Machine Learning in Digital Histopathology

Justin Pontalba, Rokshana Geread

***Abstract*—Novel method breast cancer grade classification for proliferation index and malignancy using IHC and H&E stained biopsies. Various comparisons of results obtained from Machine learning algorithms, including Random Forests, Navies Bayes, Logistic Regression, K-Nearest Neighbours, etc. Additionally, a thorough analysis of differences between standardized versus unstandardized data was explored.**

Keywords—IHC, H&E, Breast Cancer, Random Forests, Navies Bayes, K-Nearest Neighbour

I. INTRODUCTION

Female breast cancer is the second most commonly diagnosed cancer worldwide and the leading cause of cancer-related deaths in females [1]. 2.1 million newly diagnosed female breast cancer cases occurred in 2018, in which approximately 80% of cases were attributed to invasive ductal carcinoma (IDC) [1] [2]. In a study completed by Walters et al. [3], six high-income countries demonstrated a one-year net survival for individuals diagnosed at TNM stage I of 100%, compared to those diagnosed at TNM stage IV, which resulted in only 53% (UK) - 66.9% (Denmark). What can be concluded from this study is that breast cancer survival is attributed largely in part by both differences in stage at diagnosis and stage-specific survival. Prognosis of breast cancer traditionally included the use of the TNM staging system, which describes the anatomical extent of a tumor. This system analyzes the initial size of the tumor, whether the cancer has spread to the lymph nodes, and whether the cancer has metastasized [4]. Due to rapid advances in clinical and laboratory science, other factors have been integrated into breast cancer prognosis to give a more accurate progression of the disease [4]. Biomarkers such as histological grade, hormone receptor status (estrogen receptor (ER), progesterone receptor (PR)), human epidermal growth factor receptor-2 (HER2), and a marker for proliferation (Ki-67), are now used in combination with the TNM system to provide accurate staging [4].

Critical to patient management and diagnosis of breast cancer, is the evaluation of stained tissue under assessment by a pathologist. Stains such as hematoxylin and eosin (H&E), and immunohistochemical (IHC) stains such as Ki67, are essential in emphasizing characteristics of tumor malignancy. Pathologists score and provide a grade to suspected tissue based on morphological features of H&E stained specimens. In addition, Ki67 binds to certain proteins of the nuclei that are indicative of cell proliferation. Both of these biomarkers are used in updated best practices for staging breast cancer [4]. The following paper explores proxy measurements extracted of these stain images and their ability to indicate malignancy.

II. MOTIVATION

Manual whole-slide analysis is a rigorous task that requires precision and accuracy. With the use of a digital microscope and computer monitor, pathologists analyze both traditional and digital whole-slides to provide a diagnosis and prognosis of a given specimen. This requires the pathologist to manually segment or annotate regions which they discern to be malignant.

Though digitization of whole-slide specimens improved the workflow of whole-slide analysis and health-record management [4], diagnosis subjectivity and data variability remain as barriers to delivering reliable patient care. Inter-observer variability becomes apparent in breast cancer grading even with a consistent scoring system [5][6]. In addition, unstandardized laboratory practices and inherent tumor heterogeneity also contribute to variability observed in tissue specimens. Lastly, with the digitization of whole-slide specimens, noise and color artefacts consequently affects the quality of the resulting digital image [7][8][9][10]

Advancements in computer vision and machine learning (ML) leverages the potential of computational pathology to provide data standardization and computer aided diagnosis to highly variable and high dimensional data. Furthermore, computer vision and ML demonstrate potential to extract proxy measurements of tumor differentiation sooner than manual histologic or IHC analysis. [25][26].

III. Background on Digital Pathology

A. H&E Staining and Histological Grade

H&E stains are used in routine staining and increase tissue contrast by highlighting specific structures. Hematoxylin, normally dyed purple, is a stain that has an affinity to the nucleic acids contained in the nuclei [4]. Eosin, normally dyed pink, is a counter stain that binds to the cytoplasm of cells. Histological grade is determined by analyzing tissue specimens stained with H&E.

Histological grade is a proxy measure of tumor differentiation and describes three morphological features: i) tubule formation, ii) nuclear pleomorphism, and iii) mitotic count [4]. Various machine learning frameworks demonstrate that nuclear pleomorphic features such as shape, size, and texture, are essential in the prognosis for various cancers [25]. However, before these features are extracted and subsequently classified, stained nuclei from H&E images must first be isolated from the background. In addition, existing segmentations frameworks demonstrate improved accuracy when color normalization or color standardization is applied.

B. Automatic Nuclei Segmentation Using Machine Learning

Nuclei segmentation is an important step in analyzing tissue for malignancy. Individual nuclei segmentation is needed in order to extract features related to the spatial arrangement of cells and nuclear pleomorphism. Traditional nuclei segmentation methods include morphological processing [26], hand-crafted feature design and classification [27], unsupervised clustering [28], and supervised methods that classify each pixel into different categories, nuclei or background [29]. Popular methods such as active contour and watershed segmentation, and their variations, are effective in segmenting nuclei which exhibit uniform colour, texture, and edge content [25]. Yet, the performance of these methods is negatively impacted by the inherent variability of tumor tissue. Visualization of H&E stained breast tissue is greatly affected by tumor heterogeneity, and color variability that is introduced by digital whole-slide scanners, and the lack of standardization in laboratory staining practices [4][5][6]. With the popularization of deep learning, various architectures such as fully connected neural networks, U-net architectures [30], sparse-stacked autoencoders [31], and recurrent neural networks, have demonstrated the ability to produce accurate predictions even with large input data variability [29][30][31][32]. However, the current convolutional neural network based methods only classify pixels into two classes, nuclei or background. Because of this, extensive post processing is often required to separate nuclei. Naylor et al. [34] implemented a fully connected neural network to discriminate between nuclei and background, followed by the watershed method to spit overlapping nuclei. Similarly, Xing et al. [35] propose a CNN-based nuclei segmentation model to generate a probability map of prospective nuclei. Iterative region merging is then applied to the probability map for shape initializations, and nuclei is separated using a sparse shape model and a local repulsive deformable model. Though effective, these methods require extensive post processing after the initial segmentation.

A method by Kumar et al. [29] avoids extensive post-processing, by implementing a CNN-based framework that introduces a third pixel class, the boundary class. This allows for basic post-processing as overlapping nuclei can be separated by subtracting the boundary class image from the nuclei class image. The following paper employs this method of nuclei segmentation on H&E images and analyzes the effect of data variability on nuclei classification

C. Data Standardization using Colour Normalization

Colour normalization is a common preprocessing step in digital pathology used to correct the color variability observed in digital H&E stained images. However, before the stain images are normalized, the stains images must be transformed such that the resulting image is representative of only the pure stains that stain image is comprised of. This transformation is called stain or colour deconvolution (CD). Stain deconvolution removes colour variability by correcting for artefacts related to stain co-localization [36]. Only the pure constituent stain is represented in the resulting image.

Stain deconvolution, proposed in [19], provides a method for quantifying histological stains. For H&E stained specimens, the amount of stain attached or deposited in the tissues follow the Beer-Lambert law [36] (equation 1 below):

$$Ic = Io * e^{(-A * C_c)} \quad (1)$$

where Io is the intensity of light entering the specimen, Ic is the intensity of light detected after passing the specimen, and A * Cc which is the amount of stain with absorption factor C.

