Automated Localization of Breast Ductal Carcinoma *in Situ* in Whole Slide Images

by

Nikhil Seth

A thesis submitted in conformity with the requirements for the degree of Masters of Science
Department of Medical Biophysics
University of Toronto

© Copyright by Nikhil Seth 2019
Automated Localization of Breast Ductal Carcinoma *in Situ* in Whole Slide Images

Nikhil Seth

Masters of Science

Department of Medical Biophysics
University of Toronto

2019

Abstract

Duct segmentation in whole slide images (WSIs) is an important step needed to analyze breast ductal carcinoma in-situ (DCIS), an early form of breast cancer. Here, we trained several U-Net architectures – deep convolutional neural networks designed to output probability maps – to segment DCIS in WSIs and validate the optimal patch field of view necessary to achieve superior accuracy at the slide-level. A U-Net trained at 5x achieved the best test results (DSC = 0.771, F1 = 0.601), implying the U-Net benefits from seeing wider contextual information. A custom U-Net based architecture, trained to incorporate patches from all available resolutions, achieved test results of DSC = 0.759 (F1 = 0.682), showing improvement in the model’s duct detecting capabilities. Both architectures showed comparable performance to a second expert annotator on an independent test set. This is preliminary work for a pipeline targeted at predicting recurrence risk in DCIS patients.
Acknowledgments

I would like to express sincere gratitude to my supervisor Anne Martel, for encouraging me to pursue new research avenues and broaden my knowledge, as well as guiding me through my project. This work would not have been possible without her support. I would like to thank my committee members David Andrews, Martin Yaffe and Carlos Parra-Herran for pushing the boundaries of my knowledge and helping me become a more complete scientist. I would also like to thank all the other members of the Martel Lab for all their helpful tidbits of knowledge, with a specific shoutout to Shazia Akbar whose machine learning knowhow was invaluable to me. The Sunnybrook Research Institute and Department of Medical Biophysics has been an amazing place to work the last few years, filled with bright minds and great friends.

Finally, I would like to thank my wonderful parents, Hitesh and Vasu, for supporting me through my educational journey. They are truly my biggest inspirations.

The funding that made this work possible is the Canadian Cancer Society Grant #705772, Canadian Breast Cancer Federation.
# Table of Contents

1 Introduction.............................................................................................................................................................................1

1.1 Breast Cancer...........................................................................................................................................................................1

1.1.1 Breast Cancer Screening.....................................................................................................................................................3

1.1.2 Histological Assessment and Classification .......................................................................................................................3

1.1.3 Breast Cancer Treatment.....................................................................................................................................................5

1.1.4 Digital Pathology ..................................................................................................................................................................7

1.2 Machine Learning......................................................................................................................................................................8

1.2.1 Supervised vs. Unsupervised Learning...............................................................................................................................8

1.2.2 Classical Machine Learning..............................................................................................................................................9

1.2.3 Modern Machine Learning - Neural Networks & Deep Learning ......................................................................................11

1.2.4 Convolutional Neural Networks......................................................................................................................................14

1.2.5 Model Evaluation..............................................................................................................................................................16

1.3 Machine Learning in Digital Pathology.................................................................................................................................18

1.4 Thesis Outline.............................................................................................................................................................................20

2 Automated Localization of DCIS in WSIs .................................................................................................................................22

2.1 Introduction..............................................................................................................................................................................22

2.1.1 Related Work.........................................................................................................................................................................22

2.2 Materials & Methods...............................................................................................................................................................23

2.2.1 Dataset..................................................................................................................................................................................23

2.2.2 Training..................................................................................................................................................................................24

2.2.3 Resolution vs. Field of View ............................................................................................................................................24

2.2.4 Multi-resolution Network ..................................................................................................................................................25

2.2.5 Ensemble Models...............................................................................................................................................................25

2.2.6 Evaluation.............................................................................................................................................................................26

2.2.7 Post-processing.................................................................................................................................................................27
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rm</td>
<td>Custom multi-resolution U-Net architecture – 3 downsampling arms</td>
</tr>
<tr>
<td>9ch</td>
<td>Multi-resolution U-Net architecture – 9 channel image input</td>
</tr>
<tr>
<td>Adam</td>
<td>Adaptive moment estimation</td>
</tr>
<tr>
<td>ADH</td>
<td>Atypical ductal hyperplasia</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial neural network</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the ROC curve</td>
</tr>
<tr>
<td>BCS</td>
<td>Breast conserving surgery</td>
</tr>
<tr>
<td>CAD</td>
<td>Computer-aided diagnosis</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional neural network</td>
</tr>
<tr>
<td>DCIS</td>
<td>Ductal carcinoma in situ</td>
</tr>
<tr>
<td>Dropout</td>
<td>Randomly deactivating neuron connections</td>
</tr>
<tr>
<td>ECDP</td>
<td>European congress on digital pathology</td>
</tr>
<tr>
<td>ELU</td>
<td>Exponential linear unit</td>
</tr>
<tr>
<td>ER</td>
<td>Estrogen receptor</td>
</tr>
<tr>
<td>FCN</td>
<td>Fully convolutional network</td>
</tr>
<tr>
<td>FDA</td>
<td>Food &amp; drug administration</td>
</tr>
<tr>
<td>FoV</td>
<td>Field of view</td>
</tr>
<tr>
<td>FPR</td>
<td>False positive rate</td>
</tr>
<tr>
<td>GB</td>
<td>Gigabytes</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphical processing unit</td>
</tr>
</tbody>
</table>
H&E: Hematoxylin and eosin
HER2: Human epidermal growth factor receptor 2
IDC: Invasive ductal carcinoma
IF: Immunofluorescence
IHC: Immunohistochemistry
ILC: Invasive lobular carcinoma
ILSVRC: ImageNet Large Scale Visual Recognition Challenge
LCIS: Lobular carcinoma *in situ*
ML: Machine learning
MLE: Maximum likelihood estimation
MLP: Multi-layer perceptron
mRMR: Minimum-redundancy-maximum-relevance
OBSP: Ontario Breast Screening Program
PCA: Principal component analysis
PR: Progesterone receptor
ReLU: Rectified linear unit
ROC: Receiver operating characteristic
SGD: Stochastic gradient descent
SVM: Support-vector machine
TDLU: Terminal ductal lobular unit
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIL</td>
<td>Tumour-infiltrating lymphocyte</td>
</tr>
<tr>
<td>TPR</td>
<td>True positive rate</td>
</tr>
<tr>
<td>tSNE</td>
<td>t-distributed stochastic neighbor embedding</td>
</tr>
<tr>
<td>WSI</td>
<td>Whole slide image</td>
</tr>
</tbody>
</table>
1 Introduction

The research objective of this thesis was to develop an algorithmic pipeline to detect and segment breast ductal carcinoma *in situ* (DCIS) regions within a whole slide image (WSI). The purpose was to highlight relevant regions to be used as part of a larger pipeline for predicting a DCIS patient’s likelihood of local recurrence post-surgery. This will inform a patient’s need for additional therapy. The pipeline leverages digital pathology and current machine learning techniques. The following introduction section will cover the necessary breast cancer pathology and machine learning background to understand the contents of the thesis.

1.1 Breast Cancer

Breast cancer is the third most commonly diagnosed cancer in Canada, accounting for 13% of all cancers. It is also the most diagnosed cancer in women as it accounts for 25% of cancers among females [1]. In 2017, an estimated 13% of cancer-related deaths were from breast cancer. The age adjusted incidence rate has greatly increased – from 2.4 to 27.7 per 100,000 women from 1981 to 2001 – due to widespread breast screening programs [2-5]. It mostly occurs in women, although in very rare cases it can be found in men.

The female breast contains a network of ducts that join and converge at the nipple, for the purpose of delivering milk to nourish our young. The ducts are connected to terminal ductal lobular units (TDLUs), or lobules, that facilitate milk production during lactation. These ductal and lobular structures are surrounded by stromal tissue composed of fibrous ligaments and adipose tissue, which support the glandular structures and give the breast its shape. Fig. 1. depicts a diagram of the breast’s inner anatomical structure [3].

Cancer describes a class of diseases characterized by abnormal uncontrolled cell reproduction, usually the result of abnormal mutations in genes meant to regulate cell growth and reproduction. The resulting lump or mass is called a tumour. A tumour can be benign, where growth does not spread, or malignant, where the growth has the potential to invade other tissues of the body. Portions of a tumour can break off and travel through the vascular or lymphatic systems to form secondary tumour sites in different parts of the body, known as metastases. There are over 100 different types of cancer, each with a variety of symptoms. Carcinomas are the most common group, defined as cancers that originate from epithelial cells. They can be categorized into *in situ*
and invasive carcinomas. *In situ* carcinomas replace epithelial cells within glandular structures and become invasive carcinomas once they break through the gland walls. In the case of breast cancer, these epithelial cells are located in the linings of the ducts and lobules within the breast, resulting in either breast DCIS or lobular carcinoma *in situ* (LCIS). If the lesion escapes its structure of origin it is classified as either invasive ductal carcinoma (IDC) or invasive lobular carcinoma (ILC).

