their outline color represents the tissue quantity per patch. Green indicates high amounts of tissue, yellow indicates medium, orange indicates low and red indicates no tissue.

Once patches had been extracted from each whole slide image, the patches were sorted into subfolders "high" and "low" based on their recurrence risk category. A convolutional neural network was used to extract features from each image. The CNN used for feature extraction was ResNet-18. ResNet-18 is a convolutional neural network pretrained on ImageNet. It implements residual blocks which contain "shortcut connections" sometimes also referred to as "skip connections". These skip connections allow for some input to bypass some layers and directly connect to layers deeper in the network. This aids with the vanishing gradient problem sometimes seen in other neural networks [21]. Feature extraction via ResNet-18 was done using the Python library Img2Vec. Each image yielded 512 features regardless of image size. For each patch processed, the patch features, slide level label, and case number were saved. Once features were extracted for each image, they were saved

Measurement/Patch Size	256 x 256	512 x 512
Smallest # High Quality Patches	538	51
Largest # High Quality Patches	60,371	15,113
Average # High Quality Patches	17,651	4,306
Standard Deviation of High Quality Patches	11,724	3,025

Table 3.2: Patch Statistics. Patch statistics are shown for high quality tiles for patch sizes 256 x 256 pixels and 512 x 512 pixels. Patches are considered "high quality" if their overall score is at least 75%.

for processing for prediction. Feature extraction was performed on a Tesla T4 GPU and averaged approximately 8.5 iterations (images processed) per second. The total time to process feature extractions for each dataset varied based on the patch size and number of patches per whole slide. Once the features were loaded for prediction, the next step was to sort them into bags based on which case they originated from. Once all bags had been created and they were verified to contain the correct number of features, the data was split into 80% training and 20% testing.

The data underwent preprocessing to convert the three-dimensional arrays into two dimensional arrays for analysis. Data was standardized using the Python library Standard Scalar. Once the data was preprocessed, it was fed into a logistic regression model. This method of using a CNN for feature extraction and replacing the fully connected layers with another machine learning model matches the feature extractor hybrid method of transfer learning. Grid search was used to determine the best parameters for each dataset for logistic regression. The parameters used for the best scoring dataset were C=10.0, penalty='l1', and scorer='liblinear'. After the models were created, they were tested on a variety of measurements including accuracy, F1 score, and AUC. Cross validation of the training dataset as well as cross validation of the entire dataset was also used to evaluate their ability to generalize the training data in addition to the testing data. The code used in this work is publicly available and can be accessed at https://github.com/mad-rose/breastcancer-

recurrence-prediction.

Chapter 4

Results

Model performance was evaluated in two main ways. First, holdout testing was performed with 20% of the data that was originally selected for testing using the train_test_split function from Sci-kit learn. The logistic regression model was evaluated on this data. The model's accuracy, F1 score, and AUC were evaluated. These results are displayed in Table 4.1. 5-fold cross validation was performed on the training set to assess the model's ability to generalize across the training data.

Due to the size of the dataset being relatively small (n = 102), 5-fold cross validation on the entire dataset was performed. These results can be seen in Table 4.2. Due to the size of the dataset, this cross validation is likely more indicative of the model's overall performance on the entire dataset. When performing cross validation on the dataset, 5-fold

Patch Type	Accuracy	F1	AUC	Training Cross Validation
256 x 3,000	0.524	0.540	0.676	0.481 $\pm$0.104
256 x 2,000	**0.714**	**0.750**	**0.736**	**0.640 $\pm$0.067**
256 x 1,000	0.619	0.636	0.704	0.530 $\pm$0.078
512 x 2,000	0.571	0.640	0.593	0.517 $\pm$0.080
512 x 1,000	0.571	0.609	0.630	0.518 $\pm$0.066

Table 4.1: Holdout testing results. Results from holdout testing for each patch type are reported across multiple metrics including accuracy, F1 score, AUC, and cross-validation of the training data.