Beer-Lambert's (BL) law describes the linear relationship between absorbance and concentration of an absorbing species. This linear relationship can be derived as follows:

$$\frac{I_c}{I_0} = e^{-A * C_c} \quad (2)$$

$$-ln(\frac{I_c}{I_0}) = A * C_c \quad (3)$$

$$OD_c = A * C_c \quad (4)$$

By equation 3, it is apparent that stain concentration is linear. This linear correlation between stain concentration and absorption factor is referred to as optical density (OD). If more than one stain is applied to a specimen, equation 4 is additive. Therefore, if the intensity of light entering the specimen and the intensity of light detected after passing through the specimen are known, the constituent stains can be separated. Each pure stain can be characterized by the a specific optical density for the light detected in each of the RGB cameras of a digital scanner. A 3 by 3 vector will represent the RGB intensity contributions for each stain:

Table 1: Optical density matrix

R	**G**	**B**	
p11	p12	p13	**Stain 1**
p21	p22	p23	**Stain 2**
p31	p32	p33	**Stain 3**

Each column represents the detected intensity of light at each camera, whereas the rows represent the contribution of stain to the intensity detected. If the OD matrix above is **M**, and **C** is a 3 by 1 vector representative for the amounts of stains at a single pixel, then the OD levels detected at that pixel is [36]:

$$y = CM \quad (5)$$

To find the pure stains, C, at the pixel:

$$C = M^{-1} * y \quad (6)$$

where $y = -ln(\frac{I_0}{I_C})$. Practical implementation of this method is proposed in [37].

Once colour deconvolution is performed, the resulting image is colour normalized. The majority of current color normalization methods match the colour of a query image to a target or template image's colour. The differences among these methods are observed in the method in which the OD vector is obtained [7][8][9][10].

D. Immunohistochemistry Ki67 Stain

As mentioned previously, Breast cancer can be diagnosed using a variety of different types of stains. H&E stained biopsies are very common, Immunohistochemistry stains are slowly but steadily becoming a popular choice. Ki67, also known as MKI67, is a nuclear protein that is associated with cell proliferation [19]. Ki67 biomarker scores into a clinical or research workflow would be valuable to stratify patients for personalized medicine therapies, clinical trials and large research studies. Through IHC analysis, Ki67 is being investigated as a clinical marker for breast cancer tumour aggressiveness and proliferation, showing potential in predicting disease survival, recurrence, and response to various treatment options [12][13][14]. Tissue slides are stained with diaminobenzidine (DAB) to visualize Ki67 and counterstained with hematoxylin (H). Pathologists interpret Ki67 slides with scoring systems where manual nuclei counts of both Ki67 positive and negative nuclei are used to estimate a proliferation index (PI); the number of positively stained tumour nuclei divided by the total number of nuclei in a specific region [11]. A higher PI indicates many cells are undergoing cell division, which can signify a more aggressive tumor.

E. Experimental Data Acquisition

Manually annotated data for Breast Cancer biopsy slides are hard to acquire. Staining using Ki67, to indicate cell proliferation, is a relatively new area – hence, data acquisition for Ki67 is minimal, in comparison to H&E dataset. For this experiment a total of 94 TMA core images were used, 2 separate datasets, with very different

staining protocols from two different continents. This will test the methods robustness and overall ability to adapt to variable colours and staining methods.

A total of 30 canine mammary TMA core images obtained from the Ontario Veterinary College (OVC) at the University of Guelph was used to analyze the performance of the proposed method. The TMAs were scanned at the Digital Histology Shared Resource at Vanderbilt University, Lecia SCN400 Slide Scanner was used to digitize the slides. The dataset's colour varying levels of Ki67 stain expression, as shown in figure 5. Individual TMA core images were cropped from whole slide images scanned at 20X magnification with a pixel spacing of 0.5µm using Pathcore's Sedeen [24]. Manual counting was performed in ImageJ [37] by zooming in on the image and placing seed markers at nuclei locations using a computer mouse. This was repeated until the entire core was labeled. Three example TMAs, exhibiting varying biomarker expression levels, PI and colour variability are shown in figure 1.

Another dataset containing 64 Ki67 TMA core images from the Protein Atlas [20] was used to further assess the proposed approach. This open source dataset, specifically the Ki67 protein expression data was acquired from an IHC lab in Uppsala, Sweden. All human tissues were gathered through the Uppsala Biobank and scanned using the two types of Aperio scanners, ScanScope® AT and ScanScope® T2. In order to annotate the data, a custom software, developed within their Laboratory Information Management System (LIMS), was utilized. Parameters that were annotated include, overall staining, stain intensity, proliferation index, pattern and localization of immunoreactivity (nuclear, cytoplasmic or cell membrane). Once the system completed annotating the data, pathologists manually analyzed the TMA's and confirmed these images according to three categories of specific PI ranges, < 25%, 25 – 75% and > 75%. The ground truths for this Ki67 dataset consists of PI ranges, as opposed to visibly annotated nuclei. Three example TMAs from this dataset are shown in figure 2, for the three specific PI classification ranges, low, medium and high scores.

To test the robustness of the proposed framework, the two datasets acquired introduce a wide range of variability. In order to validate the accuracy of this novel method, the entire dataset containing 94 multi-institutional TMA images were scanned using three different scanners, two different continental processing laboratories with contrasting staining protocols, which results in a broad diversity of patients, stain vendors, colour variation and variable staining levels. The combined dataset makes the, total of 94 TMA images, an ideal candidate for testing the robustness of the proposed work.

(a) PI = 29.92 (b) PI = 11.57 (c) PI = 36.23

Figure 1. OVC Ki67 TMA images with scores

(a) PI = < 25% (b) PI = 25 – 75 % (c) PI = > 75%

Figure2: Protein Atlas TMA images with PI ranges

F. Challenges related to utilizing Ki67

The issue with utilizing Ki67 stains is the lack of reproducibility, depending on how the image was stained, the scanners involved to digitize the slides along with the colour variability and overstaining of the slide – it becomes difficult to produce results. Incorporating Ki67 biomarker scores into a clinical or research workflow would be valuable to stratify patients for personalized medicine therapies, clinical trials and large research studies. To enable clinical-use and large-scale studies of Ki67, automated PI quantification tools must provide robust and consistent results across all images and patients. For clinical implementation, ideally, algorithms should be vendor-agnostic, and robust to data variability, so they can be deployed in any clinical centre and lab. For Ki67 standards development and research, algorithms must analyze multi-centre, international datasets robustly and reliably. However, sources of noise and variability in Ki67 images cause challenges for automated approaches. These sources include:

- **Multi-institutional Data**: Across labs and institutions, there are different types of scanning

hardware and stain compounds used that create colour variability across images, tissues, biomarkers and patients [21]. Since algorithms are quantitative, even small changes in the colours of structures can reduce detection performance.

- **Slide Preparation**: Different slide preparation and staining protocols can create stain inconsistency, overstaining and other image artifacts. DAB stromal overstaining causes background regions to carry significant amounts of stain resulting in a reduction of contrast between nuclei and background regions, which impedes nuclei detection.
- **Biomarker Levels**: The biomarker expression level varies from patient to patient, and even within tumours due to disease heterogeneity. Some regions and images have low biomarker concentrations in cells, resulting in lightly stained Ki67 regions, whereas other images may have high Ki67 concentration in cells, resulting in a wide range of intensities and colours that need detection within one framework
- **Ki67 Detection**: The threshold at which ki67 positivity is determined is based on "how brown" a cell's nuclei is.
- **Magnification Level**: Depending on the magnification level (resolution) the images are scanned at, objects such as nuclei can be larger or smaller, and tumour cells can also vary in the size. Therefore, a robust nuclei detection scheme would need to be able to capture objects at a variety of scales.

Despite the utility, there are challenges of incorporating the Ki67 biomarker into clinical and research workflows. Manual PI measurement remains time consuming, and sensitive to inter- and intra-observer variability among pathologists [12][15][16]. There have been international standardization efforts to reduce observer variability in PI measurement, but many challenges remain [17]. Points of discussion include variability in manually counted cells [17], the selection of appropriate cut-off thresholds for protein detection [18], the number of high power fields to evaluate, and the PI ranges that correlate to prognosis, i.e. low (<25%), intermediate (25-75%), and high (>75%) levels of proliferation activity. The European Society for Medical Oncology (ESMO), the American Society of Clinical Oncology have come to the consensus that Ki67 may be a useful clinical tool if it could be standardized. Therefore, to increase clinical utility and adoption of Ki67, robust standards development, research and tools to improve the consistency of PI quantification are needed.