DCIS progression is characterized by the slow buildup of abnormal epithelial cells along the inner lining of a breast duct. Fig. 2. shows the progression from normal duct to ductal hyperplasia, atypical ductal hyperplasia (ADH) and finally to DCIS. If the growth breaks through the ductal wall it is classified as IDC. While DCIS is not considered an invasive cancer, it is usually treated because of the potential for it to progress to an invasive cancer. DCIS is considered a non-obligatory precursor to IDC. There have been several studies showing a link between the two diseases, although the mechanisms are not confirmed and there remains some controversy about
this within the research community [5]. Predicting which DCIS cases will progress to invasive cancers presents a challenge and will be discussed further in the next section.

![Diagram](image)

**Figure 2.** Diagram depicting the progression of a normal duct to ductal hyperplasia, atypical ductal hyperplasia, ductal carcinoma in situ, and finally to invasive ductal carcinoma. Progression is characterized by slow buildup of abnormal epithelial cells along the inner duct lining [7]. (Image distributed under the Creative Commons Attribution License).

### 1.1.1 Breast Cancer Screening

Breast cancer is usually detected using mammography. The Ontario Breast Screening Program (OBSP) recommends women aged 50 to 74 to undergo mammographic screening every two years. Women aged 30 to 69 that are within a high-risk screening group are recommended to get an annual mammographic and breast MRI screening (or breast ultrasound if appropriate) [8]. The breast cancer 5-year survival rate has increased to approximately 87%, due to increased early detection rates as a result of these widespread screening programs [1]. During the mammography screening process, the breasts are compressed between two plates and low-dose radiation is passed through two or 3 different planes. Regions of greater tissue density have higher radiation attenuation, creating image contrast on the mammogram. Lesions which have high density appear brighter than the surrounding fatty tissue. Once a suspect lesion is identified, its status must be confirmed via biopsy, as histological analysis is considered the “gold standard” for diagnosis [10].

### 1.1.2 Histological Assessment and Classification

Histological analysis is performed on abnormal tissue removed via biopsy or surgical excision. The excised tissue blocks must be processed for analysis and storage. This involves a fixation process using a formalin solution, followed by an embedding process into paraffin wax blocks. This allows the tissue blocks to be preserved for several years [11]. Fixed tissue blocks are then sliced into very thin (approximately 5-10 µm) sections so they can be mounted and stained. The
most commonly used standard histological stain is hematoxylin and eosin (H&E), although there are a variety of alternatives. Hematoxylin dye is basic and positively charged, binding to the cell nuclei staining them blue. Eosin dye is acidic and negatively charged, binding to the extracellular matrix and cytoplasm staining them pink. Other visible structures take on combinations of these colours, while hydrophobic components such as fat cells tend to remain clear [12]. The majority of cases only require an H&E stain, but sometimes a diagnosis requires additional information that cannot be found in only an H&E image. Some of these advanced staining techniques are immunohistochemistry (IHC) and immunofluorescence (IF). IHC is applied to tissue samples while IF is applied to cell cultures. IHC and IF stains apply specific antibodies to a sample in order to image a target antigen. The antibodies are chemically conjugated to fluoresce to an antigen so they can be viewed in the sample. Fluorescence microscopy can be used to track the specific fluorophore [13].

There are several required tasks for histological analysis and classification of breast cancers. One is determining the origin and extent of cancerous regions. As previously mentioned, most breast cancers begin from epithelial cells in either the breast lobules or ductal walls. They are classified according to the type of cellular proliferation – cohesive cells are called DCIS while non-cohesive cells are called LCIS. If the tumour region extends through the myoepithelium into surrounding tissues, it is classified as either IDC or ILC. DCIS lesions must be assigned a grade based on the appearance of their cells and nuclei. Differentiation refers to the overall variability in size and shape of tumour cells. Well differentiated cells are more uniform and almost resemble normal cells, while poorly differentiated cells have an abnormal appearance and vary in their size and shape [3]. Mitotic count refers to the number of visibly reproducing cells in the image and is meant to estimate the tumour growth rate. There are several DCIS grading systems available that combine differentiation and mitotic information, attempting to capture the aggressiveness of the tumour. Most systems recognize three grades, corresponding to well differentiated (grade 1), moderately differentiated (grade 2) and poorly differentiated (grade 3). DCIS can also be classified according to gross tumour structure. Many DCIS tumours consist of large irregularly shaped masses filled with necrotic cells and tissue, known as comedonecrosis. Comedo DCIS are so called due to necrotic cell debris that oozes from excised tumours when squeezed, resembling comedonal acne. Other types of DCIS are collectively called non-comedo but can be subcategorized according to their dominant microscopic growth pattern. These include solid, cribriform, papillary and
micropapillary, although many DCIS lesions show complex combinations of these subtypes (Fig. 3.) [2-5]. Relating to previous terminology, comedo DCIS is considered poorly differentiated. Non-comedo subtypes are variable, but most are well to moderately differentiated.

1.1.3 Breast Cancer Treatment

Patients diagnosed with breast cancer have different treatment options available depending on the imaging results, pathological review and subsequent diagnosis. Patients may also undergo genetic counseling to determine their risk for hereditary breast cancer. Most patients diagnosed with DCIS undergo a lumpectomy, also known as breast-conserving surgery (BCS), to remove only the cancerous site. If negative surgical margins (all cancerous tissue removed) cannot be obtained through resection, patients should undergo a full mastectomy to remove the whole breast, which can be followed by breast reconstruction. Patients often also receive post-surgical treatment in the form of high-energy ionizing radiation directed at the tumour bed region. This destroys leftover cancer cells by damaging their DNA. Whole breast radiation therapy following BCS reduces the

Figure 3. Ductal carcinoma in situ of the breast: different nuclear grades and comedonecrosis. A - Low grade ductal carcinoma in situ, cribriform type, showing uniform cells with mild atypia. Hematoxylin and eosin, x200. B – Intermediate grade ductal carcinoma in situ, showing cells with mild a moderate atypia and focal necrosis. Hematoxylin and eosin, x200. C – High grade ductal carcinoma in situ, solid type, showing severe atypical cells, without necrosis. Hematoxylin and eosin, x400. D – High grade ductal carcinoma in situ, solid type with extensive comedonecrosis. Hematoxylin and eosin, x200 [4]. (Image distributed under the Creative Commons Attribution License).
risk of local recurrence by approximately 50%, however patients identified with very low recurrence risk may be treated by excision alone [15].

Patients with invasive disease also undergo surgery, either lumpectomy or mastectomy depending on the extent and aggressiveness of their disease, followed by radiotherapy. Invasive patients often also receive a form of chemotherapy either before or after their surgery. Chemotherapy is a treatment option utilizing cancer-killing drugs. One type of chemotherapy is hormonal therapy, which targets hormonal receptors discussed in the previous section. Different hormonal therapies have varying effectiveness depending on the molecular subtype of breast cancer being treated. Platinum based chemotherapy agents are commonly used in a variety of cancer types. In breast cancer, they are often used to treat metastatic and triple negative breast cancers. Another type of treatment is immunotherapy, which seeks to improve the body’s immune system to identify and attack cancer cells more effectively. Active immunotherapies use antigens found in a tumour to stimulate the immune system to target those cancer cells. Passive immunotherapy introduces artificial immune system components to fight the cancer cells. When chemotherapy drugs are used to treat breast cancer post-surgery, it is known as adjuvant therapy. In some cases, chemotherapy is used pre-surgery to improve surgical results and overall outcome. This process is called neoadjuvant therapy, the results of which can be informative for estimating a patient’s prognosis [15].

Recently, there have been concerns about overtreatment in DCIS patients [16,17]. Screening programs are detecting more early DCIS cases, some of which may never progress to invasive disease. Treating such patients puts them through unnecessary risk and drains healthcare resources. Thus, there is a clinical need for a method to properly stratify patients into high- and low-risk groups to address DCIS overtreatment. One method which researchers have been developing involves using an Oncotype DX score. Oncotype DX is a genomic assay that measures the expression level of certain genes, useful for predicting a patient’s response to treatment. A repurposed version of the Oncotype DX score known as the DCIS score has been tested [18-20]. The assay has been shown to be predictive of recurrence in DCIS patients. However, due to the relatively short time since it became commercially available, there have only been small cohort studies. Additionally, the Oncotype DCIS assay is still very expensive ($3416 USD) [18]. More development is needed before clinical utility is reached, and there is still a need for a recurrence risk stratification method.
1.1.4 Digital Pathology

Digital pathology refers to the creation and analysis of histological tissue images using digital scanning and computational tools. This contrasts with traditional pathological analysis, utilizing a light microscope and glass slides holding tissue specimens. Digital slide scanners were originally created as an educational tool for pathologists but have been approved for clinical use in Canada since 2013, and recently got approved for clinical use in the US by the FDA [21,22]. Slide scanners operate by optically imaging a glass slide through a microscope. The scanner pans across the slide on a robotic staging table, and stitches together the images captured at each point, creating an ultra-high-resolution image of the slide known as a whole slide image. These images also store lower resolution information in a hierarchical data format and require specialized viewing software to view. Fig. 4. shows an image of a WSI viewer as well as a diagram of a typical multi-level image storage schema. Modern pathology labs have begun the push towards a digital workflow, which will enable analytical tools developed in research to be more easily deployed in a clinical setting. This means that in addition to high-throughput slide scanners, pathology labs will be required to invest in high-performance computers, high-speed network connections and large data storage solutions, as each high-resolution WSI can take up several gigabytes (GBs) of space. This is outweighed by the many benefits of a digital workflow, including the ability to easily access and share slides, have remote consultations, markup slides with annotations and comments, use cases for further research and deploy analytical tools to the clinic [22]. Studies have shown the non-

![Figure 4. Left image shows an example of a WSI being viewed with the Pathcore Sedeen™ viewer [24]. Annotations and markups can be seen on the slide. Right image is a diagram of the hierarchical data scheme, where images are also stored at low resolution so the viewer can display a WSI with wider context [25] (Image distributed under the Creative Commons Attribution License).](image-url)
inferiority of WSIs for diagnostic purposes, as compared with state-of-the-art light microscopes [23].

