Automatic Cellularity Assessment from Post-treated Breast Surgical Specimens

Mohammad Peikari**,1, Sherine Salama2, Sharon Nofech-Mozes2, and Anne L. Martel1,3

1Medical Biophysics, University of Toronto, Canada
2Laboratory Medicine and Pathobiology, University of Toronto, Canada
3Physical Sciences, Sunnybrook Research Institute, Canada

4th July 2017

Abstract

Purpose: Neoadjuvant treatment (NAT) of breast cancer (BCa) is an option for patients with locally advanced disease. It has been compared with standard adjuvant therapy with the aim of improving prognosis and surgical outcome. Moreover, the response of tumor to the therapy provides useful information for patient management. The pathological examination of the tissue sections after surgery is the gold-standard to estimate the residual tumor and the assessment of cellularity is an important component of tumor burden assessment. In the current clinical practice, tumor cellularity is manually estimated by pathologists on hematoxylin and eosin (H&E) stained slides, the quality and reliability of which might be impaired by inter-observer variability which potentially affects prognostic power assessment in NAT trials. This procedure is also

** Corresponding author. Email: mpeikari@sri.utoronto.ca, address: room M6-621F, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5.
qualitative and time consuming. In this paper we describe a method of automatically assessing cellularity.

**Method:** A pipeline to automatically segment nuclei figures and estimate residual cancer cellularity from within patches and whole slide images (WSIs) of BCa was developed. We have compared the performance of our proposed pipeline in estimating residual cancer cellularity with that of two expert pathologists.

**Result:** We found an intra-class agreement coefficient (ICC) of 0.89 (95% CI of [0.70, 0.95]) between pathologists, 0.74 (95% CI of [0.70, 0.77]) between pathologist #1 and proposed method, and 0.75 (95% CI of [0.71, 0.79]) between pathologist #2 and proposed method. We have also successfully applied our proposed technique on a WSI to locate areas with high concentration of residual cancer. The main advantage of our approach is that it is fully automatic and can be used to find areas with high cellularity in WSIs. This provides a first step in developing an automatic technique for post NAT tumor response assessment from pathology slides.

**Keywords**— Pathology image analysis, breast cancer, machine learning, neoadjuvant therapy

1 **Introduction**

Breast cancer (BCa) is one of the most common occurring cancers in women [42]. There are several treatment options available to treat BCa depending on its type or stage which include surgery, radiation therapy, chemotherapy, or targeted therapy. Neoadjuvant (preoperative) therapy (NAT) is a treatment approach for selected high-risk and/or locally advanced BCa patients in which chemotherapy or targeted therapy is administered before the surgery [48]. There have been several advantages reported for the neoadjuvant treatment including: a) tumor size reduction for patients with large tumors which makes them eligible for surgical resection or candidates for breast-conserving surgery instead of mastectomy; b) efficacy of the therapy can be assessed in-vivo which helps in the modification or change of treatment for patients who show little to no respond to the therapy; and c) association of tumor response to the therapy is facilitated by comparing tissue samples before, during, and after treatment to identify gene expression profiles associated with tumor response and/or development of novel
therapeutic agents [11, 23, 41, 48]. It is desirable to achieve a pathologic complete response (pCR) after the course of NAT as this is directly correlated to prognosis and efficacy of the treatment [18, 62, 31]. pCR is defined as the absence of residual cancer in the breast tissue and associated axillary lymph nodes at the completion of NAT [47]. Tumor response during NAT is monitored by physical examination (palpation), ultrasonography, and mammography [7, 27]. While accurate prediction of the residual cancer is crucial after NAT in assessing the efficacy of the therapy, clinical and radiologic assessment of the response often under- or over-estimates the amount of residual cancer [7, 41, 25].

Although reporting treatment response for post NAT is not part of the synoptic reporting system suggested by American Joint Committee on Cancer (AJCC) manual, a descriptive assessment of response is commonly reported qualitatively across all health care centres including Sunnybrook Health Sciences Centre (SBHSC). Currently, the pathologic examination of the tissue sections after surgery is the gold-standard to estimate the residual tumor. Quantitative measurement of the residual cancer after NAT is even more beneficial for drug trials to assess the success of newly developed drugs. In [40, 52], different proposed systems for evaluating the pathologic response to NAT are summarized. Currently, tumor burden is manually estimated by pathologists on hematoxylin and eosin (H&E) stained slides, the quality and reliability of which might be impaired by inter-observer variability which potentially affects pCR and prognostic power assessment in NAT trials [45, 47]. This procedure is also qualitative and time consuming in the current practice [1, 5, 34, 32].