Similar the H&E stained analysis automated image analysis and machine learning tools for IHC analysis are a great solution to these challenges. They offer efficient, reliable and objective PI measurement for large amounts of data. There are reimbursement models in the US that support using digital image analysis to improve consistency of IHC analysis for clinical use [45]. Such tools can be used to assist pathologists in obtaining reproducible PI measures [21]. Automated PI calculators could also be used to develop more consistent Ki67 scoring guidelines as well as for breast cancer research, since large amounts of images could be analyzed and correlated to clinical data efficiently.

G. IHC analysis using Colour Deconvolution

For robust and accurate automated PI estimation algorithms, and IHC analysis in general, separating the hematoxylin (H) stained nuclei from the IHC positive structures is the first and most critical piece of the pipeline. Such a tool allows for the biomarker positive regions to be analyzed separately from the negative regions, which improves robustness. Any inconsistencies in this phase can greatly affect successive nuclei detection and PI quantification techniques. Colour deconvolution, has dominated digital pathology image analysis as a pre-processing step for automated stain separation. It has been applied to Ki67 analysis for stain deconvolution and PI estimation [22] [23] [24].

Although CD has been used for IHC analysis with promising results, for high concentrations of Ki67, the DAB stain is dark and creates non-linear properties in the BL Law [12]. For darkly stained regions, DAB stains are not true absorbers of light and there is light scattering which does not follow the BL law of light absorption [13]. Figure 3 demonstrates how DAB stained images behave in the OD space. The images on the left contain the original Ki67 images, the middle images contain the 3D scatter plot of the OD RGB values (colour-coded), and the right contains the 3D scatter plot projected on the plane of best fit. At low DAB stain concentration levels, DAB and H are linearly separable in the OD space and BL is effective to separate hematoxylin, and DAB stains in this case. However, for higher Ki67 concentrations, light scattering creates non-linearities in the OD space -

brown pixels are spread over a larger region in a nonlinear fashion and there are significant quantization effects at the darkest stain levels. In these cases, the BL law may not be as effective in separating hematoxylin and Ki67 stains and errors in the estimated stain concentration images can occur.

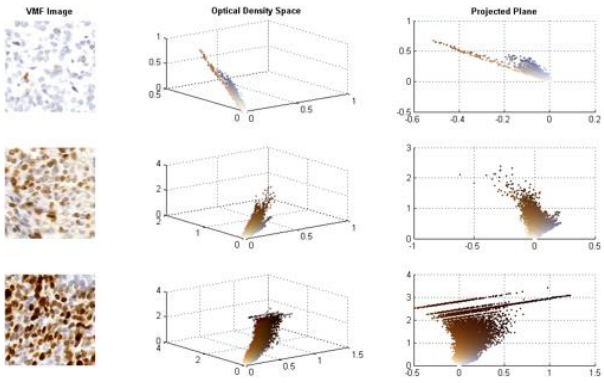

Figure 3: Intensity distribution of Ki67 stained tissue

To overcome these challenges, this paper proposes to use image preprocessing techniques such as, vector median filtering, background subtraction, colour thresholding algorithms and nuclei estimation as preprocessing techniques for IHC specific image. Use machine learning technique, Naives Bayes, to grade the tumor with a Low, Medium and High grade for proliferation index label. A dataset containing, 94 images were preprocessed before features were extracted and then fed into the machine learning algorithm.

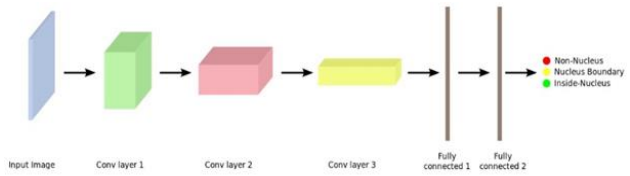

Figure 4: CNN Architecture

IV. Material and Methods

A. Hematoxylin & Eosin Framework for Nuclei Classification

The goal of the proposed framework is to classify nuclei from regions of H&E stained lymph node whole-slides. Lymph node specimens, as stated earlier, are used in the TNM staging system and indicative of breast cancer metastases. In addition, a secondary objective is to analyze the effect of data standardization on the performance of the classifier. In this work, a fully convolutional neural network, adapted from Kumar et al. [29], is used to segment nuclei. Following segmentation, basic geometric characteristics are extracted. Subsequently, the nuclei data is split into training and testing sets, followed by classification using several classifiers.

i) Data Protocol

The data used in this work is adapted from the CAMELYON 2016 Grand Challenge [46]. The original dataset contains 400 whole-slide images of sentinel lymph node from two independent sources collected in Radboud University Medical Center (Nijmegen, the Netherlands), and the University Medical Centre Utrecht (Utrecht, the Netherlands). For the scope of the paper, only 19 1000x1000 region of interests (ROI) images were used as training and testing data. Whole-slide analysis continues to be an expensive task and surpasses the scope of the project.

To represent benign or normal lymph node tissue, 9 1000x1000 ROI images (Figure 5a) were cropped from whole-slides that did not contain metastases. Similarly, to represent malignant lymph node tissue, 11 1000x1000 ROI images (Figure 5b) were cropped from annotated regions of whole-slides which contained metastases.

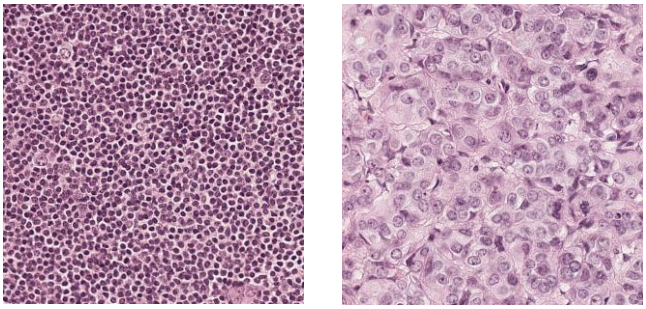

Figure 5: a) Normal ROI
b) Malignant ROI

ii) Convolutional Neural Network

The architecture of the implemented CNN can be seen in Figure 4. The network contains three convolutional layers followed by two fully connected layers. Additional metrics of the architecture can be seen in Table 2 below:

Table 2: CNN Architecture

Layer	**Filter Size**	**Activation**	**Output Size**	**Dropout Rate**
Input	-	-	51x51x3	-
Conv 1	4x4	ReLU	48x48x25	0.1
Pool 1	2x2	Max	24x24x25	-
Conv 2	5x5	ReLU	20x20x50	0.2
Pool 2	2x2	Max	10x10x50	-
Conv 3	6x6	ReLU	5x5x80	0.25
Pool 3	2x2	Max	3x3x80	-
FC 1	-	ReLU	1024	0.5
FC 2	-	ReLU	1024	0.5
Output	-	SoftMax	3	-

In short, during training, patches from RGB images and the corresponding annotation data are applied to the CNN to develop the nucleus-boundary model. It is important to note that for this implementation, the CNN model was trained on publicly available ROIs of various tissues provided by Kumar et al. [29]. Square patches with a size of 51x51 pixels are extracted from the training image with a stride of 7 in which the central pixel belongs to one of the target classes. Once all of the patches are extracted, 90 and 180 degree augmentations are applied to the patches. This is done to ensure that the resulting model is invariant to rotation and to increase the number of training samples. Subsequently, the patches were optimized to ensure that the class labels - inside, outside, and on boundaries of nuclei - were balanced.

The CNN was trained using Pytorch on an Amazon Web Services (AWS) student account. The initial learning rate for the network was 0.001 and was trained over 20 epochs.

iii) Nuclei Isolation and Feature Extraction

After nuclei segmentation for both malignant and benign ROIs, the next steps are to extract binary masks from the predicted segmentation, isolate and label nuclei, and to extract features in preparation for training and testing.

Binary masks of the three classes are extracted from the prediction image by thresholding for unique intensities representative of each class (Refer to Figure 4).