Bringing digital images into the clinic will also enable the use of computer-aided diagnosis (CAD) systems as a diagnostic aid for pathologists. Many CAD systems focus on disease diagnosis from biopsy specimens. Earlier systems utilize mathematical models of structural, shape and texture features to determine a diagnosis with high accuracy [26]. Other possible applications and research areas include the use of multiple imaging modalities overlayed on the same slide, combining H&E with IHC or IF stains. Analysis of the spectral composition of images can provide insight into the chemical makeup of a sample. Stain normalization and compensation for tissue auto-fluorescence are required color processing steps for digital images. Automated grading and localization of disease can be accomplished with more advanced algorithms [10]. Finally, 3D images of histological sections can be achieved by stacking multiple slides together and reconstructing them into a volume. This is challenging due to non-linear deformations to the tissue slices caused by the fixation process [22]. With the recent advent of deep learning and availability of digital data, systems will be able to accomplish complex tasks with even higher accuracy.

1.2 Machine Learning

Machine learning (ML) refers to the use of computational algorithms to build a mathematical model for making predictions or accomplishing a task without explicit external instructions. The algorithms use prior data, known as training data, to formulate the model. Models are then validated using data unseen by the model, known as testing data [27]. Since datasets are often very large and models can be complex, iterative numerical optimizations are needed for the training process. This always centers around updating the model’s parameters to minimize a chosen loss function.

1.2.1 Supervised vs. Unsupervised Learning

There are two broad classes of ML problems, supervised learning and unsupervised learning [27]. Supervised learning involves training data paired with the desired output label. The learning algorithm attempts to learn a relationship between input training data and the desired label. The most common problem for this type of algorithm is the classification problem. This could be binary classification where the label is a single number, either a one or a zero, indicating the presence or
absence of the class in question [27]. For example, a simple binary image classification problem could involve taking input images and trying to predict whether there is cancer present in the image, with zero indicating absence of cancer and one indicating the presence of cancer. Multi-class classification is also possible, with labels in the form of a vector having the same length as the number of classes. The vector has zeroes everywhere except one position with a one indicating the class to which the sample belongs [27]. This is known as a one-hot encoding scheme. A simple example of multi-class classification could be a tissue classification system which takes an input image and attempts to determine what tissue type it contains, from a list of predetermined tissue types. Another type of supervised learning problem is regression, where the learning algorithm predicts a continuous variable rather than a discrete one [27]. In fact, classification algorithms also output a continuous variable, between zero and one, representing the probability of that sample belonging to a particular class (pseudo-probability). In binary problems a cut-off threshold must be chosen to discretize the output for classification, while in multi-class problems the class with the highest probability output is chosen. Regression problems are not limited to outputs between zero and one, and can be used for estimation of biological variables from an image. An example of this would be using a learning algorithm to estimate the cellularity (cell density) of an input image.

The second type of ML problem is known as unsupervised learning and involves learning from data without attached labels [28]. Two main tasks in unsupervised learning are dimensionality reduction and clustering. Dimensionality reduction aims to reduce a dataset to its most prominent modes of variance. Some tools for doing so are principal component analysis (PCA) and t-distributed stochastic neighbour embedding (tSNE), both of which are used for visualizing high dimensional data in lower dimensions. Clustering aims to group unlabelled data into clusters of similar samples, to learn an intrinsic classification scheme. Cluster information can be extrapolated to infer relationships between different subsets of the data [28]. Clustering can also be used for outlier detection. Each of the described ML problems has a variety of associated algorithms and techniques, however this thesis work will focus solely on supervised binary classification methods.

1.2.2 Classical Machine Learning

Classical machine learning refers to ML algorithms that came before the recent advent of deep learning [29]. A classical ML pipeline begins with feature extraction. Features are transformations
of data, designed to provide the learning algorithm with a more meaningful data representation. Features are especially important in applications where raw data is uninterpretable to a mathematical algorithm, such as text classification. Examples of such text features could be average sentence length, relative frequency of words, etc. Image classification tasks use textural, spectral and shape features with a variety of extraction methods to compact image data into values easily interpretable by a machine [30,31]. Another issue that feature extraction addresses is the curse of dimensionality. As the dimensionality of data increases, the space occupied by the data becomes sparser. This makes statistical inference more difficult, as one needs many more data samples to achieve statistical significance. Feature extraction can act as a method for dimensionality reduction, by condensing a dataset into only its most relevant features for classification. Feature selection is another important step in an ML pipeline. Often there are more features available than necessary for a classification task, so feature selection is used to eliminate irrelevant and redundant features. One common approach is minimum-redundancy-maximum-relevance (mRMR) feature selection, which is based on the mutual information between features for each class [32]. This eliminates highly correlated features, keeping those with high predictiveness for the task. Feature selection aids with dimensionality reduction by decreasing model complexity. It also helps improve generalization error by reducing overfitting. Overfitting occurs when a model is fit too closely to the training data, such that performance on an unseen test set is very poor. By removing redundancy, feature selection can reduce the effect of overfitting.

After a feature set has been chosen, an ML algorithm must be selected to fit a model to the data. The exact fitting process differs for each algorithm, but usually centers around feeding data through the model and updating parameters to minimize some loss function, which represents the difference between the model’s prediction and the true output value. One of the most basic supervised classification algorithms is logistic regression, where the output is a linear combination of the input features [27,33]. The equation defining logistic regression is:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

(1)

The $x_1 \ldots x_k$ are features calculated from the input data. $\beta_0 \ldots \beta_k$ are the linear combination coefficients that form the model’s parameters. The output is put through a sigmoid function, as seen on the left-hand side of equation (1), resulting in a pseudo-probability representing the probability of the sample belonging to a particular class. Logistic regression coefficients are
usually fit using maximum likelihood estimation (MLE), whereby the model’s parameters are iteratively optimized with respect to the likelihood function, using a numerical optimization technique such as stochastic gradient descent (SGD) [33]. It is also possible to estimate logistic parameters using iteratively reweighted least-squares, which performs a numerical least-squares minimization rather than MLE.

A more powerful classical supervised ML algorithm is the support-vector machine (SVM) [26]. SVMs attempt to find a transformation of feature space, known as a kernel function, such that the two classes can be separated by a hyperplane. More specifically, a hyperplane that maximizes distance to the nearest data point on either side, called the maximum-margin hyperplane. Fig. 5. shows a diagram of a feature space transformation and subsequent hyperplane separation of the data. There are a variety of possible kernel functions, including linear, polynomial and radial functions, each useful for different data arrangements [27]. Suffice to say SVMs are a powerful classical ML tool, useful in a variety of situations.

![Diagram of a dataset being transformed in feature space by some kernel function \( \phi \), such that a hyperplane can be found to separate the two classes. The plane represents the maximum-margin hyperplane generated by an SVM algorithm, having the largest separation between the two classes [34]. (Image distributed under the Creative Commons Attribution License)](image)

1.2.3 Modern Machine Learning - Neural Networks & Deep Learning

Machine learning in the last few years has been revolutionized by the popularity of deep learning [35,36]. Deep learning algorithms are a type of machine learning algorithm making use of artificial neural networks (ANNs). ANNs were first developed in the 1960s and 1970s and were originally...
called feedforward multi-layer perceptrons (MLPs). Described as universal function approximators, their design takes inspiration from sequential neuronal and synaptic connections in the brain [37]. The subsequent development of the backpropagation algorithm enabled ANNs to learn parameters from data using numerical optimization [38]. While this worked, ANNs were unable to outperform other popular ML algorithms of the time, such as SVMs, due to prohibitively long training times and relative lack of available data. That is, until Dr. Geoffrey Hinton and his research group famously won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 with a convolutional neural network (CNN), an ANN specifically designed for image data input [39]. One driver of this resurgence is the abundance of data available now, which ANNs require to achieve good performance. Second, the availability of high-performance computing power has greatly increased, particularly using graphical processing units (GPUs). GPUs are well suited to the large matrix calculations required for deep learning and have seen a reduction in cost as the technology has improved [35].