Patch Type	Accuracy	F1
256 x 3,000	0.462 ±0.075	0.471 ±0.095
256 x 2,000	**0.628 ±0.044**	**0.639 ±0.060**
256 x 1,000	0.588 ±0.044	0.596 ±0.037
512 x 2,000	0.521 ±0.117	0.521 ±0.126
512 x 1,000	0.529 ±0.053	0.515 ±0.072

Table 4.2: 5-fold cross-validation testing results. Accuracy and F1 scores are reported for each patch type when the entire dataset (102 cases) was used for 5-fold cross-validation.

cross validation was used and was measured on accuracy and F1 score. The results for each model are indicated for the model's best parameters as determined by grid search. The best performing patch type across all evaluations was the patch size 256 x 256 pixels with 2,000 patches per whole slide image. Generally, a patch size of 256 x 256 pixels performed better than a patch size of 512 x 512 pixels, with the exception of the 256 x 256 x 3,000 model that utilized patch oversampling. A larger patch quantity of 2,000 patches appeared to perform better in comparison to only using 1,000 patches for the patch size 256 x 256. However, the 3,000 patch model performed more poorly than both the 1,000 and 2,000 patch models. The performance of both the 1,000 and 2,000 patch models was similar for the patch size 512 x 512.

Chapter 5

Discussion

Overall, the patching method of using 2,000 patches of size 256 x 256 performed the best across all measurements. This includes both the holdout testing and the 5-fold cross validation of the entire dataset. The lower scores for cross validation when compared to the accuracy for the holdout data could suggest that the selected holdout data was easier to predict or may be more similar to the training data used. The cross-validation scores for the entire datasets are likely a better indicator of overall model performance since multiple folds of testing were completed. Interestingly, the 256 x 256 x 3,000 images utilizing oversampling had the lowest accuracy and cross validation scores. There are a few reasons this method might have performed so poorly. First, if the relevant patches needed for recurrence prediction were already captured within the first 2,000 patches extracted, extracting 3,000 patches may have only introduced extra noise into the dataset. The oversampling method might have also caused some irrelevant patches to be oversampled which would also introduce noise into the dataset.

In the 5-fold cross validation for the entire dataset, the standard deviation values for the 256 x 256-pixel sized patch datasets, excluding the 256 x 3,000 oversampling model, are smaller than the standard deviation for the datasets of the 512 x 512-pixel sized patches. This could suggest that the models using the 256 x 256 patch size are able to generalize better across more data, since the scores for each fold are closer to the cross validation

mean. One reason for this could be that the smaller patches may be able to capture important information near the tissue edge. When a larger patch size is used, this could become a lower quality patch due to increased background presence since the tissue is right at the edge between the sample and the background.

One factor that could be hindering the model performance is the varying size of tissue samples in each case. These statistics can be seen in Table 3.2. The number of available patches of high quality (patches with a score of at least 75%) varied greatly among cases. With the patch size of 256 x 256 pixels, the whole slide image with the fewest number of high quality patches contained 538 high quality patches. The image that contained the most high quality patches had 60,371 high quality patches. This is a range of 59,833 patches. Given the average number of high quality patches at 17,651 and the standard deviation of 11,724, it can be concluded that there is high variability of high quality patches amongst the images in the dataset. For the patch size of 512 x 512 pixels, the smallest and largest number of high quality patches from images in the dataset were 51 and 15,113 respectively. This represents a range of 15,062 patches. The average number of high quality patches for this dataset was 4,306 and the standard deviation was 3,025. These high standard deviations contribute to a large difference in the total high quality patches available for analysis, which can make it difficult to determine the total number of patches that should be extracted from each image.

Both ends of this spectrum present their own challenges. On the smaller side, for cases with fewer available high quality patches, the number of these available patches is sometimes less than the total number of patches used in each dataset for training. This would result in some patches being included in the dataset that are of lower quality and include more irrelevant data. This could introduce noise into the dataset which could impact model accuracy. On the other side, having such a large number of high quality tiles makes it less likely that a sample of 1,000 or 2,000 patches would include the tumor region of the image.

This could mean that the selected patches are not as relevant to recurrence and important patches for predicting recurrence were not captured in the dataset which could also impact the model's accuracy.