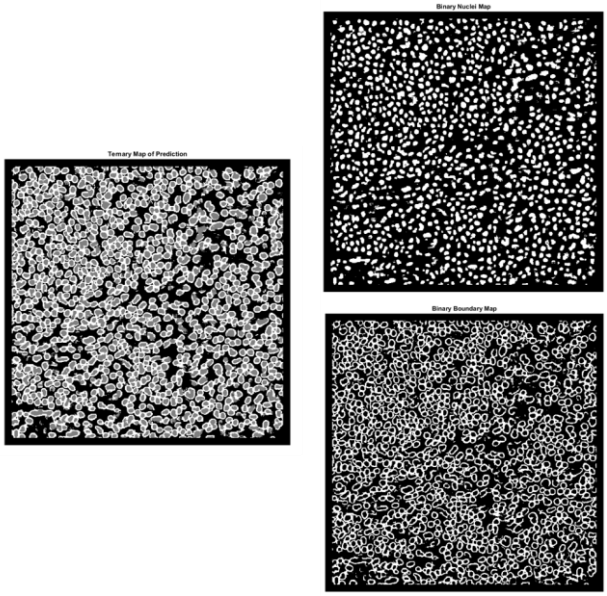

Figure 4: Ternary maps to binary maps

Once the binary maps are obtained, overlapping nuclei are separated by simply subtracting the boundary class image from the nuclei class image. Subsequently, additional morphological processing, such as erosion, and hole filling are applied.

Using built-in Python libraries, nuclei are isolated and individual geometric characteristics are extracted. Area, eccentricity, equivalent diameter, extent, orientation, and inertial tensor eigen-values are the initial selected features.

- ***area:*** number of pixels of a region
- ***eccentricity:*** ratio of focal distance over the major axis length
- ***equivalent diameter:*** diameter of a circle with the same area as the region
- ***orientation:*** angle between the 0th axis and the major axis of the ellipse that has the same moments as the region

- ***inertial tensor eigen-values:*** eigen values of the inertial tensor

Manually labeling of nuclei is then applied. This is possible because each region has a known crop location and source (benign or malignant slide). The resulting feature matrix is then separated into training and testing sets, 80% and 20% respectively.

iv) Nuclei Classification

The feature matrix are applied to numerous classifiers, including: logistic regression, random forests classifier, support vector machine, and naive bayes. All the classifiers were trained using default parameters , with the exception of random forests which used 128 estimators. In addition, the class weight for each classifier was set to *balanced*. This parameter ensures that the class labels are weighted with respect to class proportions. This results in balanced classes during training.

B. Immunohistochemistry Framework for Proliferation Index Calculation

As previously mentioned, the Immunohistochemistry framework will be using various pre-processing steps to help reduce noise and variability to the image along with extracting vital features that will be used for the machine learning classification.

First off, each TMA image will be inputted into the feature extraction preprocessing framework. Before the preprocessing steps three colour spaces will be analyzed, RGB, HSV and LAB. Each channel of the mentioned colour spaces will be extracted, and the mean value will be determined and be utilized as a feature. Once the mean values per channel of each colour space are determined, the preprocessing steps are begun. The background of each TMA will be removed, a VMF filter will be applied, with the hopes of removing noise. Then, the image will use a balanced histogram approach to achieve the thresholding value at which brown and blue separate on the b*channel of the LAB colour space. Finally, the thresholded images will be send into an automatic radius estimator, where each stain, Hematoxylin and Ki67 will have an estimated radius prediction. Each preprocessing step will be explained in depth in the following sections. Using the features extracted, the machine learning classifiers are compared, and the highest accuracy rate is determined. Depending on the accuracy rate of each classifier, the highest value will be implemented and a comparison of the predicted versus actual will be completed.

Incorporating Ki67 biomarker scores into a clinical or research workflow would be valuable to stratify patients for personalized medicine therapies, clinical trials and large research studies. To enable clinical-use and large-scale studies of Ki67, automated PI quantification tools must provide robust and consistent results across all images and patients. For clinical implementation, ideally, algorithms should be vendor-agnostic, and robust to data variability, so they can be deployed in any clinical centre and lab. For Ki67 standards development and research, algorithms must analyze multi-centre, international datasets robustly and reliably. However, sources of noise and variability in Ki67 images cause challenges for automated approaches.

For clinical adoption, automated Ki67 approaches must handle these factors that create variation in noise and colour characteristics to provide maximum reliability. To overcome these challenges, a new Ki67 PI quantification framework that is robust to data variability and different magnification factors is proposed.

i) Vector Median Filter

Due to the noise and variability in intensities/colour in histopathology images, a de-noising procedure is necessary. Histopathology images consist of colour (RGB) images and the application of a scalar de-noising filter to each channel individually may result in the loss of correlation between channels and the skewing of colours [17]. Instead, a vector median filter [17] was applied to the images to make colours appear more uniform in the images by preserving colour proportions and content. The vector median filter calculates the average colour vector within a defined pixel neighborhood for every pixel in the image. By treating the pixels as vectors instead of scalars, the resultant filtered image retains the relationship between colour channels. Since edges are defined across RGB channels in colour images, dependent on the size of the filtering window, edges can be maintained better than the scalar counterpart, while smoothing in low frequency regions. Figure 5 illustrates the visual before and after progress of utilizing the VMF filter.

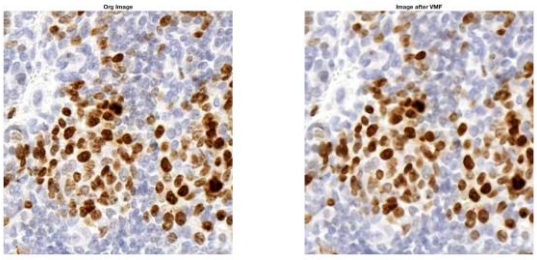

Figure 5: ROI image from a Ki67 stained TMA before (left) and after being processed through the VMF Filter (after)

ii) Background Subtraction

After denoising, a second preprocessing step known as background subtraction was performed to isolate tissue and cell regions and suppress background regions. A local mean adaptive threshold was applied to each image in the grayscale luminance channel because both DAB and H stained pixels are darker than their surroundings allowing application of a single threshold to capture both stains. The local mean adaptive method is an iterative algorithm that measures a local threshold determined by the mean grayscale intensities of a neighbourhood [18]. This method was preferred over other popular methods such as Otsu's method [19] because it is local and can adapt to different tissue compositions.

iii) Colour Separation via Thresholding

Many frameworks use the CD method for colour separation, which was mentioned previously. The CD method cannot be incorporated into IHC analysis due to the law of Beer Lambert. Hence, this proposed framework uses a thresholding algorithm to determine at which value of b* (the threshold) that image converges from blue to brown.

Manual nuclei counting, and proliferation index estimation performed by a pathologist relies on the human visual system's (HVS) perception of staining intensity and colour characteristics. Therefore, the proposed colour separation method utilizes the perceptually linear L*a*b* colour space to separate hematoxylin and DAB stains in a way that mimics human perception and discrimination of blue and brown colours. The figure below, figure 6, visually shows an illustration of a TMA image outputted from the L*a*b colour space, specifically isolated b*channel. The two preprocessing steps previously mentioned, background subtraction and the VMF filter ensure that the calculated thresholded value is indeed accurate for the machine learning algorithm to process.

The lightly contrasted colours represent the positive b values (yellow, brown) and the darker colours represent the negative b values (blue, purple).

