## Introduction

The field of pathology is concerned with identifying and understanding the biological causes and effects of disease through the study of morphological, cellular, and molecular features in samples of tissue. Examination of prepared tissue samples is typically performed by a pathologist looking through an optical microscope. Thin slices of patient tissue samples are stained with contrast agents and fluorescently labelled molecular markers before viewing, which provides contextual information useful in diagnosis.

A grading system is used to communicate the severity of the ailment present in a particular tissue sample, where a higher grade indicates a worse prognosis [1]. A histological grade is assigned based on features that may be quickly identified by the examining pathologist. Here, we restrict our focus to features associated with cancerous tumors. One so called hallmark of cancer is uninhibited cellular replication [4, 5], which leads to increased mitotic activity and poorly differentiated tissue. Increased mitotic activity is indicated by counting the number of cells undergoing the late stages of cellular division, which is indicated by the presence of additional chromosomes in the nucleus. Poorly differentiated tissue does not have identifiable structures (like membranes or blood vessels), and instead looks like an amorphous matrix of cells. The cells in a poorly differentiated tumor have irregular size and shape, with no clear indication of the intended function of the cell.

The histological grade of a tumor has been associated with patient survival [2]. As a result, the grading system has remained the gold standard for the past 80 years. However, the histological grading system has a few important shortcomings. First, the feature set available for diagnosis and prognosis is limited, since the features must be reliably identified from a visual inspection of the tissue sample. This leaves an enormous number of potentially relevant features unexamined. Second, the grade assigned to the same tissue sample may vary by pathologist, due to ambiguity in the features themselves or the variation in pathologist experience, performance, etc [6]. Third, pathology labs are under enormous pressure to analyze large volumes of slides while maintaining diagnostic accuracy. While the time required for a trained pathologist to examine a slide depends on the stains used during slide preparation, the standard hematoxylin and eosin stain (cell nuclei blue, cytoplasm pink) takes a few minutes to examine (specialized stains may require fifteen minutes or more).

The issues associated with the histological grading system may be addressed using digitally assisted analysis of tissue samples, which is enabled by the rapid improvement in imaging quality, memory size, and computational power over the past twenty years. Digital processing of histological data enables the examination of arbitrary features previously unsuited for visual analysis. These new features may have diagnostic or prognostic value in the clinical setting, as well as providing direction for basic research. Furthermore, digital processing of histological

data enables consistent feature definition and analysis, reducing misinterpreted or miscommunicated analysis results. Finally, digital processing of histological data enables greater throughput and allows pathologists to focus on higher level tasks (rather than counting nuclei, for example).

**Objective**

The impact of machine learning on histological analysis was recently reported by Andrew Beck et. al. in *Science Translational Medicine* [3]. In this work, a model was trained to classify breast cancer patients as "high-risk" or "low-risk" of death within the next five years using histological slides from the Netherlands Cancer Institute (NKI, n = 248) and the Vancouver General Hospital (VGH, n = 238). The binary classifier was trained using L1-regularized logistic regression. Preprocessing was used to remove background and segment the image into smaller sections called superpixels, which were classified as either epithelial tissue (forms a membrane) or stromal tissue (forms a matrix in which epithelial cells are embedded) using a binary classifier trained using labeled histology data. Relational features between superpixels were additionally calculated (intensity, size, shape…) to form the prognostic classifier. In total 6642 features were used, however, only 11 of those features were required to produce a robust prognostic model.

The objective of this project is to replicate the work of Beck et. al. to provide a personal foundation for future work in this area. The goals are to 1) implement superpixel segmentation, 2) identify nuclei within each superpixel, 3) replicate the epithelial vs stromal classifier, and 4) construct the top three dominant features used in the prognostic classifier. Essentially, repeating the machine learning component of [3] to verify my understanding. The histological slide data used in [3] is located in the Stanford Tissue Microarray Database (TMAD). The raw images and images with labeled epithelial and stromal regions are available for both the NKI and VGH data sets.

**Results**

The results shared here concern the classifier tasked with identifying regions of background/background tissue, epithelial tissue, and stromal tissue. Two approaches to implement this classifier and are outlined visually in Figure 1. Both approaches are exercises in supervised learning, but differ in image segmentation, the types of features used, and the algorithms used to train the models.