It's possible that higher performance on certain subsets of the data could indicate that there are some cases in the data which are more similar to each other. Therefore, the model is able to learn from one of these cases and accurately predict the other. However, lower scores in other subsets could indicate that some cases in the dataset are more unique or there are not as many similar cases and therefore the model is not able to learn these features as well which results in poorer performance on these cases.

The results in this study can be difficult to compare to other works discussed due to differences in data and prediction goals. For example, one of the previously mentioned studies saw a similar cross validation score, but focused on predicting three year recurrence based on actual patient outcome. Their 10-fold validation accuracy was found to be 62.4%, which is similar to the best scorer in this study which scored 62.8% accuracy across 5-fold validation of the entire dataset. However, since this study and the related study look to predict different recurrence metrics, it is difficult to compare them [47].

Another more similar study also discussed attempted to predict Oncotype DX recurrence risk category, however it is also somewhat difficult to compare to the work in this study due to differences in the data used. BCR-Net uses an intelligent sampling method of patches from within an annotated region of interest. BCR-Net achieved 0.700 ± 0.030 accuracy across 5-fold validation. This score is similar to the score from this study testing the 256 x 256 x 2,000 model on the hold out data, with 0.714 accuracy. However, the 5-fold cross validation for this model was $0.628 \pm .044$. This indicates that although the models may see similar results among certain data subsets, the BCR-Net model is able to generalize and perform better across more folds of data [51].

Some limitations of this study include the data availability. The dataset used was rela-

tively small (n = 102). While more cases were available, all available high-risk cases were utilized, and a select number of low-risk cases were used to keep the dataset balanced. Since high-risk cases are less common, this was the limiting factor on the dataset. Another limitation is the data used in this study came from one hospital and one whole slide image scanner. More analysis is needed using data from multiple hospitals and multiple whole slide image scanners to increase variability.

Chapter 6

Conclusion

Since whole slide image scanners were introduced, there has been a strong push towards using machine learning methods within digital pathology. Great strides have been made on tasks such as breast cancer diagnosis and metastasis detection, however there is not as much research into outcome predictions, including recurrence risk. Many studies that attempt to predict recurrence scores or categories use clinical data or image data containing region of interest annotations. If an image-based model could be developed, this could greatly decrease the cost associated with recurrence score testing since H&E-stained images are routinely collected for breast cancer diagnosis [47]. Additionally, if these predictions could be done without the need for pathologist annotations of tumor regions of interest, additional pathology work could be reduced.

Within this research area there are many potential future work opportunities. One thing that would be interesting would be to use a CNN finetuned on the patches themselves to perform feature extraction. This might increase model accuracy by allowing the CNN to extract more relevant features from each of the image patches. It would also be interesting to investigate more patch sizes. Since in this study, the smaller patch size generally performed better, more investigation should be done into even smaller patch sizes, possibly even as small as one cell per patch, to see if an improvement in accuracy continues. Additionally to add onto this work, one future possibility is obtaining more varied data from

different hospitals and scanners to test. If this were done it is possible techniques for color normalization would be needed to be added to account for this variability. Also, including available patient data such as age and tumor grade may also help the model perform better with its predictions.

Despite not having any region of interest annotations from a pathologist in this dataset, another possible avenue for future work could include using this dataset with preexisting models available for tumor segmentation. This way, a region of interest map/mask could be used without requiring extra work from a pathologist. This could allow for better patch selection.

Another possible area for future work would be to evaluate a regression model. Since recurrence scores have wide ranges, there could be significant differences between a recurrence score of 1 and a recurrence score of 25 despite both being considered low-risk. Creating a model to try and predict the actual recurrence number might also help discover similarities between some cases.

With the emergence of vision transformers, this presents an exciting future research opportunity. A vision transformer could potentially be used to find connections between the patches to increase patch selection quality.