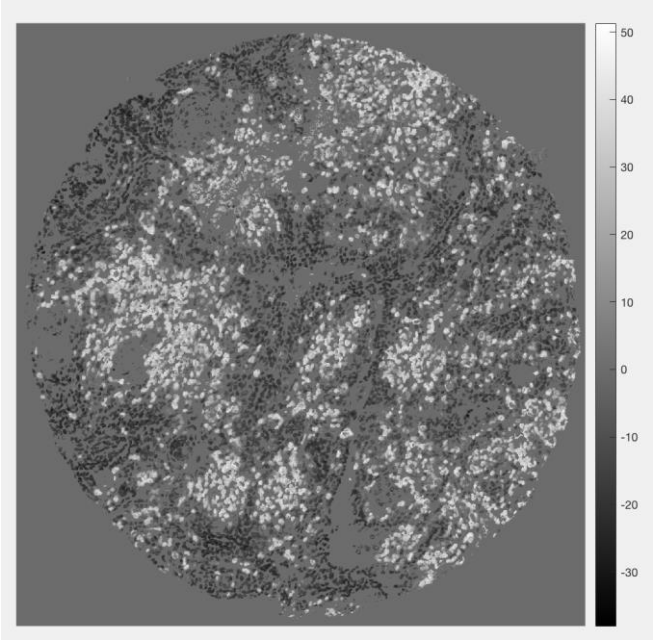

Figure 6: Ki67 stained TMA from the bchannel

For colour separation of hematoxylin and Ki67, we have found a novel way to differentiate colour intensities associated with each stain by focusing on the yellow/blue characteristics within the b* chrominance channel of IHC images. Possible b* channel intensities for all RGB vector combinations lie on the approximate range [-128, 127] where the negative region represents "pure" blue, zero represents no colour content, and the positive region defines "pure" yellow. Colours of interest in hematoxylin and Ki67 images are blue and brown, where brown is described by yellow content (dominating red and green components) and the blue colour associated with hematoxylin is defined in the negative blue region.

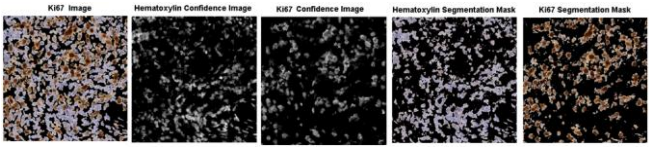

Figure 6: ROI images from a Ki67 stained TMA, the different thresholding algorithms used to obtain an accurate colour separation method.

In order to separate IHC images into hematoxylin and IHC components, colour separation is achieved through application of an adaptive threshold value T on the b* channel. The result is two confidence images which represent the hematoxylin and Ki67 content respectively. Using the b channels histogram and an adaptive thresholding method, the optimal threshold of brown versus blue is found. Figure 7 illustrates this histogram of an ROI. Using this the threshold value is found. This threshold value is extracted as a feature for each patient's TMA, as it is an important piece of information regarding the stain variation of the image. The thresholding value per TMA image changes as the stain variation and staining protocols differ.

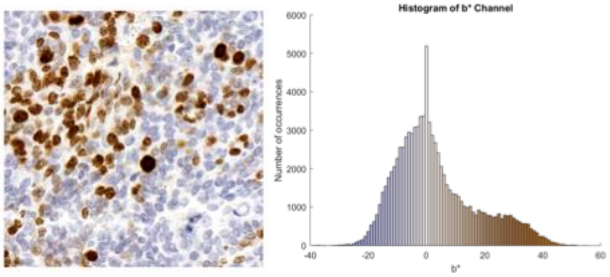

Figure 7: Colour mapped histogram of a TMAs region of interest, this is incorporated to precisely determine the threshold between brown and blue in the b*channel

iv) Estimated Radius

The F1 performance metric was used to evaluate the effectiveness of the nuclei detection algorithm with an automatic cell radius estimator. Since the proposed framework separates nuclei content into two colour separated images, the F1 score and cell radius estimate were determined for each stain independently. Cell radii estimates were validated using a circular validation window with a diameter equal to 2X the automatically determined radius. To examine the results of the cell radius, estimate and balanced threshold T, the F1 score for various user defined cell radius estimates were observed for three different thresholding methods. The balanced threshold was compared to Otsu's method [19] and the natural boundary b*=0 discussed previously. The results of the experimentation are summarized in figure 8, where the F1 score performance for both Ki67 and hematoxylin are shown against the automatic cell radius estimator and possible user selected parameters.

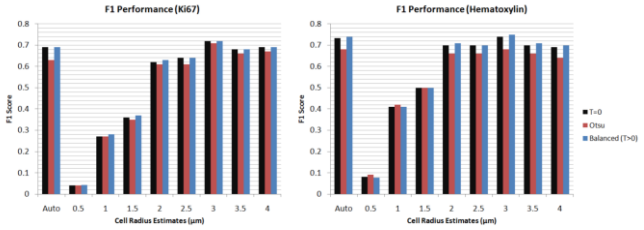

Figure 8: F1 score vs performance for both stains, Ki67 (left) and Hematoxylin (right) – different thresholding techniques were applied to determine the optimal threshold

The ideal arbitrarily selected estimates yielded average F1 scores of 0.72 and 0.75 for Ki67 and hematoxylin respectively while the automatic cell radius estimate produced F1 scores of 0.69 and 0.74 respectively. Although the cell radius estimator achieved slightly lower results, any deviation from the ideal radius estimate as a result of manual error or user subjectivity would result in suboptimal performance. Therefore, the cell radius estimator is able to automatically achieve comparable results to ideal radius selections without relying on manually defined parameters.

v) Naive Bayes Machine Learning Algorithm

After the preprocessing techniques were completed, the features decided upon were exported into excel, along with the cancer label. The data was split into 80% training data and 20% testing data. The features decided upon were, mean values of each colour space that achieved promising results. The mean of each channels colour space (HSV, LAB and RGB) the thresholding value of each patients TMA along with the estimated radius of the hematoxylin stained & Ki67 stained nuclei, were all used as features.

Using the Exploratory data analysis, the features were normalized. In order to decide the type of algorithm to use, a scatter matrix of the different features, with relationship between each other, were obtained, figure 9. The matrix illustrates the features relationship in correlation to the labels, with a histogram along the diagonal. The scatter plot of the low grade labels is green, medium grade in yellow and high grade in red. This was used to confirm that the features used definitely have a relationship with the graded labels. Then each classification machine learning algorithm was

computed, and the accuracy was outputted, as shown in Table I.

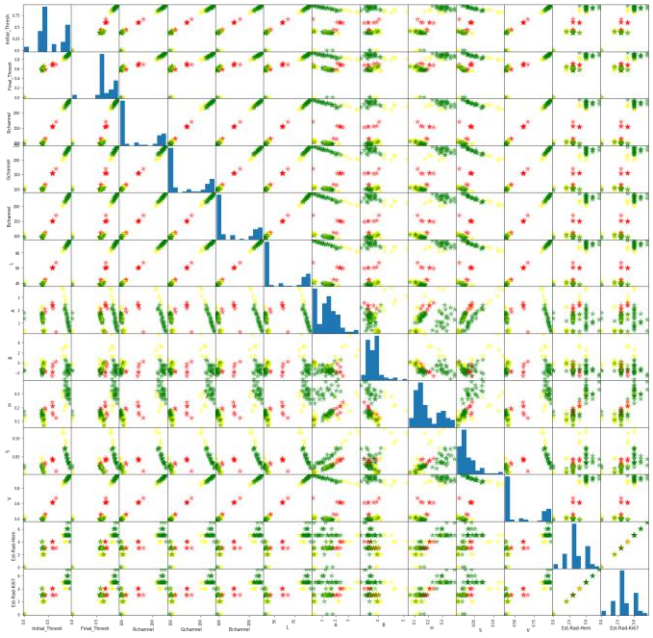

Figure 9: Matrix of each feature with the three corresponding labels and a histogram along the diagonal (Green = Low, Yellow = Medium, Red = High)

Looking at the matrix, it is clear that the features have an evident relationship with the graded tumors. Hence, this matrix gives us the "go-ahead" to continue with using the extracted features. When comparing the different algorithms, Naive Bayes had the highest accuracy rate of 73.7%, hence this classification model was used.