Neural networks are composed of neurons organized into stacked layers (Fig 6.). The value of each neuron is a linear combination of neurons from the preceding layer. During forward propagation, the output of each neuron is transformed by an activation function before being fed to the next layer. Essentially, each neuron is performing a logistic regression as described in equation (1),
although the activation functions may differ. The parameters of each neuron correspond to the $\beta_0...\beta_k$ coefficients in equation (1). The output layer also uses an activation function, usually a sigmoid, to transform the output into a pseudo-probability representing the probability that the input sample belongs to a certain class. Training the ANN’s parameters is achieved through the backpropagation algorithm, which leverages the chain rule for derivatives. The gradient of each neuron can be calculated from the gradient of its output neuron, which can be calculated from the gradient that neuron’s output, etc. This results in an iterative, recursive algorithm for calculating the gradient of each neuron, starting from the final output value. However, it requires that the derivative function for each neuron’s activation be known beforehand, so it can be computed using automatic differentiation. Each training sample is fed forward through the network, and the error between the predicted and true output is measured using a loss function. Backpropagation is used to calculate the error term gradient, which is fed backwards through the network. Finally, a numerical optimization such as SGD is used to update each weight value accordingly [35,36]. This process allows ANNs to iteratively learn from data. It is important to note that data fed into the ANNs does not need to be a feature space representation, as is the case with classical ML. Inputs are simply raw data (usually with some preprocessing). Since ANNs are universal approximators, they can learn a feature space representation optimized for the specific task at hand. This is what makes ANNs so advantageous over classical ML algorithms. Classical ML requires in depth domain knowledge and time to engineer a rich and informative feature set for each classification problem, whereas an ANN can bypass that step by automatically learning a feature representation.

Figure 7. Example of a simple CNN, with two convolution layers, two pooling layers and finally some fully connected layers leading to an output prediction. This network is classifying natural images. The probability outputs on the left indicate the CNN has correctly identified the input image as a boat [41]. (Image distributed under the Creative Commons Attribution License)
1.2.4 Convolutional Neural Networks

CNNs are an extension of ANNs, designed specifically to handle images as input (Fig. 7). They take advantage of the spatial arrangement of image data, and use weight sharing to reduce the computational load [35,36]. CNNs center around the convolution operation, in which a kernel strides across an image in steps. At each step a linear combination is computed between the kernel’s weights and the corresponding image pixel values, followed by an activation function. The output at each location is joined together to produce an output feature map. The parameters are confined to the convolution kernel and shared across the image, rather than having a parameter per pixel. This greatly reduces the memory and computational time required to operate a CNN. After a convolution there is typically a pooling layer. Max-pooling is the most common pooling operation, in which only the maximum value in a 2x2 region is passed to the next layer. This condenses the feature map, preserving the most activated features. Many convolution and pooling layers can be stacked to form the structure of a CNN (seen in Fig. 7.), after which the output is fed into fully connected layers (ANN layers as previously described) for classification.

Most final output layers use the sigmoid activation function, which is defined as:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$  \hspace{1cm} (2)

The sigmoid function sets the output between zero and one to represent a pseudo-probability, and so is useful for the final classification layer. However, there are many other activation functions available. One of the most commonly used activations in CNNs is the rectified linear unit (ReLU), defined as:

$$R(x) = \max (0, x)$$  \hspace{1cm} (3)

This function simply sets all negative values to zero. This function violates the need for activations to be differentiable as the ReLU has a discontinuity at zero. However, in practice no neurons will sum to identically zero, so it is not an issue. Another common activation is the tanh function, which is similar to the sigmoid except it sets the output between -1 and 1. Finally, the exponential linear unit (ELU) is similar to a ReLU, but with a smooth transition replacing the discontinuity [42].

The previous sections have mentioned loss functions several times. A loss function is used to estimate the error between a training sample’s output value and the expected value (ground truth).
The error is used by the optimizer to update the ANN’s parameters accordingly. A common loss function in binary classification is binary cross-entropy, defined as:

$$CE = -y \log(p) - (1 - y) \log(1 - p)$$  \hspace{1cm} (4)

$y$ is the ground truth label, either zero or one. $p$ is the score output from the classifier, between zero and one. Cross-entropy sums the logarithmic probabilities for each of the two classes. The optimizer’s goal is to minimize this cross-entropy term, resulting in a large separation between the two classes.

Another term mentioned previously is optimization, which is the process of finding a parameter set which minimizes the loss function. ANNs will not have a closed-form solution for their parameter values, so iterative numerical optimization techniques must be used. One such technique commonly used in deep learning is SGD [35]. SGD feeds data through the network in small batches, calculates the gradient on the error of a batch and uses it to update the weights. This is repeated for the entire training dataset (a single pass through the dataset is known as an epoch), after which the data is shuffled, and the entire process is repeated with new batches. Several epochs are typically performed before the optimization algorithm converges. The learning rate is usually decreased at higher epochs to assist with convergence. Many extensions have been developed for SGD. A simple but effective one is the addition of a momentum term that depends on the previous update, which reduces oscillations and helps the algorithm reach convergence faster. Another popular extension to SGD is Adaptive Moment Estimation (Adam) [43]. Adam uses the running average of both the gradients and the second moments of the gradients to calculate parameter updates. It has proven to be very fast and effective for training CNNs.

One problem ANNs have in general is that they very easily overfit to a dataset if trained for too long. Many techniques have been developed to increase their generalizability. One is known as dropout, which involves randomly turning off some neuron connections during each training step [44]. The fraction of connections switched off is controlled by a hyperparameter called the dropout rate. This forces ANNs to learn more robust features that are not dependent on one another, thus increasing generalizability. Another common technique is batch normalization, which normalizes all hidden layers in an ANN for each batch during training [45]. For each batch output from a hidden layer, the batch mean is subtracted out and the batch standard deviation is divided out. The
normalized output is then fed to the next layer. This reduces internal covariate shift, a phenomenon in which parameters increase or decrease too quickly when the learning rate is too high, preventing the ANN from converging. Keeping the hidden layers normalized allows one to train with greater learning rates. Batch normalization also acts as a regularization technique by placing a constraint on features that the network can learn, thus increasing generalizability. Finally, data augmentation is a technique CNNs can leverage to artificially increase the amount of training data available [46]. The convolutional features a CNN learns are spatially dependent, so if the input image is spatially altered the CNN will think it is receiving a new sample. This property can be exploited by applying random transformations to the images in a training set, such as translations, rotations and flips. Some applications can even use image stretching and brightness alterations. The CNN considers each permutation as a new input and will learn to recognize a wider variety of images, which helps it increase generalizability. The limitation is that all image modifications must result in a valid image for the application, in order to keep augmented data consistent with real data.

1.2.5 Model Evaluation

As mentioned previously, data to be used for a machine learning problem is always split into training and testing subsets. In deep learning, a third subset called the validation set is required to monitor a model’s performance as it trains, and to determine when a model has converged and is

![Figure 8. Example of an ROC curve comparing two different models. Test B is superior because the curve is closer to the top left corner, which represents optimal performance. The yellow chance line depicts the curve a model would have if it randomly guessed a class [47]. (Image distributed under the Creative Commons Attribution License)]](image.png)
sufficiently trained. After training, metrics can be calculated on the test set to evaluate a model’s performance. For binary classification tasks, it is common to construct a receiver operating characteristic (ROC) curve [47,48]. The model must be run on all test data to produce prediction probabilities. The cutoff threshold determining what probability level is positive vs negative must be incrementally increased. For each threshold value, the true positive rate (TPR) and false positive rate (FPR) are calculated and plotted. TPR is defined as:

\[ TPR = \frac{TP}{TP+FN} \]  

where TP is the number of true positives, and FN is the number of false negatives. FPR is:

\[ FPR = \frac{FP}{FP+TN} \]  

where FP is the number of false positives, and TN is the number of true negatives. Plotting these values against each other results in a curve representing the model’s performance trade-off at different thresholds, an example of which can be seen in Fig. 8. Curves closer to the top left corner are superior, as that corner represents perfect performance (highest TPR, lowest FPR). A straight line drawn from bottom left to top right corner represents classification models that randomly guess a class, which is the lowest possible performance. A useful metric to calculate from an ROC curve is the area under the curve (AUC). Models having higher AUC are closer to the top left, and are thus superior [47,48]. The random chance line will produce an AUC of 0.5, which is considered the lowest possible AUC value. AUC represents the probability that the classifier will rank a random positive sample higher than a random negative sample, making it indicative of overall classification performance.

A useful technique used to evaluate ML models is cross-validation [49,50]. There are several types of cross-validation, but this thesis will only use k-fold cross-validation. In k-fold cross-validation, the data is randomly partitioned into k equal subsets, called folds. One of these folds is held out as the test set, while the other k-1 folds are used to train a model. Evaluation metrics are then calculated using the test fold. This process is repeated k times, with each fold being held out for testing once. The results can be averaged across folds to produce a single testing quantification. Stratified k-fold cross-validation ensures that each fold contains an approximately equal ratio of
positive to negative labelled samples. It is common to use 5- or 10-fold cross-validation, but in general k is a flexible parameter [49,50].