Although there is no standardized definition of the factors considered to assess post-treatment tissue sections for pCR [26, 36], the cellularity fraction of cancer within tumor bed (TB) seems to be a better prognostic indicator of the efficacy of the treatment compared to other factors [37, 26, 25]. The objective of this paper is to present a pipeline to automatically assess the cellularity of residual tumor from post-treatment pathology image patches and whole slides of breast cancer. An automated method could provide pathologists a second opinion assessment on the tissue regions and provides bases to gain more information from post-NAT tissue micro-environment [46].
1.1 Background

Many features used in the analysis of pathology slides are inspired by visual markers characterized by pathologists. A large subset of these features are derived from nuclei objects present within the tissue micro-environment. Therefore, before attempting to assess the tumor cellularity, it is an important first step to detect and segment nuclei objects from regions of interest (ROIs). While some studies focus on producing high quality nuclei segmentation for subsequent automated processing of pathology images [57, 20], others claim perfect nuclei segmentation does not necessarily lead to a better classification performance [4]. Most of the currently present nuclei segmentation techniques in the literature utilize one or a combination of watershed segmentation [60, 9, 50, 30], graph-cuts [2], matching-based [28], Laplacian of Gaussian (LoG) [2], active contours [57, 30, 51], or pixel classification [51, 22, 44] methods combined with pre- or post-processing phases [33].

Once the nuclei figures are segmented, features defining their morphology, texture, and spatial relationship with one another are calculated to represent different nuclei classes in a multi-dimensional feature space [24, 14]. Morphological features characterize size and shape of nuclei figures such as area, perimeter, major axis, minor axis, shape factor, etc [54], texture features are used to describe gray level pixels inter-relations with each other or provide cell interpretations in a multi-scale fashion such as Daubechies wavelet features [19] or Gabor wavelet features [10], and spatial relationship features are derived to summarize the way nuclei figures are located with respect to each other to form disease micro-environment such as graph-based features [6, 10, 13, 29] or spin-context [3, 49].

After deriving appropriate features characterizing nuclei figures, statistical models such as support vector machines (SVM) [15], Naive Bayes [12], neural networks [15], and decision trees [58] are employed to learn and distinguish the differences between patterns of nuclei figures. Nuclei classification approaches have been used as a processing step toward various pathology automated analysis projects in the literature including identification of intraductal proliferative lesions [59], correcting for copy number profiles between samples of BCa [61], nuclei grading [56, 35], and other [55].

Recently, deep learning based methods have gained much attention in the medical image
analysis community due to their better performance in some applications compared to the traditional machine learning approaches [21, 39, 43]. They work by iteratively improving learned data representations by deriving features that are originally calculated from the data itself (usually by applying some texture filters on the training images) with the aim of achieving maximum class separability. Compared to the traditional machine learning approaches, deep learning methods do not require designing and calculating hand-crafted features, scale well to large datasets, and can easily be adopted to other applications. However, deep learning methods are criticized in medical image analysis community for their need for large datasets to outperform traditional learning methods, and not being easy to correlate their findings with theoretical medical foundations.

In this study, we chose to use traditional machine learning approaches to assess residual cancer cellularity due to our limited dataset size and also better interpretability of these techniques. Therefore, the approach taken to assess residual cancer cellularity is to divide whole slide images into smaller patches, and from within the smaller patches or regions of interest (ROI) detect and segment nuclei figures. Segmented malignant nuclei figures are then distinguished from lymphocyte and normal epithelial figures using an SVM classifier. The cellularity estimation is then performed using distinguished malignant epithelial figures for every ROI.