VI. Results

C. Lymph node nuclei classification

Nuclei classification results on unstandardized data are depicted in Table 3a. The best performing classifier with respect to accuracy was both Random Forest and Naive Bayes classifiers, with an accuracy of 72%. However, SVM and Logistic Regression classifiers demonstrated a better true positive rate for malignant nuclei, 58% and 51% respectively. Random forest and naive bayes demonstrated high true positive rates for benign nuclei, but also demonstrated high false positive rates at 60% and 66% respectively. (Refer to appendix Figure 1 for confusion matrices)

Table 3: Nuclei classification results

Classifier	Accuracy	
	a) Unstandardized	b) Standardized
Logistic Regression	0.66	0.69
Random Forest	0.72	0.70
Support Vector Machine	0.67	0.67
Naive Bayes	0.72	0.70

When comparing the performance of the classifiers on unstandardized to standardized data, unstandardized data demonstrates slightly higher accuracy. However, the performance of the classifiers actually increased for correctly identifying malignant nuclei. Logistic regression, Random Forest, and SVM, all showed an increased true positive rate for malignant nuclei at 59%, 42%, and 65% respectively. (Refer to appendix - Figure 1 for confusion matrices)

Before coming to the conclusion of utilizing the Naive Bayes Machine Learning Algorithm, various machine learning algorithms were used, and the accuracy was calculated, as shown in the table below. Navies Bayes had the highest accuracy; hence it was implemented. Since, the dataset was small, 75 images were designated for training and 19 for testing. The confusion matrix below illustrates the accuracy of the model. The labels were, Low, Medium and High grade.

Table 4: Performance Comparison to Machine Learning Classifiers without Cross validation.

	Gaussian Naïve Bayes	K-Nearest Neighbor	Logistic Regression	Stochastic Gradient Descent	Support Vector
Accuracy	0.7368	0.6842	0.5789	0.5789	0.5263

Table 5: Confusion Matrix : Navies Bayes

Confusion Matrix	**Low**	**Med**	**High**
Low	**9**	1	0
Med	4	**4**	0
High	0	0	**1**

The diagonal along the confusion matrix illustrates to the user how many predictions were accurately executed. For the tested dataset, there was an 26% error rate, 5/19 of the predictions were incorrect.

i) Cross Validation Implementation

Machine Learning classifiers are great tools for computational pathology, when there is a substantial amount of data to accurately train the algorithm. Unfortunately, the utility of Ki67 is still controversial [14] and therefore large datasets with labeled grades per image are very hard to come by. A common method used, especially in clinical settings, to resolve this issue is using cross-validation (CV) [46][47]. This divides the data into *k* sub-datasets and allows for the training and testing data to be fully utilized to help overcome the issue of using a small dataset. Using the combined dataset, each TMA image was labeled with the PI range, i.e. low(<25%), medium (25 – 75%), and high (> 75%) levels of proliferation activity.

Table 5: Performance Comparison to Machine Learning Classifiers with *k=10* Cross validation.

K=10	Gaussian Naïve Bayes	K-Nearest Neighbor	Logistic Regression	Stochastic Gradient Descent	Support Vector
Mean Accuracy	0.62667	0.80778	0.66889	0.656667	0.58556

The accuracy of the following classifiers was explored, as shown in Table II, Gaussian Naïve Bayes (NB), K-Nearest Neighbor (KNN), Logistic Regression (LR), Stochastic Gradient Descent (SGD) and Support Vector (SVM) Classifier. Each classification algorithm used the exact dataset, 94 TMA images, and the same extracted features, in order to maintain consistency throughout the comparison. CV was used to accurately analyze the different classifiers, when a dataset is significantly small in size, CV can be used to efficiently evaluate the accuracy. In order to eliminate the trade-off between number of training and testing data between the datasets, CV method was used for each classifier mentioned in Table 4, the results of CV can be seen in Table 5. The data was partitioned into *k=10* bins of equal size and *k* separate learning experiments were conducted, where a single randomized sub-dataset was utilized as the testing data. The accuracy rates of each *k* runs were stored, and the mean value and standard deviation were obtained. This ensures that the classifier's prediction is compared to the test data and the resulting accuracy is precise. In order to see the complete range of accuracy's per classifier, Figure 10 illustrates a boxplot for each type of classifier.

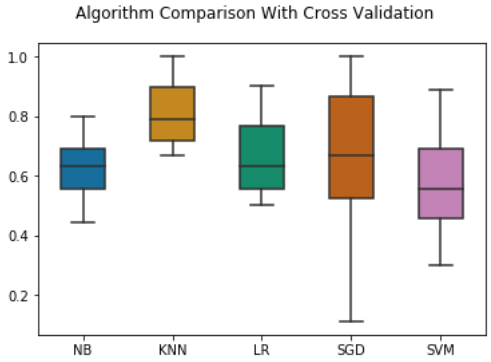

Figure 10: Boxplot representation of the ranges of accuracies for each classifier combined with the Cross Validation method.

The boxplot represents the variation and accuracy of each classifier for *k=10* runs . From Figure 10, it is evident that KNN does the best in terms of highest, least varying accuracy. Hence, instead of using navies bayes, KNN would be implemented. Before implementing KNN, the ideal value of k neighbors must be analyzed. After analyzing the accuracy rates from *k=1 to k=20*, figure 11, illustrates the idea value of k. Using five or six neighbors achieves an accuracy of approximately 80%, therefore, k =5 was used.

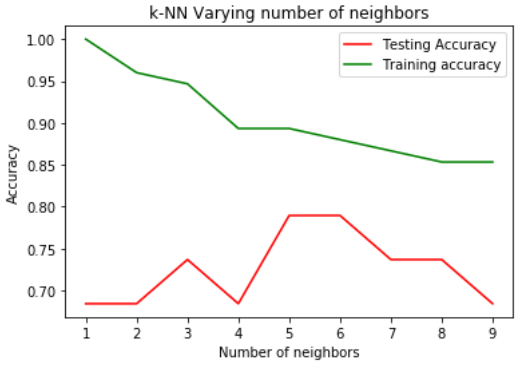

Figure 11: Analyzing the number of neighbors used to achieve an ideal accuracy rate

With the help of CV to precisely calculate the accuracy of the different methods using *k=10* subsets, we found that Navies Bayes is not as accurate as KNN method. Table 6 represent the confusion matrix obtained from the KNN model with the optimal number of k as 5.

Table 6: Confusion Matrix: KNN

Confusion Matrix	**Low**	**Med**	**High**
Low	***10***	*0*	0
Med	4	**4**	0
High	0	0	**1**

There is a slight implemented on the number of correctly classified grades, Navies Bayes classified 14 correctly, whereas KNN classified 15 correctly. Furthermore, as mentioned previously, Navies Bayes had an error rate of 26% and KNN with k=5, obtained an error rate of 21%.

VII. Discussion

D. *Classification of benign and malignant lymph node nuclei*

The classifier which demonstrated the highest accuracy in predicting non-cancerous and malignant nuclei were Random Forest and Naive Bayes. These classifiers achieved the highest accuracies on both the standardized and unstandardized datasets. However, as previously stated, these classifiers also showed the highest false positive rate and low true positive rates for predicting malignant nuclei. The over detection of malignant nuclei could result in unnecessary treatment and negative patient outcomes. For instance, over detection of cancerous tissue can result in unnecessary excision of healthy tissue.

In analyzing the results of the classifiers, accuracy could have been affected by the quality of the nuclei segmentation. Because ground truth data were not available for the segmentation, the paper is limited in quantifying the accuracy of segmentation on the standardized and unstandardized datasets. Though this may be true, it is obvious that color normalization affects the performance of the demonstrated classifiers. The improvement in true positive rate for colour normalized malignant nuclei demonstrates that colour variability impacts the ability for supervised learning frameworks to effectively classify highly variable data.

E. *Immunohistochemistry – Proliferation Index Calculation*

Overall, the Machine learning algorithm obtained an accuracy rate of 80%. The use of CV helped determine the optimal classier that should be used in order to obtain the highest accuracy with lowest variation. Switching from Navies Bayes to KNN gave the classification accuracy an approximate 10% increase in accuracy rate.