In summary, in this study a dataset containing only H&E-stained images was used to train a machine learning model with the goal of predicting breast cancer recurrence. Image patches were collected from each whole slide image and features were extracted from each image patch using a ResNet-18 architecture. Image features were then aggregated together into bags, where each slide represented a bag and the patch features represented instances inside the bag. These bags were used to train a logistic regression model which was evaluated across multiple metrics including accuracy, and F1. The best scoring model across all metrics used 2,000 patches of size 256 x 256 pixels from each whole slide image. This model observed a mean accuracy of 0.628 ± 0.044 in 5-fold cross validation across the en-

tire available dataset. Analysis of various patch sizes and patch quantities is also performed to determine their impact on the overall accuracy of the model. A few possible factors that could contribute to poorer model performance are discussed in this study, including high variance amongst available tissue in each whole slide image. Also, the dataset used in this study was limited in size and variance due to coming from only one hospital and one whole slide image scanner. More research is needed across larger and more varied datasets to further assess the possibilities of using only H&E-stained images for ODX recurrence risk category prediction. Overall, in this study patch-based methods were used to perform machine learning to predict breast cancer recurrence risk categories.

BIBLIOGRAPHY

[1] Camelyon17 - grand challenge. https://camelyon17.grand-challenge.org/Home/. Accessed: 2023-11-27.

[2] ALKABBAN, F. M., AND FERGUSON, T. Breast cancer. In *StatPearls* (2020), Treasure Island (FL):Stat Pearls Publishing.

[3] ALOMARI, R. S., ALLEN, R., SABATA, B., AND CHAUDHARY, V. Localization of tissues in high-resolution digital anatomic pathology images. In *Medical Imaging 2009: Computer-Aided Diagnosis* (2009), vol. 7260.

[4] ALZUÂBI, A., NAJADAT, H., DOULAT, W., AL-SHARI, O., AND ZHOU, L. Predicting the recurrence of breast cancer using machine learning algorithms. *Multimedia Tools and Applications 80* (4 2021), 13787–13800.

[5] ARAUJO, T., ARESTA, G., CASTRO, E., ROUCO, J., AGUIAR, P., ELOY, C., ..., AND CAMPILHO, A. Classification of breast cancer histology images using convolutional neural networks. *PLoS ONE 12*, 6 (2017).

[6] BANCROFT, J. D., AND LAYTON, C. 10 - the hematoxylins and eosin. In *Bancroft's Theory and Practice of Histological Techniques*, S. K. Suvarna, C. Layton, and J. D. Bancroft, Eds., eighth edition ed., vol. 1. Elsevier, 2019, pp. 126–138.

[7] BEJNORDI, B. E., VETA, M., DIEST, P. J. V., GINNEKEN, B. V., KARSSEMEIJER, N., LITJENS, G., LAAK, J. A. V. D., ..., AND VENÃ¢NCIO, R. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - Journal of the American Medical Association 318* (2017).

[8] BHUSHAN, A., GONSALVES, A., AND MENON, J. U. Current state of breast cancer diagnosis, treatment, and theranostics. *Pharmaceutics 13*, 5 (2021).

[9] CARLSON, J. J., AND ROTH, J. A. The impact of the oncotype dx breast cancer assay in clinical practice: a systematic review and meta-analysis. *Breast Cancer Research and Treatment 141* (8 2013), 13–22.

[10] CHAN, J. K. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. *International Journal of Surgical Pathology 22*, 1 (2014), 12–32.

[11] CIGA, O., AND MARTEL, A. L. Learning to segment images with classification labels. *Medical Image Analysis 68* (2021).

[12] CIGA, O., XU, T., NOFECH-MOZES, S., NOY, S., LU, F. I., AND MARTEL, A. L. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific reports 11* (2021).

[13] COUTURE, H. D., WILLIAMS, L. A., GERADTS, J., NYANTE, S. J., BUTLER, E. N., MARRON, J. S., ..., AND NIETHAMMER, M. Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype. *npj Breast Cancer 4* (2018).

[14] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

[15] DIMITRIOU, N., ARANDJELOVIÄ, O., AND CAIE, P. D. Deep learning for whole slide image analysis: An overview. *Frontiers in Medicine 6* (2019).

[16] ERIKSSON, D. python-wsi-preprocessing. GitHub repository, 2017.