2 Dataset

2.1 Pathology Whole Slide Images

The dataset used to train and validate the tumor cellularity assessment consists of $n = 121$ H&E stained post neoadjuvant whole slide images (WSIs) from 64 patients. They were prepared and scanned at 20X magnification (0.5 μm/pixel) in the Department of Anatomic Pathology at Sunnybrook Health Sciences Centre (SBHSC), Toronto, Canada after receiving appropriate ethical approval from the hospital and were annotated by expert pathologists using Sedeen Viewer (Pathcore, Toronto, Canada). Both collaborating pathologists in this study are specialist in anatomical pathology with breast pathology fellowship training at
the University of Toronto with more than 10 years of experience in practice. 67 slides (corresponding to 31 patients) were used to train and 25 slides (corresponding to 15 patients) were used to validate the nuclei classification statistical models (training part- also referred to as sets A and B respectively in the text). Furthermore, 29 slides (corresponding to 18 patients) were used to validate the cellularity assessment pipeline (validation part- also referred to as set C in the text). Table 1 illustrates the histopathologic characterization of patients included in the training and validation sets used in this study. The training and validation sets were chronologically gathered from cases with residual disease between years ??? for training set and ??? for validation set. As shown in table 1 patient data used for this study include different hormonal status, grade and clinical response categories. There were 8 patient cases specifically added to the training set only to include the complete pathologic response category. Some of the eligibility criteria for selecting patients for NAT trials include having clinical stage of II or III, presence of objective measurable lesion(s), and previously untreated disease. Here, we briefly explain the dataset preparations used for nuclei object classification and cellularity assessment pipelines.

2.1.1 Nuclei Figure Annotation

Nuclei figures were marked by an expert pathologist by clicking on the centroid of individual nuclei from within a rectangular ROI on the tissue slides of the training part (sets A and B). Size and location of ROIs were arbitrary and chosen from within the tumor bed by the expert pathologist in such a way that they contained a mixture of nuclei figure types (benign or malignant epithelial, and lymphocyte). Figure 1 depicts an example of a rectangular ROI from within which nuclei centroids were marked by expert pathologist. Marked Nuclei figures were the ones for which the pathologist was certain about their types. More than 30000 nuclei \( (n = 3868 \text{ lymphocyte}, n = 10407 \text{ benign epithelial}, \text{ and } n = 16419 \text{ malignant epithelial figures})\) were marked by the pathologist from 166 ROIs in sets A and B which included a mixture of lymphocyte, benign, and malignant epithelial nuclei figures.
2.1.2 Cellularity Assessment within Patches

In order to evaluate the proposed cellularity assessment pipeline, ROI patches from image slides in the validation part (set C) were extracted by an expert pathologist (pathologist #1) and their individual cellularity was reported as a scalar value between 0 and 100%. All image patches were selected from within tumor bed regions and were 512x512 pixels (corresponding to an area of about 1 mm\(^2\)). To ensure a balanced selection between various patches with different cellularity portions, the pathologist was asked to select representative ROIs in four roughly estimated cellularity categories of 0 (normal), 1-30 (low cellularity), 31-70 (medium cellularity), and 71-100\% (high cellularity). \(N = 1121\) ROI patches were reviewed by the pathologist from 29 WSIs. Figure 2 shows a WSI from which a subset of cellularity assessment dataset used for this study were chosen with a representative high cellularity patch. A second pathologist (pathologist #2) was also asked to score the same patches marked by pathologist #1 and their results were stored to obtain the inter-rater variability among pathologists and compare with that of the performance of the presented automated method.

3 Methodology and Experimental Setup

Segmentation and classification of nuclei is an important step to analyze pathology slides. In this study, in order to assess the cellularity of tumor bed after NAT we particularly need to be able to distinguish between lymphocyte, benign and malignant epithelial nuclei. Therefore the following multi-step process was considered to estimate the cellularity of residual tumor from pathology images: 1) segment nuclei figures from regions within tumor bed, 2) learn a cascaded statistical model to distinguish between nuclei classes (malignant epithelial, benign epithelial, and lymphocytes); use set A to train and tune the statistical models and set B to validate their performances, and 3) estimate the cellularity of residual cancer from patches using classification results of malignant epithelial figures in step 2; use set C to validate cellularity assessment pipeline.
3.1 Nuclei Figure Segmentation

In [33], we proposed a nuclei segmentation method to extract nuclei figures from pathology slides. Briefly, the segmentation technique works by decorrelating and stretching the original RGB colorspace in image slides. Next, the now decorrelated and stretched RGB colorspace was converted into Lab colorspace to reduce dependency of the pixel values to intensity information and used multilevel thresholding and marker controlled watershed algorithms to segment foreground and separate overlapping nuclei figures. The F1-score of the nuclei segmentation technique was 0.9 when evaluated on 7931 nuclei from 36 images from a publicly available benchmark dataset in [57].