1.3 Machine Learning in Digital Pathology

Given the recent advances detailed above in both digital pathology imaging and machine learning, these two fields are primed to intersect. This section will cover some recent machine learning methods and techniques being applied in digital pathology, as well as problems faced in the field [51]. As mentioned in the previous section, digital pathology WSIs are extremely large, taking up several GB of storage space each. This presents a challenge as one cannot process an entire WSI in memory. ML methods must divide images into patches for analysis, and devise techniques for aggregating patch level outputs into a slide level decision. This issue extends to pathological image labels as well. Because images are so large, it is challenging to gather detailed labels on WSIs. Most datasets only have slide level labels for the diagnosis, so patch analysis must be performed with only global labels. Large datasets with detailed region delineation exist but are more difficult to curate [52]. One advantage of patch extraction is that histology patches are rotation invariant, as changing the orientation of a patch does not alter its biological interpretation. Thus, patches can be rotated and flipped freely, allowing for a more extensive augmentation process.

A prominent task being addressed by ML in healthcare is CAD, where an automated diagnosis system can act as an aide for pathologists in the clinic. This type of task seeks to map one or more WSIs to a disease category or identify the severity of disease (automated grading or scoring). The overall goal of such CAD systems is to aid with diagnosis by reducing variability introduced by human pathologist observers [52]. Previously, classical machine learning methods have been used for histopathological image classification. This includes traditional textural and spectral feature extraction, followed by disease classification with a classical ML algorithm such as an SVM. Classification can also be done between disease grades. These methods focus on patch level classification rather than WSIs [53,54]. While results are good, drawbacks include the amount of time needed to engineer a set of hand-crafted features, which may not be generalizable to other datasets, as well as the lack of available data at the time. Since then, ANN and CNN based methods have overtaken classical ML performance on histopathological classification tasks [52,55]. As previously mentioned, these methods learn an optimal feature representation, making it much easier to achieve state-of-the-art performance. More recent methods have developed techniques to
incorporate imaging features gathered at high and low resolutions to perform more robust disease classifications [56].

Another common ML task related to diagnosis is segmentation, which seeks to delineate diseased regions within an image. This task maps an input image to an output image, where each pixel in the output image represents a disease or histological tissue type classification. As mentioned previously, segmentation labels are much more difficult to curate than classification, but aggregating patch segmentation results presents less of a challenge. Again, classical ML methods have been developed using colour and texture features and classical classification algorithms such as SVMs. One method divides a WSI into superpixels, which are small patches divided along natural image boundaries. These patches can be classified and stitched together to form a segmented image [57,58]. This is a sliding window approach to segmentation, where a decision is made at patch level and stitched together to form a segmented image. Since then, CNN based methods have also been developed to perform segmentation at pixel level, with a classification decision being made at each pixel in the output image. This type of network is known as a fully convolutional network (FCN), in reference to the replacement of densely connected layers with a convolutional output layer [59]. The final output is a probability map with the same dimensions as the input image and can be used for segmentation. The CNN segmentation revolution came with the U-Net, which builds on the FCN architecture by adding skip connections to preserve spatially relevant information for performing the final segmentation [60]. This network structure will be discussed further in Chapter 2.

A challenge associated with ML in digital pathology is the large difference in information content between different available resolutions. Cellular morphological and textural information is captured at high power, whereas larger scale glandular and structural information is captured at low power [51]. Different magnifications are best suited for different ML tasks, and there is no overall agreement over what magnification is best [61]. Some researchers have developed methods of combining this information to improve performance [62,63]. This technique shows promise for tasks requiring features extracted from both high and low power image fields. Another challenge for ML in digital pathology is the colour and stain variation that can occur across different tissue processing protocols, scanners and pathology labs. Models trained on data from one institution may not generalize to another. Some methods have been developed to address this issue, the most prominent of which are colour normalization and stain augmentation. Colour normalization aims
to map the colour distribution of a source image to a reference image, thereby reducing variations due to stain colour in the source image [64,65]. However, this has the potential to introduce other errors as it relies on the accuracy and robustness of the stain mapping method. Stain augmentation is a type of data augmentation which adds slight random alterations to the hue, saturation and brightness of images during the training process [65,66]. This technique allows a CNN to learn a wider colour space, thus increasing generalizability to images from a different domain. However, this depends on the augmentation’s ability to include meaningful colour alterations (without fundamentally changing the appearance of a feature) and cover all possible variability. Either technique can be used to increase robustness of machine learning models in digital pathology but comes as a cost.

Finally, machine learning is a tool that can be used to uncover new clinical-pathological relationships. There is a vast amount of pathology, CT, MR and genomic data being made available to researchers, analysis of which may uncover new clinical relationships [51]. However, these datasets are far too vast for clinical researchers to observe and document relationships. This is where machine learning can be used to find associations between different modalities within large datasets. For example, it was recently uncovered that morphological features taken from stromal regions have an association with recurrence risk and survival, through a learned prognostic model [67,68]. This relationship was discovered using classical feature extraction and a logistic regression model and has yet to be recreated with an ANN based method. This thesis will explore preliminary methods for predicting recurrence risk from WSI data using a CNN based model.

1.4 Thesis Outline

This introduction section has explored the typical workflow used for breast cancer diagnosis and treatment, while outlining some clinical needs in the field. I have also explored technological developments in digital pathology and machine learning, as well as their intersection. The next chapter will present in detail the machine learning pipeline constructed to segment DCIS regions within a WSI. This will include a comparison of models developed at different resolutions, as well as multi-resolution models leveraging multiple image fields. Finally, statistical comparisons are made between the proposed segmentation models and expert pathologist segmentations. Chapter 3 describes the future development and next steps for the pipeline. It will also establish connections with the larger project of predicting recurrence risk, and the research field in general.
The work described in Chapter 2 of this thesis was presented at the European Congress on Digital Pathology (ECDP) in April 2019. The conference paper is available online [69].
2 Automated Localization of DCIS in WSIs

2.1 Introduction

DCIS is a common early form of non-invasive breast cancer, accounting for approximately 2,500 new cases per year in Canada [1]. Most DCIS patients undergo BCS to remove the lesion, and many receive post-surgery radiotherapy to reduce the risk of developing local recurrences. Since prediction of absolute risk of recurrence based on traditional histopathologic evaluation is limited, it is not possible at present to identify patients with very low risk of recurrence in which radiotherapy can be omitted. Radiotherapy is also an aggressive form of treatment and there have been concerns about overtreatment of early forms of breast cancer, such as DCIS [16,17]. Improved stratification of patients into low- and high-risk recurrence groups would be of great benefit for a guidance-based treatment approach. Histopathologic evaluation of excised tissue is an important step for planning additional treatment and understanding the underlying biology of tumours which can help to some extent identify recurrence risk. As digital slides are becoming more accessible in practice, automation can be adopted to analyze large datasets of archived tissue. Previous work has shown that quantitative features extracted from digital pathology images containing DCIS may improve prognostication [67,68]. Typical digitized whole slide images are extremely large (approximately 3-8 GB each) and have information stored at several different magnifications. As such, it is impractical to extract quantitative features from an entire WSI due to computational and computer memory constraints. To efficiently extract relevant information, scalable methods are needed to accurately localize DCIS regions in a WSI. This is a non-trivial problem due to high intra-class variability within DCIS and comparatively low inter-class variability between DCIS and normal ducts, as illustrated in Fig. 9.

2.1.1 Related Work

This is not a new problem, and many previous methods have approached this using patch classification. As mentioned in the previous chapter, classical ML methods include using image processing, feature extraction and a classical ML algorithm like SVMs to classify histology patches as benign or malignant [53,54,58]. These methods often show excellent performance on small datasets, but the engineered features have trouble generalizing to larger datasets. Advancements in deep learning make it possible to leverage large datasets of WSIs to overcome
these issues and recent methods have shown promise utilizing deep learning techniques for patch classification [55, 56]. Patch classification allows for coarse WSI segmentation by tiling the classification results. Finer segmentations can be achieved at the pixel level by using an FCN style CNN. In this thesis, I describe a fully automated pipeline for segmentation of ducts containing DCIS which encompasses CNNs trained on patches extracted from WSIs.

2.2 Materials & Methods

2.2.1 Dataset

Our dataset consists of 202 women who were diagnosed with DCIS between 2012/2013 and underwent BCS. Excised tissue specimens were handled per routine tissue processing procedures to produce formalin-fixed paraffin embedded tissue blocks and H&E stained sections. Representative sections were imaged in an Aperio digital slide scanner at 20x resolution. An expert pathologist reviewed each WSI and marked annotations via a pen tool around the DCIS ducts using the Pathcore Sedeen™ viewer [24]. These annotations served as ground truth segmentations for training CNNs. WSIs were split into subsets for training (n=111), hyperparameter tuning (n=72) and testing (n=19). An additional 10% of the training set was held out for use as a validation set, to monitor the CNN for early stopping.