[17] FERNANDEZ, G., PRASTAWA, M., MADDURI, A. S., SCOTT, R., MARAMI, B., SHPALENSKY, N., ..., AND DONOVAN, M. J. Development and validation of an ai-enabled digital breast cancer assay to predict early-stage breast cancer recurrence within 6 years. *Breast Cancer Research 24* (2022).

[18] GADERMAYR, M., AND TSCHUCHNIG, M. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations amp; future potential. *Computerized Medical Imaging and Graphics 112* (3 2024), 102337.

[19] GUPTA, J., PATHAK, S., AND KUMAR, G. Deep learning (cnn) and transfer learning: A review. *Journal of Physics: Conference Series 2273*, 1 (2022), 012029.

[20] GUPTA, V., MISHRA, V. K., SINGHAL, P., AND KUMAR, A. An overview of supervised machine learning algorithm. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (2022).

[21] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (2016).

[22] HOU, L., SAMARAS, D., KURC, T. M., GAO, Y., DAVIS, J. E., AND SALTZ, J. H. Patch-based convolutional neural network for whole slide tissue image classification. In *IEEE Conference on Computer Vision and Pattern Recognition* (2016).

[23] HOWARD, A. G., ZHU, M., CHEN, B., KALENICHENKO, D., WANG, W., WEYAND, T., ..., AND ADAM, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR abs/1704.04861* (2017).

[24] HUANG, P. W., OUYANG, H., HSU, B. Y., CHANG, Y. R., LIN, Y. C., ..., Y. A. C., AND PAI, T. W. Deep-learning based breast cancer detection for cross-staining histopathology images. *Heliyon 9*, 2 (2023).

[25] KHENED, M., KORI, A., RAJKUMAR, H., KRISHNAMURTHI, G., AND SRINIVASAN, B. A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific Reports 11* (2021).

[26] KIM, H. E., COSA-LINAN, A., SANTHANAM, N., JANNESARI, M., MAROS, M. E., AND GANSLANDT, T. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging 22*, 1 (2022).

[27] KÖRBER, N. Mia is an open-source standalone deep learning application for microscopic image analysis. *Cell Reports Methods 3*, 7 (2023).

[28] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM 60*, 6 (2017).

[29] KRUPINSKI, E. A., JOHNSON, J. P., JAW, S., GRAHAM, A. R., AND WEINSTEIN, R. S. Compressing pathology whole-slide images using a human and model observer evaluation. *Journal of Pathology Informatics 3* (2012), 17.

[30] LECUN, Y., BOTTOU, L., BENGIO, Y., AND HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[31] LECUN, Y., HINTON, G., AND BENGIO, Y. Deep learning. *Nature 521* (2015), 436–444.

[32] LEE, S., CHO, J., AND KIM, S. W. Automatic classification on patient-level breast cancer metastases, 2021.

[33] LENAIL, A. Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software 4*, 33 (2019), 747.

[34] LI, J., YANG, S., HUANG, X., DA, Q., YANG, X., ..., Z. H., AND LI, H. Signet ring cell detection with a semi-supervised learning framework. In *Information Processing in Medical Imaging* (2019), vol. 11492, pp. 842–854.

[35] LITJENS, G., BANDI, P., BEJNORDI, B. E., GEESSINK, O., BALKENHOL, M., BULT, P., ..., AND VAN DER LAAK, J. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: The camelyon dataset. *GigaScience 7*, 6 (2018).

[36] MI, W., LI, J., GUO, Y., REN, X., LIANG, Z., ZHANG, T., AND ZOU, H. Deep learning-based multi-class classification of breast digital pathology images. *Cancer Management and Research 13* (2021).

[37] NOUNOU, M. I., ELAMRAWY, F., AHMED, N., ABDELRAOUF, K., GODA, S., AND SYED-SHA-QHATTAL, H. Breast cancer: Conventional diagnosis and treatment modalities and recent patents and technologies. *Breast Cancer: Basic and Clinical Research 9s2* (2015).

[38] O'SHEA, K., AND NASH, R. An introduction to convolutional neural networks. *ArXiv e-prints 10* (2015).

[39] OTSU, N. Threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics 9*, 1 (1979), 62–66.

[40] PANTANOWITZ, L., FARAHANI, N., AND PARWANI, A. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International 7* (2015), 23–33.