For this study, the effect of color variation within slides was minimized before segmenting nuclei figures by standardizing their color to a reference image. The reference image was chosen from an image slide not used in this study and was selected due to its good color contrast between different image components corresponding to various tissue regions (fat, stroma, and nuclei). The reference and selected images were both converted to Lab color space and selected (source) image color channels were standardized to those of reference image using the following equations:

\[
\begin{align*}
l_t &= (l_s - l_s) \frac{\sigma^l_s}{\sigma^l_r} + l_r; \\
a_t &= (a_s - a_s) \frac{\sigma^a_s}{\sigma^a_r} + a_r; \\
b_t &= (b_s - b_s) \frac{\sigma^b_s}{\sigma^b_r} + b_r
\end{align*}
\]

where \(l_t, a_t,\) and \(b_t\) are the \(l, a,\) and \(b\) channels of the target image respectively; \(l_s, a_s,\) and \(b_s\) are the \(l, a,\) and \(b\) channels of the reference image respectively; \(l_r, a_r,\) and \(b_r\) are the means of \(l, a,\) and \(b\) channels of the source image respectively; \(\sigma^l_s, \sigma^a_s,\) and \(\sigma^b_s\) are standard deviations of the \(l, a,\) and \(b\) channels of the source image respectively; and \(\sigma^l_r, \sigma^a_r,\) and \(\sigma^b_r\) are standard deviations of the \(l, a,\) and \(b\) channels of the reference image respectively. The target Lab image was then converted back to the RGB colorspace.
3.2 Nuclei Figure Labeling

In order to train a classifier to differentiate between normal, malignant, and lymphocyte nuclei, it was necessary to label each segmented figure using the manually labeled data. Therefore, to ensure only one label is assigned to every segmented figure, centroids of segmented nuclei figures were matched with at least one entry in the groundtruth sets A and B. To match the centroids in any given patch, the euclidean distance between the segmented figure and all nuclei centroids in the groundtruth training set were compared. Given the euclidean distance of any two nuclei centroids in segmented group and groundtruth training set being lesser than 10 pixels (5 μm), the label of the nucleus in the groundtruth training set was assigned to the closest segmented figure. The assigned label was then removed from the pool of available groundtruth nuclei to avoid assigning label of one nucleus to multiple segmented figures. More than 72% of nuclei figures ($n = 21779$) in the groundtruth training set were matched with one segmented nuclei.