Three areas of improvements are to utilize more features, using the values outputted from a histogram for each image, implementing a cross validation method and using more training data with an even distribution. The bar graph below illustrates the distribution of Low, Medium and High grade images. The labels were unfortunately, not distributed evenly, due to the limitation of obtained images with Proliferation Index values.

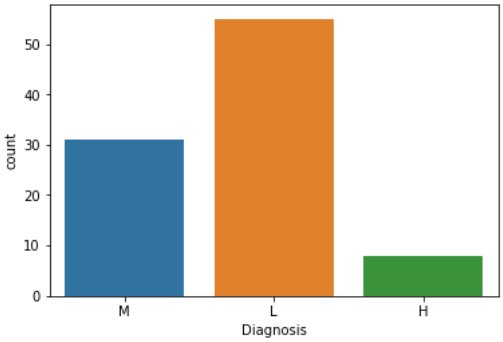

Figure 12: Distribution of Labeled Dataset

VIII. Conclusion

Stains such as hematoxylin and eosin (H&E), and immunohistochemical (IHC) stains such as Ki67, are critical components to achieving an accurate diagnosis and prognosis, specifically for breast cancer patients. Using automated grading systems can decrease human error and increase pathologist workflow, which is essential for improving a patient's quality of care. With the addition of machine learning, we have the ability to explore different proxy measurements, features and colour spaces in order to achieve minimal

error and higher accuracy rates within digital histopathology.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. Siegel, L. Torre and A. Jemal, "Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *Cancer Journal for Clinicians*, vol. 68, pp. 394-424, 2018.

[2] "Breast Cancer," 16 October 2018. [Online]. Available: https://www.breastcancer.org/symptoms/types/idc. [Accessed 22 November 2018].

[3] S. Walters, C. Maringe, J. Butler, B. Rachet, P. Barett-Lee, J. Bergh and J. Boyages, "Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: a population-based study," *British Journal of Cancer*, vol. 108, pp. 1195-1208, 2013.

[4] G. Hortobagyi, J. Connolly, C. D'Orsi, S. Edge, E. Mittendorf, H. Rugo, L. Solin, D. Weaver, D. Winchester and A. Giuliano, "Breast," in *AJCC Breast Cancer Staging System*, Illinois, The American College of Surgeons, 2018, pp. 1-50.

[5] S. Ismail, A. Colclough, J. Dinnen, D. Eakvins and D. Evans, "Observer variation in histopathological dianosis and grading of cervical intraepithelial neoplasia," *British Medical Journal*, vol. 298, pp. 707-710, 1989.

[6] A. Andrion, C. Magnani, P. Betta, A. Donna and A. Mollo, "Malignant mesothelioma of the pleura: interobserver variability," *Journal of Clinical Pathology*, vol. 48, pp. 856-860, 1995.

[7] M. Niethammer, D. Borland, J. Marron, J. Woosley and N. Thomas, "Appearance normalisation of histology slides," *Machine Learning in Medical Imaging*, pp. 58-66, 2010.

[8] M. Mackenko, M. Niethammer, J. Marron, D. Borland, J. Woosley, X. Guan, C. Schmitt and N. Thomas, "A Method for Normalizing Histology Slides for Quantitative Analysis," in *IEEE International Symposium on Biomedical Imaging* .

[9] D. Magee, "Colour Normalisation in Digital Histopathology Images," Proc. Optical Tissue Image Anal. Microsc., Histopathol. Endosc., pp. 20-24, 2009.

[10] A. Khan, N. Rajpoot, D. Treanor and D. Magee, "A Nonlinear Mapping Approach to Stain Normalization In Digital Histopathology Images Using Image-Specific Color Deconvolution," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729-1738, 2014.

[11] D. M. Schonk et al., "Assignment of the gene(s) involved in the expression of the proliferation-related Ki-67 antigen to human chromosome 10," Hum Genet, vol. 83, no. 3, pp. 297–299, Oct. 1989.

[12] T. Mungle, S. Tewary, I. Arun, B. Basak, A. Sanjit, A. Rosina, C. Sanjoy, K. Asok, and C. Chandan, "Automated characterization and counting of Ki-67 protein for breast cancer prognosis : A quantitative immunohistochemistry approach," Comput. Methods Programs Biomed., vol. 139, pp. 149–161, 2017.

[13] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," Anal. Quant. Cytol. Histol., vol. 23, pp. 291–299, 2001.

[14] S. Di Cataldo, E. Ficarra, A. Acquaviva, and E. Macii, "Automated segmentation of tissue images for computerized IHC analysis," Comput. Methods Programs Biomed., vol. 100, no. 1, pp. 1–15, 2010.

[15] J. Konsti, M. Lundin, H. Joensuu, T. Lehtimäki, H. Sihto, K. Holli, T. Turpeenniemi-hujanen, V. Kataja, L. Sailas, J. Isola, and J. Lundin, "Development and evaluation of a virtual microscopy application for automated assessment of Ki-67 expression in breast cancer," BMC Clin. Pathol., vol. 11, no. 1, p. 3, 2011.

[16] F. Varghese, A. B. Bukhari, R. Malhotra, and A. De, "IHC Profiler : An Open Source Plugin for the Quantitative Evaluation and Automated Scoring of Immunohistochemistry Images of Human Tissue Samples," PLoS One, vol. 9, no. 5, 2014.

[17] J. Van der Laak, M. Pahlplatz, A. Hanselaar, and P. Wilde, "Hue-Saturation-Density (HSD) Model for Stain Transmitted Light Microscopy," Cyto,etry, vol. 39, pp. 275–284, 2000.

[18] C. R. Taylor and R. M. Levenson, "Quantification of immunohistochemistry — issues concerning methods , utility and semiquantitative assessment II," Histopathology, vol. 49, pp. 411–424, 2006.

[19] M. Uhlén et al., "Tissue-based map of the human proteome," Science, vol. 347, no. 6220, p. 1260419, Jan. 2015.

[20] S. Krishnamurthy et al., "Multi-Institutional Comparison of Whole Slide Digital Imaging and Optical Microscopy for Interpretation of Hematoxylin-Eosin–Stained Breast Tissue Sections," Archives of Pathology & Laboratory Medicine, vol. 137, no. 12, pp. 1733–1739, Aug. 2013.

[21] R. S. Cook, P. Kay, and T. Regier, "The World Color Survey Database: History and Use," p. 22.

[22] C. M. van der Loos, "Multiple Immunoenzyme Staining : Methods and Visualizations for the Observation With Spectral Imaging," J. Histochem. Cytochem., vol. 56, no. 4, pp. 313–328, 2008.

[23] W. Koziey-Kronas, "PathcoreSedeen," Pathcore.

[24] "Fiji: an open-source platform for biological-image analysis | Nature Methods." [Online]. Available: https://www.nature.com/articles/nmeth.2019. [Accessed: 29-Nov-2018].

[25] C. Lu, D. Romo-Bucheli, X. Wang, A. Janowczyk and S. Ganesan, "Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers," Springer Nature, 2018.

[26] M. Veta, A. Huisman, M. Viergever and P. van Diest, "Marker-Controlled Watershed Segmentation of Nuclei in H&E Stain Breast Cancer Biopsy Images," IEEE, 2011.

[27] J. Vink, V. Van Leeuwen and C. Van Deurzen, "Efficient nucleus detector in histopathology images," Journal of Microscopy, vol. 249, no. 2, pp. 124-135, 2013.

[28] W. Cai, S. Chen and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local informative information for segmentation," Pattern Recognition, vol. 40, pp. 825-838, 2006.

[29] N. Kumar, V. Ruchika, S. Sharma, S. Bhargava, A. Vahadane and A. Sethi, "A Dataset and Technique for Generalized Nuclear Segmentation for Computational Pathology," IEEE Transactions On Medical Imaging , pp. 1-11, 2017.

[30] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolution Networks for Biomedical Image Segmentation," Computer Vision and Pattern Recognition, pp. 1-8, 18 May 2015.

[31] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images," IEEE transactions on medical imaging, vol. 35, no. 1, pp. 119–130, 2016.