In Method A, 150 raw histology images are separated into background/epithelium/stroma using pixel masks generated through k-means operating on the images that have pathologist labeled sections of epithelial and stromal tissue. The separated tissue sections are then segmented into smaller objects using a watershed algorithm, which finds borders between local minima in the grayscale version of the histology image. Intensity features are calculated in each object and used to train two classifiers through logistic regression. The first classifier separates background tissue

from tissue that is either epithelial or stromal. Given the output of the first classifier predicts an object is not background, the second classifier makes the final separation between epithelial tissue and stromal tissue. The resulting model is tested using an addition 50 histology images that were not used in training.

In Method B, 30 raw histology images are segmented into a grid of 15 pixel x 15 pixel boxes. For training, each box is labeled as background, epithelium, or stroma using the same k-means separation as in Method A. A box may overlap multiple tissue regions, so each box is assigned to

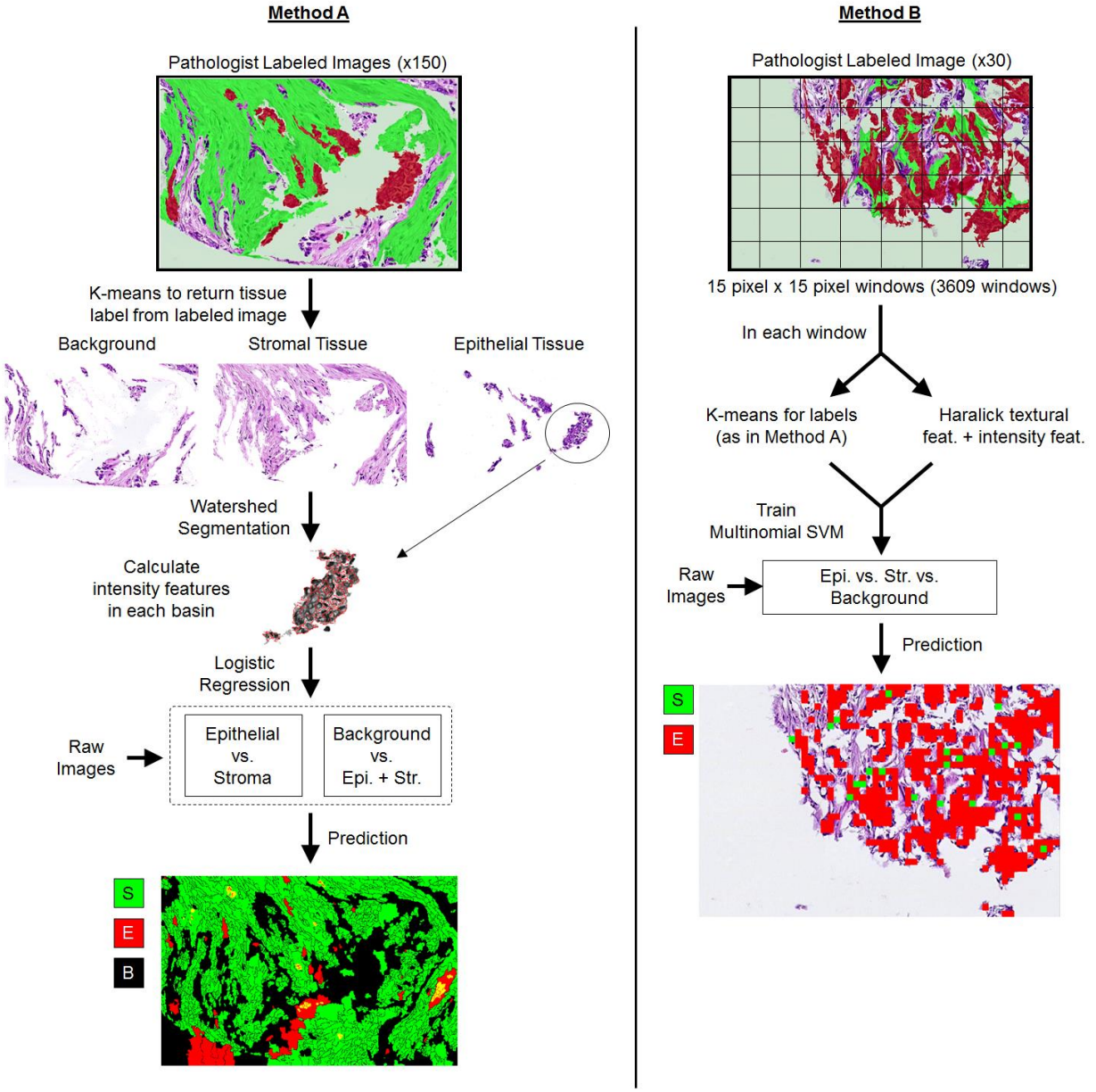



Figure 1. Visual outline of the two methods used to classify background, epithelial tissue, and stromal tissue in raw histology images.

the region of maximum overlap. Textural features based on [7] are calculated for each box in addition to intensity features similar to those used in Method A. The labels and features of each box are used to train a multinomial classifier using the SVM algorithms available through [8]. The models produced by Method B are tested using leave-one-out cross validation on the 30 images.