3.3 Nuclei Figure Classification

After nuclei were segmented from color standardized pathology images, they were converted to gray scale and their intensity level was normalized to have zero mean and standard deviation of one. For every nuclei figure, 125-dimensional features describing shape (mean and standard deviation of area, perimeter, eccentricity of all nuclei within a square with edges of size $S = 0.125$ mm and with the nuclei figure’s centroid as its center, and shape factor $\frac{\text{Perimeter}}{2\times\sqrt{\text{area}}}$), intensity (mean, mode, median, standard deviation, skewness, and kurtosis of the gray scale histogram of pixels within the segmented nuclei figure), texture (10 dimensional normalized histogram of uniform rotation-invariant local binary patterns [53] from 8-neighborhood pixels from within the segmented nuclei figure, and 88 dimensional Haralick textures [17] at 0, 45, 90, and 135 degrees from within the segmented nuclei figure), and 14 dimensional spatial features (comprised of 12 for number of nuclei present around the considered nuclei in spiral growing circles with diameters starting from 0.04 mm to 0.125 mm with a step size of 0.04 mm, including mean and standard deviation of the values from all 12 circles) were calculated from each individual segmented nuclei figures.
The nuclei features extracted from 92 slides (sets A and B) were divided into two parts: set A \((n = 13821\) nuclei objects from 67 slides) to train statistical models to distinguish between three nuclei classes (lymphocyte, malignant epithelial, and benign epithelial) and set B \((n = 7958\) nuclei objects from 25 slides) to check generalization performances of statistical models learned from set A. We trained a cascaded classifier from set A to first distinguish lymphocyte from epithelial figures \((L \text{ vs. } BM, n = 2260\) lymphocyte, and \(n = 11561\) epithelial figures) and then to separate epithelial figures into benign versus malignant figures \((B \text{ vs. } M, n = 3157\) benign, and \(n = 8404\) malignant figures). A five-fold cross-validation scheme was used for both cascades to train SVM classifiers with non-linear radial basis function (RBF) kernels. We used the SVM implementation present in the libsvm [8] library. In each fold, the optimum hyperplane parameters were found by testing different possible combinations of the classifier’s trade-off parameter \((C)\) and kernel width \((\gamma)\) in order to maximize accuracy of classification. A final model for each classifier cascade was then generated by considering the whole set A (for each cascade) and combining the hyperplane parameters corresponding to the best classification performance by taking the median of the five parameters (corresponding to 5-folds). The overall trained model generated for each cascade was then used to validate the classification performance on nuclei objects present in the independent and unseen set B. Stromal nuclei were eliminated from the segmented figures both for tumor cellularity assessment and nuclei figure classification by removing the objects with the ratio of major to minor axes equal or more than 3.

3.4 Residual Cancer Cellularity Assessment

Once statistical models were trained to classify nuclei figures to any of lymphocyte, normal, and malignant epithelial classes, they were used to assess the cellularity of image patches in set C. In order to reduce the false-positive and false-negative detection rates on images in set C, a k-nearest neighbor approach (with \(k = 8\)) was used to correct nuclei labels that were mostly surrounded by one class of nuclei. Using this post-processing approach, about 2% improvement in the performance accuracies was observed from the images in set B.

Once the individual nuclei figures were classified to the three known classes and the malignant epithelial candidates were identified, a morphological dilation with a structuring
element of size $s = 11$ was applied on the malignant figures to expand the malignant figures to account for the presence of cytoplasm around each nuclei. The cellularity of any given image patch was then calculated as the area covered by malignant figures divided by the area of the image patch. In order to correct for variation between the area calculated from morphologically expanded figures and the actual scores given by the pathologist in set C, the information on a portion of patches ($n = 758$) in set B that was previously annotated by pathologist #1 was used to find correction parameters using a linear regression.

4 Results

4.1 Nuclei Object Classification

Table 2 summarizes the mean performance of the 5 fold cross-validation on set A consisting of $n = 13821$ nuclei objects from 67 slides and Table 3 summarizes the performance of applying the model generated from set A on the unseen and independent testing set B. Comparing the performances reported in Tables 2 and 3 indicates that the developed classification pipeline is generalizable on independent sets to classify nuclei objects with $AUC_{LvsBM} = 0.87$ and $AUC_{BvsM} = 0.86$. The same model was used in subsequent analysis to assess residual cancer cellularity. Figure 3 shows the performance of nuclei classification on two representative patches from the validation dataset (set C) one covered by normal epithelial and another by lymphocyte and malignant epithelial components.

4.2 Residual Cancer Cellularity Assessment

In order to compare the performance of the proposed cellularity assessment pipeline with that of the expert pathologists, they were reported both in terms of categorical and correlational agreements. Table 4 summarizes the confusion matrices of different cellularity categories comparing pathologists and the automated method. Considering the first confusion matrix in Table 4 (pathologist #2 vs. pathologist #1), it can be seen that the two pathologists were mostly in agreement when they were looking at normal or low cellularity patches, while their level of agreement decreased when annotating medium and high cellularity patches with
pathologist #2 tending to assign a higher value of cellularity. The Cohen’s Kappa agreement between pathologist #1 vs. pathologist #2 is $\kappa = 0.66$. On the other hand, comparing the automated method with pathologists #1 and #2, we see that while accuracy for the medium cellularity patches is very good, the automated method seems to perform less accurate for normal and low cellularity patches. Furthermore, the automated method’s performance is very close to those of pathologists #1 and #2 for high cellularity patches. The Cohen’s Kappa agreements between pathologist #1 vs. automated method and pathologist #2 vs. automated method are $\kappa = 0.42$ and $\kappa = 0.43$ respectively.