[32] H. Sharma, N. Zerbe, I. Klempert, O. Hellwhich and P. Hugnagl, "Deep convolution neutral networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology," Computerized Medical Imaging and Graphics, vol. 61, pp. 2-13, 2017.

[33] W. Rawat and Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," Neural Computation, vol. 29, p. 2352=2449, 2-17.

[34] P. Naylor, M. Lae, F. Reyal, and T. Walter, "Nuclei ´ segmentation in histopathology images using deep neural networks," in Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on. IEEE, 2017, pp. 933–936.

[35] Xing, F., Xie, Y., Yang, L., 2016. An automatic learning-based framework for robust nucleus segmentation. IEEE Trans Med Imaging 35 (2), 550–566.

[36] A. Ruifrok and D. Johnston, "Quantification of histchemical staining by color deconvolution," Anal Qual Cytol Histol, vol. 23, pp. 291-299, 2001.

[37] G. Landini, "Colour Deconvolution," ImageJ, 29 April 2017. [Online]. Available: https:// imagej.net/index.php?title=Colour_Deconvolution& action=info. [Accessed 12 November 2018].

[38] M. Dowsett, T. O. Nielsen, R. A. Hern, J. Bartlett, R. C. Coombes, J. Cuzick, M. Ellis, N. L. Henry, J. C. Hugh, T. Lively, L. Mcshane, S. Paik, F. Penault-llorca, L. Prudkin, M. Regan, J. Salter, C. Sotiriou, I. E. Smith, G. Viale, J. A. Zujewski, and D. F. Hayes, "Assessment of Ki67 in Breast Cancer : Recommendations from the International Ki67 in Breast Cancer Working Group," J Natl Cancer Inst, vol. 103, no. 22, pp. 1656–1664, 2011.

[39] P. Jalava, T. Kuopio, T. Kotkansalo, P. Kronqvist, and Y. Collan, "Ki67 immunohistochemistry : a valuable marker in prognostication but with a risk of misclassification : proliferation subgroups formed based on Ki67 immunoreactivity and standardized mitotic index," Histopathology, vol. 48, pp. 674–682, 2006.

[40] S. M. Veronese, M. Gambacorta, O. Gottardi, F. Scanzi, M. Ferrari, and P. Lampertico, "Proliferation Index as a Prognostic Marker in Breast Cancer," Cancer, vol. 72, no. 12, pp. 3926–3931, 1993.

[41] E. Gudlaugsson, I. Skaland, E. A. M. Janssen, R. Smaaland, Z. Shao, A. Malpica, F. Voorhorst, and J. P. A. Baak, "Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer," Histopathology, vol. 61, no. 6, pp. 1134–1144, 2012.

[42] R. Shui, B. Yu, R. Bi, F. Yang, and W. Yang, "An Interobserver Reproducibility Analysis of Ki67 Visual Assessment in Breast Cancer," PLoS One, vol. 10, no. 5, pp. 1–11, 2015.

[43] A. Khademi, "Image Analysis Solutions for Automatic Scoring and Grading of Digital Pathology Images," Can. J. Pathol., vol. 5, no. 2, pp. 51–55, 2013.

[44] C. Albarracin and S. Dhamne, "Ki67 as a Biomarker of Prognosis and Prediction : Is it Ready for Use in Routine Pathology Practice ?," Curr Breast Cancer Rep, vol. 6, pp. 260–266, 2014.

[45] "Healthcare FAQs." [Online]. Available: https://digitalpathologyassociation.org/healthcare-faqs. [Accessed: 09-Dec-2018].

[46] "Grand-Challenged," [Online]. Available: https://camelyon16.grand-challenge.org/Data/.

[47] R. Love, P. B. Mangu, L. Mcshane, K. Miller, C. K. Osborne, and S. Paik, "American Society of Clinical Oncology / College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer," J. Clin. Oncol., vol. 28, no. 16, pp. 2784–2795, 2010.

[48] S. Di Cataldo, E. Ficarra, A. Acquaviva, and E. Macii, "Automated segmentation of tissue images for computerized IHC analysis," Comput. Methods Programs Biomed., vol. 100, no. 1, pp. 1–15, 2010.

[49] J. Konsti, M. Lundin, H. Joensuu, T. Lehtimäki, H. Sihto, K. Holli, T. Turpeenniemi-hujanen, V. Kataja, L. Sailas, J. Isola, and J. Lundin, "Development and evaluation of a virtual microscopy application for automated assessment of Ki-67 expression in breast cancer," BMC Clin. Pathol., vol. 11, no. 1, p. 3, 2011.

[50] F. Varghese, A. B. Bukhari, R. Malhotra, and A. De, "IHC Profiler : An Open Source Plugin for the Quantitative Evaluation and Automated Scoring of Immunohistochemistry Images of Human Tissue Samples," PLoS One, vol. 9, no. 5, 2014.

Appendix - Additional Figures

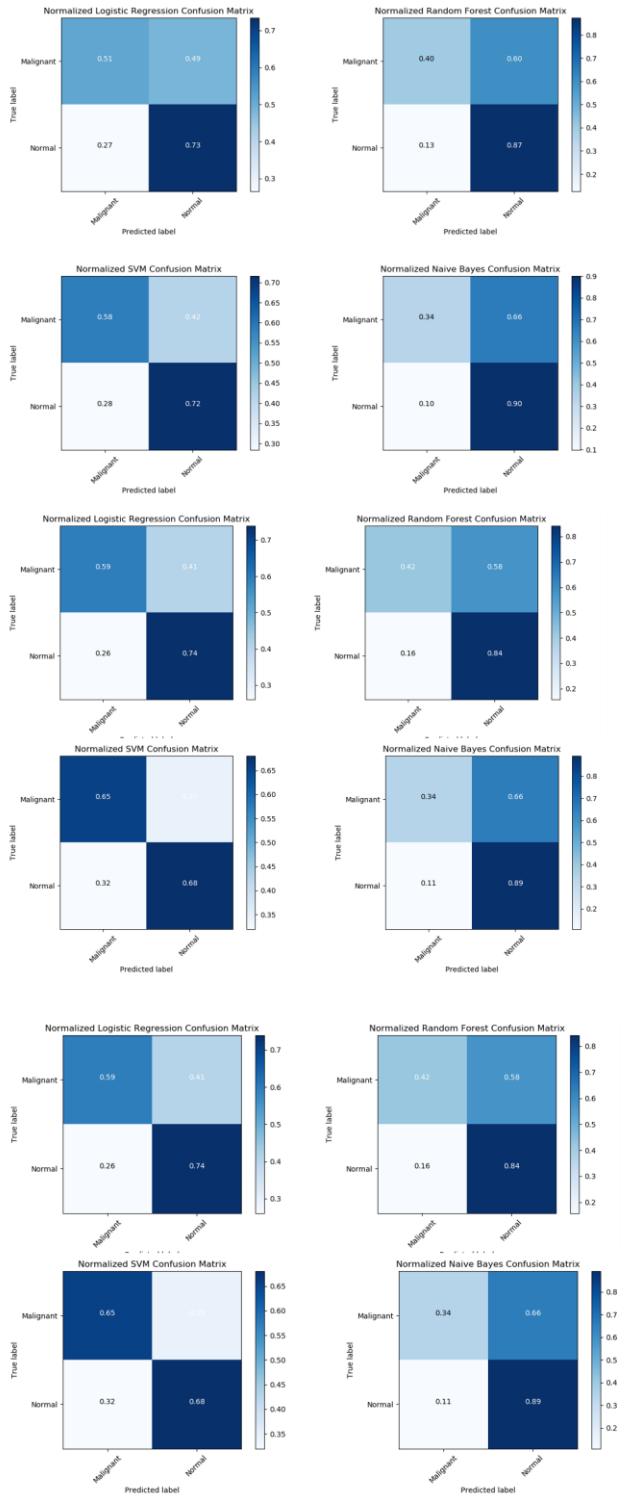

Figure 1: Comparing confusion matrices for unstandardized (top) and standardized data (bottom)