Figure 9. Example images of ducts taken from various WSIs. Panel a) depicts a normal duct without DCIS. Panels b) through d) depict ductal regions containing DCIS, with the blue outlines depicting a pathologist’s manual annotations of the lesions.
2.2.2 Training

Due to computational costs associated with reading WSIs into memory, the data must be subdivided into patches for training. To account for high variability within our dataset, each patch was augmented via random rotations at either 0, 90, 180 or 270 degrees, thus increasing the effective size of the training set. The CNNs are trained using the mean pixelwise cross-entropy loss function and Adam optimizer [43]. Training runs for 100 epochs each, and the model with the lowest validation loss from the final 10 epochs is selected as the final trained model. This training scheme allows each network sufficient training time, whilst ensuring they do not finish with a stochastically unfavorable update.

![Figure 10. Example patches from each of the three training datasets. The blue squares show the FoV of the patch one resolution step up within the context of the lower resolution patch. Patch sizes are 128µm (20x), 512µm (10x), 2048µm (5x).](image)

2.2.3 Resolution vs. Field of View

Our WSIs are formatted such that the following resolutions were available: 5x, 10x and 20x. Using a constant patch size of 256x256, patches extracted at higher resolutions will have a narrower field of view (FoV). Here we set out to determine which resolution to FoV ratio is most informative for identifying DCIS. Example patches from these different resolution patch sets can be seen in Fig. 10, and Table 1 describes the dimensions and pixel spacing of each patch set. While there are many CNNs available in the literature, we opted to train a U-Net [60], as they have proven to be effective for segmentation tasks whilst learning complex features in a fully-automated manner. The U-Net is an FCN with feature map concatenation between the downsampling and subsequent upsampling arm, which preserves spatially relevant information for segmentation. I made modifications to the original U-Net architecture to use ELU activation functions and batch normalization on all
convolutional layers, as both have been shown to train networks faster [42,45]. We use padded convolutions to ensure that the output segmentation maps have the same dimensions as the input image.

Table 1. Patch resolution and FoV from each data paradigm to be tested.

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Pixel Spacing (µm/px)</th>
<th>Patch FoV (µm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20x</td>
<td>0.5</td>
<td>128</td>
</tr>
<tr>
<td>10x</td>
<td>2</td>
<td>512</td>
</tr>
<tr>
<td>5x</td>
<td>8</td>
<td>2048</td>
</tr>
</tbody>
</table>

2.2.4 Multi-resolution Network

In order to overcome the resolution vs. FoV trade off, we also designed a multi-resolution network. Multi-resolution networks have previously been used in digital pathology for nuclei segmentation [63]. Instead of feeding only one FoV at a time, all three FoVs (e.g. all three patches in Fig. 10.) are treated as a single input sample, giving the network access to both high resolution details and wider contextual information. Two architectures were designed to handle these inputs. One in which three patches are concatenated into a single nine channel image (9ch) and fed through the previously described U-Net architecture. The second is a custom architecture (Fig. 11) that splits the U-Net down-sampling arm into three convolutional branches – one for each input FoV – before recombining via concatenation and feeding into the up-sampling arm (3rm). The label images used to train these networks correspond to the high-resolution input images.

2.2.5 Ensemble Models

Ensemble models are commonly used in machine learning and have been shown to help improve model performance. The core idea is that a set of weak learners can be combined in some way to create a strong one. To this end, a set of 5x resolution U-Nets were trained using a 10-fold cross-validation scheme. 10 networks are trained, each with 9 folds of data and the last fold held out for validation. Each trained model is run individually on a WSI, and the resulting output probability maps are averaged across the models. The averaged probability maps are used as the ensemble model predictions.
2.2.6 Evaluation

Patches were extracted from each WSI at the appropriate resolution to create segmentation maps during test time. They were fed sequentially through each CNN, and the outputs were stitched together to create an output probability image. These probability maps were thresholded to create binary DCIS segmentation masks. The ground truth masks were downsampled to the appropriate size for evaluating the segmentation maps. We used the following metrics to validate the performance of each CNN.

2.2.6.1 Dice Coefficient

The dice similarity coefficient is a measure of positive overlap between two binary images. It is defined as:

\[
\text{DSC} = \frac{2 \times TP}{2 \times TP + FP + FN}
\]  

where TP is the number of true positive pixels in the images, FP is the number of false positive pixels, and FN is the number of false negatives. It is a similarity measure ranging from zero to one.
– one meaning the two images are identical. The dice coefficient is a useful metric for WSI labels because the dice coefficient does not depend on true negative values. This means that, unlike other metrics such as binary accuracy, the dice coefficient is not inflated by the high number of background pixels.

2.2.6.2 Modified F1 Score

The F1 score is the harmonic mean of precision and recall, defined as:

\[
F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]  

\[
\text{precision} = \frac{TP}{TP + FP}
\]

\[
\text{recall} = \frac{TP}{TP + FN} = TPR
\]

Recall has the same definition as TPR from equation (5). The F1 score and dice coefficient are mathematically identical. However, for the F1 score, precision and recall are calculated at the object (duct) level from a binarized DCIS mask, whereas the dice coefficient is computed at the pixel level. Typical object detection metrics favour a 1-1 ratio of predicted regions to ground truth regions. Our expert labels often group collections of ducts together as one region, whereas the networks often segment each of those ducts individually. We still want those individually identified ducts to be counted correctly, so we modified precision and recall to avoid penalizing multiple predicted regions within a single ground truth region. The downside is that the F1 score is no longer symmetric; swapping the predicted and ground truth images will result in a different value for the F1 score. Henceforth, all mentions of the F1 score refer to this modified definition of the metric.

2.2.7 Post-processing

A simple post-processing pipeline was created to refine the output segmentations, which come out grainy, non-smooth and contain grid tiling artifacts. The pipeline consists of a thresholding operation on the probability maps followed by a morphological opening (removes small positive grains) and finally a morphological closing (fills small negative gaps). The morphological operations use a disk-shaped structuring element to smooth the segmentation images. Each step in the pipeline is illustrated in Fig. 12. The threshold and morphological disk radius are the
parameters for the post-processing pipeline, tuned sequentially using metrics calculated on the tuning set. The threshold maximizing the dice coefficient is found first. This threshold is then fixed to find the morphological disk radius maximizing the F1 score.

2.2.8 Random Parameter Search

A random search paradigm [70] was used to tune U-Net hyperparameters, whereby the parameter space is randomly sampled and used to train the model a set number of times. This allows us to set a search budget independent of the number of parameters and possible values. Also, the search efficiency does not decrease by adding extra parameters that do not affect the performance. Some preliminary tests were performed to identify valid ranges for each parameter being included. The search ran 100 times, each with a set of hyperparameters randomly sampled from their valid ranges. Each model was trained with a 5-fold cross-validation. A list of parameters tested can be seen in Table 2.
2.2.9 Annotator Similarity

In order to gauge the effectiveness of these automated systems, it is useful to compare to human level performance. A second expert pathologist was asked to independently annotate all the DCIS on the held-out test set of 19 WSIs. The dice coefficient and F1 score were measured between pathologists to assess inter-observer variability in our dataset. Fig. 13 shows an example of such paired independent annotations. This image illustrates the significant variability between observers. These annotator similarity comparisons will serve as a benchmark for the described automated techniques.
2.2.10 Model Uncertainty

Another important analysis of this machine learning system involves quantifying uncertainty and identifying situations in which the model is unlikely to make a correct prediction. A prominent method of measuring model uncertainty, known as Monte Carlo dropout [71], involves turning on neural network dropout [44] during test time and repeatedly running the sample to acquire a distribution of output values. Our dropout rate was set to 0.5 and each sample was run 100 times. The standard deviation of these outputs was calculated to obtain a measure of network uncertainty for the sample. This was performed on all patches from WSIs in the test set to create uncertainty maps that complement the WSI output probability maps. These uncertainty maps provide insight into histological structures and image regions that the models struggle with and are important for flagging difficult cases for further analysis or future work.

2.3 Experimental Results

2.3.1 Resolution and Architecture Evaluation

Test results for each resolution architecture, including their post-processing pipelines, can be seen in Table 3. Each network was run on the independent test set used to compare annotators and evaluated against the annotator who created the original training labels. The 5x model achieves the greatest dice coefficient and F1 score of the single resolution networks. This implies that the segmentation quality benefits from patches with a greater FoV, as the network can make use of the wider contextual information. This is consistent with the way that pathologists determine which

Figure 13. Example of a group of DCIS ducts annotated independently by two different pathologists. They annotated images in two categories; definite DCIS (blue) and probable DCIS (yellow). The remainder of each image would be considered not DCIS.
regions contain DCIS by observing the WSI at low resolution, using higher resolutions to fine tune their decisions. The 5x U-Net also benefits from reduced training and run times, due to the reduced WSI size at this resolution. The 3rm multi-resolution architecture had the best F1 score of all the U-Net architectures, and a dice coefficient comparable to the 5x U-Net. Thus, the 3rm network was successfully able to combine information from high and low resolution to increase the segmentation accuracy and duct detecting capabilities of the model. However, this comes at the cost of much longer processing times. An ROC analysis, as seen in Fig. 14, further supports the conclusion of the 5x model performing best, with an AUC of 0.987. The ensemble model produced a nearly identical ROC curve and dice score to the 5x model, but also has increased processing time. The 3rm achieves a slightly worse AUC than the 5x and ensemble models, though it is still an improvement over all the other models.