Table 5 shows the degree of absolute intraclass agreements (also known as intraclass correlation - ICC) among the expert pathologists and the proposed automated method in this paper.

Figure 4 shows the heat scatter plot of the agreement between pathologists and the automated method. It is clear from this figure that the cellularity of most of the patches were assessed closely to those of the pathologists as most of the patches laid close to the correlation (black) line. There are some other patches for which the assessment pipeline fails to estimate their true cellularity assigned by the pathologists (located at the top-left or bottom-right of the graphs in Fig. 4) due to failure in nuclei segmentation or classification, and presence of large difference between the cellularity reported by one of the pathologists and that of estimated by the automated pipeline. Figure 5 shows some examples where the automated method fails to correctly estimate the cancer cellularity compared to the pathologists due to the aforementioned reasons.

We have also applied the proposed automated method on a whole slide image to find regions of high residual cancer (Fig. 6). As seen in Fig. 6, our proposed method was able to find regions with high residual cancer cellularity while neglecting regions that contained normal components.

5 Discussion and Conclusion

In this study, we developed a framework to automatically estimate residual cancer cellularity from patches within whole slide images of BCa. We extensively validated the framework using
a dataset consisting of 64 patients (n = 121 WSIs). This automated tool could potentially decrease assessment inter-observer variability from pathology slides of post NAT.

The agreement coefficient between the pathologists and automated method (Table 5) could have been much better if some of the outliers in Fig. 4 (right) were not present. From the 124 patches (11% of patches) where the difference between the assessed cellularity by automated method and at least one of the pathologists were more than 30%, 52 (42%) were caused by mis-classification of nuclei objects in an entire patch (Fig. 5(a, b)), 18 (14%) were caused by failure in nuclei segmentation procedure due to out-of-focus regions or identifying wrong nuclei boundaries (Fig. 5(a)), and 54 (43%) were due to large differences (more than 30%) between automated assessment method versus at least one of the pathologists (Fig. 5(c)). It is possible that using better segmented nuclei figures results of classification performance could be improved however, in [4] authors found that perfect nuclei segmentation does not ensure better classification performance after using both perfectly and imperfectly segmented nuclei figures to train a classifier to detect malignant objects from breast cancer tissues.

In order to investigate the consistency of the SVM classifier performance in dealing with imbalanced datasets for training nuclei figure classification, the weight values of the classes with fewer data points were increased to penalize mis-classifications in the minority class. The performance of the SVM classifier were consistent with and without changing the weight value parameter of the classifier.

To test the performance of our pipeline with top ranked features that are most effective for nuclei classification task, we have used various feature selection methods that are compared in [38]. We found that classification performance of using different proportions of selected features using various feature selection methods were no better than our original 125-dimensional features. However, in order to show the most relevant features for nuclei classification task we have reported the top 10 features chosen by the Fisher’s method [16] (which performance has been reported to be the best in [38] compared to some other methods) in table 6. As shown, top ranking features correspond to Haralick texture features. This finding also matches with other research paper findings in the literature [59].

While the aim of this study was to develop an automated method to assess the residual
tumor cellularity in image patches of BCa, it is worth mentioning that it is clinically desirable to be able to identify patients who have an estimated residual disease of lesser than 20% after the completion of NAT (near complete or complete pathologic response). Using our proposed automated pipeline, we were able to correctly identify 68%, and 75% of the patches with residual tumor cellularity of lesser than 20% identified by pathologist #1 and pathologist#2 respectively.