[41] PHAN, N. N., HSU, C. Y., HUANG, C. C., TSENG, L. M., AND CHUANG, E. Y. Prediction of breast cancer recurrence using a deep convolutional neural network without region-of-interest labeling. *Frontiers in Oncology 11* (2021).

[42] PINCHAUD, N. Camelyon17 challenge. 2019.

[43] POUYANFAR, S., SADIQ, S., YAN, Y., TIAN, H., TAO, Y., REYES, M. P., ..., AND IYENGAR, S. S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys 51*, 5 (2018), 1–36.

[44] ROSE, M., GERADTS, J., AND HERNDON, N. Deep learning in digital breast pathology. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 1* (2024), pp. 404–414.

[45] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., ..., AND FEI-FEI, L. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision 115* (2015), 211–252.

[46] SARI, C. T., AND GUNDUZ-DEMIR, C. Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. *IEEE Transactions on Medical Imaging 38*, 5 (2019), 1139–1149.

[47] SHI, Y., OLSSON, L. T., HOADLEY, K. A., CALHOUN, B. C., MARRON, J. S., GERADTS, J., ..., AND TROESTER, M. A. Predicting early breast cancer recurrence from histopathological images in the carolina breast cancer study. *npj Breast Cancer 9* (11 2023), 92.

[48] SIEGEL, R. L., MILLER, K. D., WAGLE, N. S., AND JEMAL, A. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians 73*, 1 (2023), 17–48.

[49] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In *ICLR* (2015).

[50] SPARANO, J. A., AND PAIK, S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology 26* (2 2008), 721–728.

[51] SU, Z., NIAZI, M. K. K., TAVOLARA, T. E., NIU, S., TOZBIKIAN, G. H., WESOLOWSKI, R., AND GURCAN, M. N. Bcr-net: A deep learning framework to predict breast cancer recurrence from histopathology images. *PLoS ONE 18*, 4 (2023).

[52] SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ..., AND RABINOVICH, A. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition* (2015).

[53] SZEGEDY, C., VANHOUCKE, V., IOFFE, S., SHLENS, J., AND WOJNA, Z. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2818–2826.

[54] VETA, M., PLUIM, J. P., DIEST, P. J. V., AND VIERGEVER, M. A. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering 61*, 5 (2014), 1400–1411.

[55] WAKILI, M. A., SHEHU, H. A., SHARIF, M. H., SHARIF, M. H. U., UMAR, A., KUSETOGULLARI, H., ..., AND UYAVER, S. Classification of breast cancer histopathological images using densenet and transfer learning. *Computational Intelligence and Neuroscience 2022* (2022).

[56] WATKINS, E. J. Overview of breast cancer. *Journal of the American Academy of Physician Assistants 32*, 10 (2019), 13–17.

[57] WHITNEY, J., CORREDOR, G., JANOWCZYK, A., GANESAN, S., DOYLE, S., TOMASZEWSKI, J., FELDMAN, M., GILMORE, H., AND MADABHUSHI, A. Quantitative nuclear histomorphometry predicts oncotype dx risk categories for early stage er+ breast cancer. *BMC Cancer 18* (12 2018), 610.

[58] XU, X., XU, S., JIN, L., AND SONG, E. Characteristic analysis of otsu threshold and its applications. *Pattern Recognition Letters 32*, 7 (2011).

[59] ZARELLA, M. D., BOWMAN, D., AEFFNER, F., FARAHANI, N., XTHONA, A., ABSAR, S. F., ..., AND HARTMAN, D. J. A practical guide to whole slide imaging a white paper from the digital pathology association. *Archives of Pathology & Laboratory Medicine 143* (2019), 222–234.

[60] ZHANG, A., LIPTON, Z. C., LI, M., AND SMOLA, A. J. *Dive into Deep Learning*. Cambridge University Press, 2021. `https://D2L.ai`.

[61] ZUJEWSKI, J. A., AND KAMIN, L. Trial assessing individualized options for treatment for breast cancer: the tailorx trial. *Future Oncology 4* (10 2008), 603–610.