Table 3. Test set (n=19 slides) results for each architecture run on the test set. P-values are from Wilcoxon sign-rank tests comparing each model to the second annotator. Asterisks indicate statistical significance. Time is average run time per WSI.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dice Score</th>
<th>Dice P-Value</th>
<th>F1 Score</th>
<th>F1 P-Value</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second Annotator</td>
<td>0.73</td>
<td>N/A</td>
<td>0.43</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>5x</td>
<td>0.77</td>
<td>0.31</td>
<td>0.60</td>
<td>0.08</td>
<td>6.35</td>
</tr>
<tr>
<td>10x</td>
<td>0.56</td>
<td>0.01*</td>
<td>0.48</td>
<td>0.57</td>
<td>93.16</td>
</tr>
<tr>
<td>20x</td>
<td>0.62</td>
<td>0.18</td>
<td>0.55</td>
<td>0.15</td>
<td>2247.89</td>
</tr>
<tr>
<td>3rm</td>
<td>0.76</td>
<td>0.40</td>
<td>0.68</td>
<td>0.01*</td>
<td>3065.07</td>
</tr>
<tr>
<td>9ch</td>
<td>0.69</td>
<td>0.55</td>
<td>0.66</td>
<td>0.02*</td>
<td>3177.02</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.77</td>
<td>0.21</td>
<td>0.64</td>
<td>0.03*</td>
<td>331.45</td>
</tr>
</tbody>
</table>
Figure 14. ROC curves for each model, run on the testing set. ROC analysis is performed at the pixel level for each slide. The ROC analysis requires the raw probability values, and thus does not include the post-processing pipeline for each model.

Figure 15. Models resulting from the random parameter search on the 5x U-Net, sorted by mean test AUC. Blue represents top set of models showing no difference from the Kruskal-Wallis test. Orange is all other models. The green x represents the model with the baseline hyperparameter set.
2.3.2 Comparison with Second Annotator

A Wilcoxon sign-rank test was performed to compare each models’ results to the annotator similarity metrics. When comparing dice coefficients, only the 10x network was found to be significantly different (worse) than the second annotator. No other networks show statistical differences from the second annotator, implying they have comparable performance. When comparing the F1 scores, both multi-resolution models and the ensemble model showed statistically significant improvement over the second annotator. Fig. 16. shows boxplots of the dice and F1 scores for each model. Both the 5x and 3rm models show reduced dice score variance compared to the second annotator. The 3rm model greatly reduces the F1 score variance compared to the second annotator. It is important to note the small sample size of dually annotated WSIs in this test set. More labelled test data would be helpful in teasing out the performance differences between each model.

Figure 16. Box plots of test set (n=19) results from each model, including the second annotator. Dice score is shown on the left while F1 score is shown on the right. Orange solid lines indicate the median of each set, while green dotted lines indicate the mean.
Figure 17. Example WSI from the test set. Cyan outlines indicate a pathologist labelled DCIS region. The colour overlay represents the probability map output by the CNN. Red indicates high probability and blue is low – zero values are set to be transparent. This image was segmented using the ensemble model described above.

Figure 18. Uncertainty map created at 5x on the same WSI sample as shown in Fig. 17. U-Net dropout was set to 0.5, each patch was run 100 times and the standard deviation of probability values was calculated and shown as a colour overlay. Cyan outlines represent pathologist DCIS labels.
2.3.3 Hyperparameter Search Results

The 5x U-Net was chosen for the random parameter search due to its superior performance and drastically reduced run time. The randomly generated models were each run and evaluated on the tuning set (Table 3). The AUC was calculated and averaged across all folds for each model. A series of iterative Kruskal-Wallis tests was performed on the model set. At each iteration, the test is performed. If it is significant (p < 0.05), the lowest mean AUC model is removed from the set. This process is repeated until the test is not significant, resulting in a set of the top models that are not significantly different from one another. The results of this analysis can be seen in Fig. 15. The set of top models includes the base model with original hyperparameters (Table 2). Most models produced by the search showed worse performance than the base model. Some models (n=19) are shown to have comparable performance to the base model, as measured by the Kruskal-Wallis test. This implies that the U-Net architecture functions well for this problem as designed, and small hyperparameter alterations have minimal effects on the performance. Thus, the baseline hyperparameters were chosen to move forward.

2.3.4 Sample Output

Fig. 17. shows an example of predictions generated by the ensemble U-Net on an entire WSI from the test set, compared to annotations from one of the pathologists. The predicted segmentation regions are generally very close to the annotation borders, with large drops in probability marking the edges of predicted DCIS regions. There are some instances in Fig. 17. where the U-Net prediction overlaps the DCIS, but the label surrounds a duct containing empty space, resulting in an erroneous prediction. Fig. 18. shows an uncertainty map created using Monte Carlo dropout, overlaid on the same WSI as in Fig 17. Most ducts are filled with low uncertainty regions while the boundaries show slightly higher uncertainty. High uncertainty regions also seem to be correlated with incorrectly predicted regions and regions with midrange probability, implying that the model uncertainty and U-Net probability outputs can be complementary. Further tests need to be performed to quantify the predictiveness of the model uncertainty analysis.
2.4 Discussion

For the intended purpose of this pipeline, we prefer the model to be sensitive to DCIS boundaries for subsequent feature extraction in stromal regions surrounding the DCIS. This is achieved by optimizing the pipeline to have the best dice coefficient. Optimizing for the best F1 score would yield an algorithm suited to correctly identifying all DCIS ducts, while allowing for errors on the boundaries. Furthermore, we prefer an algorithm with high specificity over sensitivity - we can accept missing some DCIS ducts in favour of eliminating all normal ducts. This is not a diagnosis problem; we know there is disease present in the image and need to localize it. Thus, subsequent feature extraction should only occur within and surrounding DCIS regions, ignoring irrelevant normal ducts. The model uncertainty analysis could be included in the pipeline to flag images with high uncertainty to be checked by a pathologist. With more refinement, our pipeline could be implemented as a processing step on slide scanners to identify DCIS regions immediately as slides are processed, aiding pathologists in their analysis. A previously mentioned drawback of this analysis is the small set of dually annotated test data, making it more difficult to tease out the differences between models. Another limitation of the pipeline is the lack of IDC data. While actual analysis of IDC is beyond the scope of this project, it would be valuable to include a method to recognize and flag regions of IDC to increase the clinical utility of the pipeline.
3 Discussion & Future Work

3.1 Discussion Review

Here we presented a series of U-Net architectures to solve the problem of localizing DCIS on WSIs automatically. It was found that when using a traditional U-Net architecture, there are both speed and accuracy benefits to training and running the model using low resolution patches. This is due to the patches having a greater FoV, giving the U-Net access to wider contextual information. It is evident that in DCIS segmentation, wider contextual information is more valuable to the U-Net than higher resolution information. Two novel multi-resolution architectures were presented that combined patches from all three available resolutions to overcome the resolution vs. FoV trade-off. However, this comes at the expense of greatly increased training and running times for the networks. Finally, a comparison was made between two independent pathologist annotators, showing that the best networks have performance comparable to the second annotator on the segmentation task. A noted limitation of this work is the relatively small dually annotated test set, limiting our ability to properly quantify differences between the models.

3.2 Future Work

Future refinements to the pipeline should include a method to deal with stain variations, either through stain normalization or stain augmentation [64-66]. As previously mentioned, stain normalization aims to map any stain colour variations to a reference stain, reducing the possibility of misclassification due to stain variations. However this technique depends on the efficacy of the stain normalization method, which has the possibility of introducing new stain errors. Stain augmentation introduces small random variations to the stain colour, teaching the neural network to adapt to stain variability. However this depends on the augmentation’s ability to produce meaningful colour variations without fundamentally changing the features in an image. Either technique would help increase the algorithm’s robustness to different scanners and staining protocols from other institutions, but careful consideration must be given to the potential drawbacks.

There are many types of data augmentation in addition to stain augmentation, which could be further explored to increase the pipeline’s robustness [46]. Our augmentation scheme only uses right angle rotations, but much more is possible with random rotations, crops, flips and brightness...
adjustments. Random gaussian blur filters can also be used to simulate out of focus regions on a slide.

Several different loss functions were tested over the course of the project, but only the binary cross-entropy loss proved effective. Some recent research suggests combining losses from different functions can be more effective, and there are methods for aggregating a set of different losses according to network uncertainty [72,73]. This could be applied to combine cross-entropy and dice coefficient loss functions into a more robust segmentation loss. Additionally, boundary distance-based loss functions are becoming popular in segmentation, which could be applied either solely or as part of a combined loss [74].

Finally, there are always new neural network designs being developed, including for segmentation. The U-Net is often considered the best general-purpose segmentation network, as the 2018 Medical Segmentation Decathlon grand challenge was won by a U-Net with a modified boundary-based loss function [75]. However, other new network designs may be better suited to this particular task. Mask R-CNN is a region proposal segmentation network that, although is difficult to train, is showing promising results and may be worth implementing [76]. Suffice to say there will always be new machine learning developments and techniques to experiment with, that may improve upon what came before.