In order to have a fully automated system to assess residual tumor after NAT, areas of tumor bed and ductal carcinoma in-situ (DCIS) also need to be identified. Tumor bed regions are where tumor cells are currently present or might have been present when the neoadjuvant treatment started. To identify tumor bed regions pathologists look at features indicating the process of tissue healing after it was affected by the therapy. Regions with features like high concentration of fibroblast cells within fibrosis, foamy macrophages with pigments, calcifications, aggregates of lymphocytes, areas of tumor necrosis, thin collagen bundle strands, and small capillaries are indications for the tissue that is healing and hence are considered as a part of tumor bed. It is technically a difficult task to identify each of the mentioned features in the pathology slides using image analysis methods with hand-crafted features due to their high variability and subjectivity in their identification. However, deep learning based methods could be employed to tackle this problem due to their ability to derive suitable features from the data itself.

Acknowledgment

This research is funded by the Canadian Cancer Society (grant number 703006) and with added support from the National Institutes of Health (grant number U24CA199374-01).

References

Table 1: Histopathologic characterization of the 64 patients in training and validation sets. IDC: intra-ductal carcinoma, ILC: intra-lobular carcinoma, CPR: complete pathological response, PDR: probable or definite response, and NDR: no definite response.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Training Set (n = 92)</th>
<th>Validation Set (n = 29)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>50 (± 12.6)</td>
<td>51 (± 12.7)</td>
</tr>
<tr>
<td>Tumor bed size (mm)</td>
<td>51.3 (± 35.7)</td>
<td>84.7 (± 82.2)</td>
</tr>
<tr>
<td>Hormonal status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER</td>
<td>63 (68%)</td>
<td>19 (65%)</td>
</tr>
<tr>
<td>PR</td>
<td>46 (50%)</td>
<td>17 (57%)</td>
</tr>
<tr>
<td>HER2</td>
<td>58 (63%)</td>
<td>19 (65%)</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>12 (13%)</td>
<td>4 (14%)</td>
</tr>
<tr>
<td>2</td>
<td>49 (53%)</td>
<td>14 (48%)</td>
</tr>
<tr>
<td>3</td>
<td>18 (19%)</td>
<td>11 (38%)</td>
</tr>
<tr>
<td>N/A</td>
<td>13 (14%)</td>
<td>0</td>
</tr>
<tr>
<td>Histology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDC</td>
<td>65 (71%)</td>
<td>27 (93%)</td>
</tr>
<tr>
<td>ILC</td>
<td>10 (11%)</td>
<td>2 (7%)</td>
</tr>
<tr>
<td>Other</td>
<td>16 (17%)</td>
<td>0</td>
</tr>
<tr>
<td>Clinical Response</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPR</td>
<td>8 (9%)</td>
<td>0</td>
</tr>
<tr>
<td>PDR</td>
<td>77 (84%)</td>
<td>27 (93%)</td>
</tr>
<tr>
<td>NDR</td>
<td>5 (5%)</td>
<td>2 (7%)</td>
</tr>
<tr>
<td>Other</td>
<td>1 (1%)</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 1: Subset of nuclei annotations by expert pathologist. Centroids of different nuclei classes (malignant epithelial in this demonstration) were marked by the pathologist. Regions of interest were selected from within the tumor bed regions (black contour).

Table 2: Mean performance of 5-fold cross-validation on set A (n = 13821 nuclei objects from 67 slides).

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Accuracy (SD) %</th>
<th>Sensitivity %</th>
<th>Specificity %</th>
</tr>
</thead>
<tbody>
<tr>
<td>L vs BM</td>
<td>0.97</td>
<td>95 (± 0.1)</td>
<td>79</td>
<td>99</td>
</tr>
<tr>
<td>B vs M</td>
<td>0.86</td>
<td>83 (± 0.5)</td>
<td>57</td>
<td>92</td>
</tr>
</tbody>
</table>

Table 3: Performance of applying the generated model from data used in Table 2 on n = 7958 nuclei objects from the independent testing set B from 25 slides.

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy %</th>
<th>Sensitivity %</th>
<th>Specificity %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lymphocyte (L)</td>
<td>92</td>
<td>80</td>
<td>94</td>
</tr>
<tr>
<td>Benign Epithelial (B)</td>
<td>75</td>
<td>50</td>
<td>92</td>
</tr>
<tr>
<td>Malignant Epithelial (M)</td>
<td>77</td>
<td>91</td>
<td>63</td>
</tr>
</tbody>
</table>
Figure 2: Subset of cellularity assessment dataset from image patches annotated by expert pathologist. 512x512 pixels patches were selected by the pathologist from within the tumor bed (black contour) and were given a score to depict the cellularity of the malignant components within the given patch. Four categories of patches were chosen by the pathologist: 0% (pink), 1-30% (blue), 31-70% (yellow), and 71-100% (green). The image patch shown in this figure was scored as 80% by pathologist #1.
Figure 3: Examples of nuclei figure segmentation and classification on two representative patches of size 512x512 pixels from set C. On top is an image patch covered by lymphocyte and malignant epithelial components and on bottom is an image patch covered by benign epithelial nuclei. Their classified mask images are shown to their right. Green represents lymphocyte, blue represents benign epithelial, and red represents malignant epithelial classified objects.
Table 4: Categorical agreements between pathologists #1, pathologist #2 and the proposed pipeline (automated) for \( n = 1121 \) patches of size 512x512 pixels (set C).

<table>
<thead>
<tr>
<th></th>
<th>Pathologist 1</th>
<th>Pathologist 2</th>
<th>Automated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Normal</td>
<td>237</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>202</td>
<td>23</td>
</tr>
<tr>
<td>Medium</td>
<td>0</td>
<td>102</td>
<td>197</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>3</td>
<td>155</td>
</tr>
</tbody>
</table>

Table 5: Intra-class correlation coefficients (ICC) comparing the agreement levels between pathologist #1, pathologist #2 and the proposed pipeline (automated) for \( n = 1121 \) patches of size 512x512 pixels (set C).

<table>
<thead>
<tr>
<th></th>
<th>ICC Coefficient (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathologist1 vs Pathologist2</td>
<td>0.89 [0.70, 0.95]</td>
</tr>
<tr>
<td>Pathologist1 vs Automated</td>
<td>0.74 [0.70, 0.77]</td>
</tr>
<tr>
<td>Pathologist2 vs Automated</td>
<td>0.75 [0.71, 0.79]</td>
</tr>
</tbody>
</table>
Figure 4: Heat scatter plot of the agreement between the pathologists and the automated method. The color bar indicates the number of patches for each point/circle. As can be seen, slopes of correlation lines for automated method versus pathologist #1 and #2 plots are close to each other and are different from that of the two pathologist’s plot.

Table 6: Top 10 features chosen by Fisher’s feature selection method [16] for the nuclei classification task. As shown top features correspond to Haralick texture features.

<table>
<thead>
<tr>
<th>Top 10 feature ranks for nuclei classification task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L vs. BM</strong></td>
</tr>
<tr>
<td>Cluster Shade at 135°</td>
</tr>
<tr>
<td>Cluster Shade at 0°</td>
</tr>
<tr>
<td>Cluster Shade at 90°</td>
</tr>
<tr>
<td>Cluster Shade at 45°</td>
</tr>
<tr>
<td>Autocorrelation at 135°</td>
</tr>
<tr>
<td>Sum variance at 135°</td>
</tr>
<tr>
<td>Autocorrelation at 45°</td>
</tr>
<tr>
<td>Sum variance at 135°</td>
</tr>
<tr>
<td>Sum variance at 90°</td>
</tr>
<tr>
<td>Sum variance at 0°</td>
</tr>
</tbody>
</table>
Figure 5: Image examples where the automated pipeline fails to correctly assess the cellularity of patches with reference to that reported by pathologists: (a) part of a patch with normal epithelial tissue structure where segmentation fails to correctly identify true nuclei boundaries due to them being out-of-focus, (b) part of a patch with cancer cells are within fatty tissue; this type of tissue structures were not present in the training set and so was new to the classification pipeline, (c) an example of a patch where the difference between the assessed cellularity by automated method was large compared to at least one pathologist.
Figure 6: Cellularity assessment heat map from within a tumor bed region (black contour) on a whole slide image with two representing benign and malignant regions.


[16] Quanquan Gu, Zhenhui Li, and Jiawei Han. Generalized Fisher Score for Feature Selection. 2012.


*Annals of Oncology*, 23(SUPPL. 10), 2012.

[49] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks 
and 3D brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine 

[50] Mitko Veta, Paul J van Diest, Robert Kornegoor, André Huisman, Max a Viergever, 