### 3.2.1 TIL Feature Extraction

The next step for this project involves extracting relevant features from DCIS regions identified by the described pipeline. One project currently leveraging the DCIS prediction maps aims to quantify tumour-infiltrating lymphocytes (TILs) surrounding DCIS. TILs are immune cells that have exited the bloodstream and migrated towards a tumour to combat it. The presence, shape and arrangement of TILs has been shown to have an association with patient outcome [77]. A new pipeline is being developed to extract TIL population features from regions surrounding DCIS, combining a separate cell classification algorithm with masks output from the described algorithm. An example image of lymphocytes detected in an H&E stained image is shown in Fig. 19. The DCIS mask is dilated to capture the periductal region surrounding DCIS, which is combined with the lymphocyte map to create a map of TILs surrounding a DCIS duct. An example of this process can be seen in Fig. 20. This ductal TIL map can be used to calculate some TIL features, such as count, density, ratio to DCIS ducts. Higher level TIL spatial arrangement and graph features have
also been shown to have prognostic significance [78]. These can be combined with features from the DCIS masks such as duct count and ductal area per slide. This work is ongoing, and we hope to incorporate it into a predictive model for recurrence likelihood per patient.
3.2.2 Predictive Model for Recurrence

Another project underway incorporating the DCIS masks involves using a neural network to predict recurrence using features extracted from DCIS regions. This moves away from a traditional feature extraction and machine learning model framework in favour of a neural network, as has been the trend in the last several years. Given an H&E WSI, the DCIS localization algorithm is run to obtain a DCIS mask. The mask is again dilated to capture the periductal region surrounding DCIS. An example of this sequence can be seen in Fig. 21. A neural network is then trained to predict likelihood of recurrence using data sampled from the given ductal and periductal regions. This allows the network to consider only relevant regions around the DCIS, while ignoring the rest of the slide. When making predictions at test time, the model predictions of each input patch can be stitched together into an image representing predictiveness of recurrence. This step is important for identifying which parts of the WSI the model determines are indicative of recurrence or non-recurrence. An example of such images can be seen in Fig. 22., as part of a comparison between models trained using ductal regions, periductal regions, and a combination of both. Predictions from all relevant regions must be aggregated to form a slide level prediction. This can be accomplished through either a mean or a majority voting scheme, or more involved methods which are being investigated. Preliminary experiments show the model using combined ductal and periductal regions achieves the best performance, however the pipeline needs more refinement before conclusions can be drawn.

Figure 21. From left to right, examples of an H&E WSI, DCIS prediction mask, dilated DCIS mask to include periductal regions. This sequence is used as a preprocessing step for training a neural network to predict recurrence, by only looking at the relevant regions of the WSI.
3.3 Summary of Contributions

The primary contribution of this work is a fast (6.35s/WSI), fully automatic DCIS localization pipeline with high accuracy (DSC=0.77, AUC=0.987), trained and tested on a set of WSIs annotated by an expert pathologist. The algorithm operates on patches extracted from the WSIs, as fitting entire WSIs into CPU memory is not feasible. An important finding of the work is that patches extracted from low resolution (5x) result in higher segmentation performance, due to the U-Net valuing wider contextual information over higher resolution information. This allows the algorithm to run as fast as it does, since far fewer patches need to be fed through the U-Net to cover an entire WSI.

Another major contribution is the development of a multi-resolution U-Net architecture that combined patches from all available magnifications (5x, 10x and 20x) to overcome the resolution vs. FoV trade-off. This custom architecture splits the downsampling arm in three, one for each available resolution, and outputs prediction maps with size equivalent to the highest resolution input (Fig. 11.). This architecture improved upon the object level DCIS detection capabilities of the model (F1=0.68), at the expense of greatly increased training and running times.

Other contributions include the test set comparison to a second expert pathologist annotator, to measure the variability in ground truth segmentations. Our best models showed comparable performance to the second annotator on the segmentation task, with the caveat of having a relatively small amount of dually annotated test data. A modified F1 score was developed as a metric, to correct for DCIS ducts begin grouped together in the ground truth annotations. This score does not penalize the model for identifying multiple DCIS ducts within a single ground truth
region, providing a more accurate assessment of a model’s object detection capabilities. Our hyperparameter search resulted in a set of top models with no significant differences between them, which included the base model with the original hyperparameters. Thus, we chose to move forward with the base hyperparameters, implying the U-Net hyperparameters were satisfactory as originally designed. Spatial maps of model uncertainty were created using Monte-Carlo dropout. These detail regions of a WSI in which the model is highly uncertain and may make an incorrect prediction. We propose that with more development, this technique could be used to flag images in need of further examination by a pathologist. It also highlights situations in which more training data may be needed to increase the model’s robustness. Finally, this work has laid the foundation for targeted feature extraction from WSIs containing DCIS. Several projects underway are utilizing the DCIS masks produced by this pipeline to extract DCIS imaging features and make predictions of a patient’s likelihood of recurrence.
References


[7] Huckfinne [Public domain], via Wikimedia Commons


[41] Aphex34 [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons


4 Appendix

4.1 Early Stopping

Early stopping is a CNN training technique used to prevent overfitting networks to a dataset. Our early stopping method was described briefly in the Training subsection. Each network is trained for 100 epochs while models from the final 10 epochs are saved separately. From those, the model with the lowest validation loss is selected as the final trained model. This method prevents training from stopping on a stochastically unfavourable update while allowing each network to train for a sufficient time. Fig. 23. shows the training curves for the 5x and 3rm U-Nets. Both validation loss curves show large variability, including a sharp increase at the last epoch. This early stopping scheme reduces the risk of stopping at an unfavourable update but does not guarantee an optimal solution as having the lowest validation loss may not be representative of test set performance.

![5x Training Curve](image1)

![3rm Training Curve](image2)

Figure 23. Training and validation loss at each epoch, for the 5x (left) and multi-resolution 3rm (right) U-Nets. The training loss descends steadily over time, but the validation loss is more erratic. The last epoch in both figures shows a jump in validation loss.

4.2 Dataset Distributions

Fig. 24. shows the distribution of each data subset according to DCIS grade and comedo necrosis status (grading system and comedo definitions found in background). Grade and comedo necrosis status were evaluated by an expert pathologist, although that information was not available when the data subsets were created. Fig. 25. shows the test set performance of select models, as well as the second annotator, separated by grade and comedo necrosis status. The dice scores are similar across these subclasses, implying those models have strong performance across a variety of DCIS
subtypes. It must be noted that some of the subclasses (Grade 1&3 and Some Ducts) have very low representation in the test set, so this analysis is limited and inconclusive.

4.3 Object Precision-Recall

Table 4. shows some supplemental data to that shown in Table 3. The first two columns are means of precision and recall used to calculate the F1 score, shown here and in Table 3. These values reflect the modified versions of each metric, described in section 2.2.6. This table provides a contextual breakdown for the F1 score of each model. These results further the case for the 3rm multi-resolution model, as it achieves the highest precision. As mentioned in the discussion, this problem favours an algorithm with high precision that can filter normal ducts from being detected. The results also show the ensemble model’s improvement in precision over the 5x model.
Table 4. Supplementary table to table 3, showing modified versions of precision and recall results for each model, including the second annotator. Also includes previously shown F1 Score and P-Value data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>F1 P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second Annotator</td>
<td>0.65</td>
<td>0.35</td>
<td>0.43</td>
<td>N/A</td>
</tr>
<tr>
<td>5x</td>
<td>0.57</td>
<td>0.78</td>
<td>0.60</td>
<td>0.08</td>
</tr>
<tr>
<td>10x</td>
<td>0.43</td>
<td>0.74</td>
<td>0.48</td>
<td>0.57</td>
</tr>
<tr>
<td>20x</td>
<td>0.50</td>
<td>0.76</td>
<td>0.55</td>
<td>0.15</td>
</tr>
<tr>
<td>3rm</td>
<td>0.79</td>
<td>0.66</td>
<td>0.68</td>
<td>0.01*</td>
</tr>
<tr>
<td>9ch</td>
<td>0.70</td>
<td>0.74</td>
<td>0.66</td>
<td>0.02*</td>
</tr>
<tr>
<td>ensemble</td>
<td>0.65</td>
<td>0.72</td>
<td>0.64</td>
<td>0.03*</td>
</tr>
</tbody>
</table>

4.4 Multi-resolution Output

Fig. 26. shows another example WSI from the tuning set with outputs from the 3rm multi-resolution network overlaid. The top image shows probability map output while the bottom shows an uncertainty map, as described in the model uncertainty section. The probability map shows generally good performance with high probability regions in concordance with the pathologist annotations. Some erroneous regions are detected with lower probability while part of the ground truth region is missed by the network. The probability map also contains many tiling artifacts, which get removed during post-processing. The uncertainty map shows very low uncertainty in most tissue regions, only showing high uncertainty in inked regions outside of the main tissue, highlighting the lack of inked regions in the training data. There appears to be limited correlation between the correctness of a prediction and the associated uncertainty value, limiting the usefulness of the technique for the multi-resolution network.
Figure 26. Sample slide from the tuning set outputs from the 3rm multi-resolution network displayed as a colour overlay. Top image shows probability map output (no post-processing). Bottom image shows uncertainty map overlay, created as described for Fig. 10.