Example classifications are shown in Figure 1 for both Methods A and B. Qualitatively, the classification produced by ModB more closely represents the image under consideration (top of Figure 1). ModA is a particularly poor classifier, since it predicts nearly all tissue sections are stromal, with a few background sections even being labeled as epithelial. The qualitative results observed in Figure 1 are confirmed quantitatively in Figure 2, where the classification accuracy is shown for ModA (Fig. 2a) and ModB (Fig. 2c). The generalization accuracy of ModA is better than the training error of ModA, which is attributed to the observation that background tissue

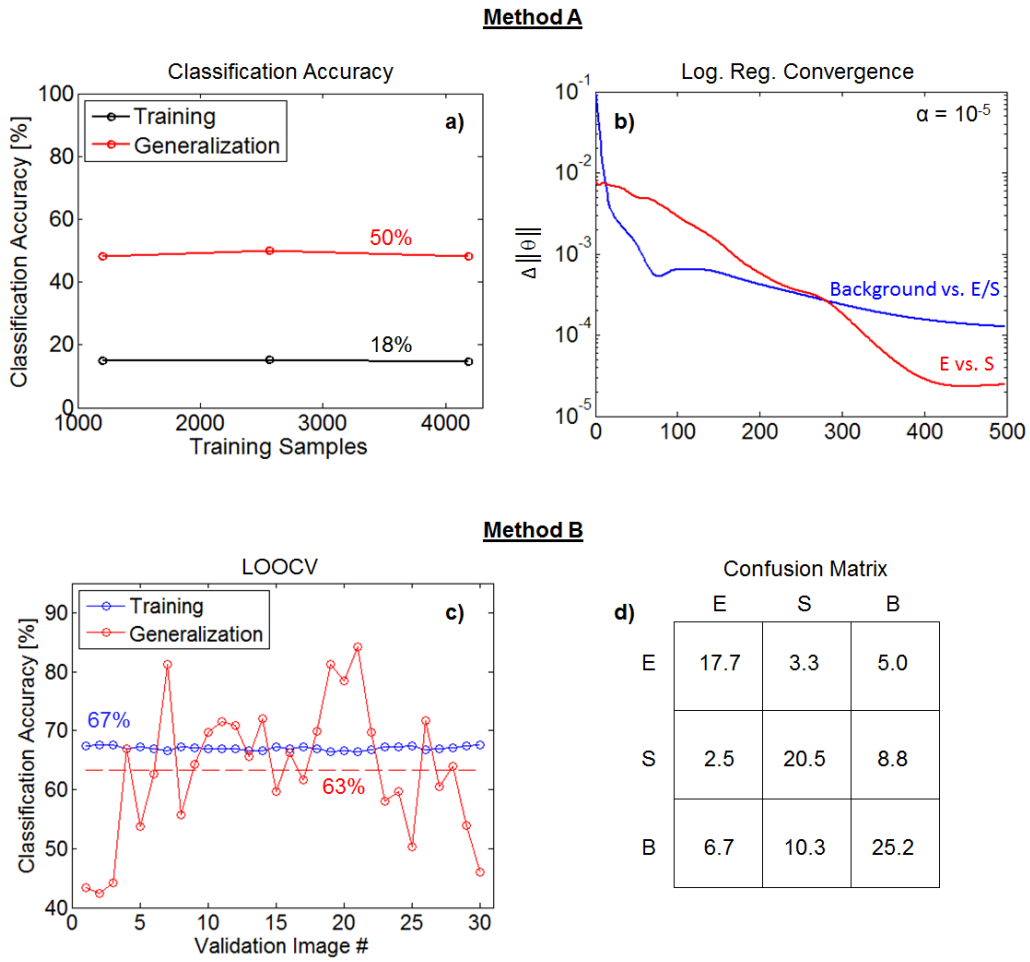
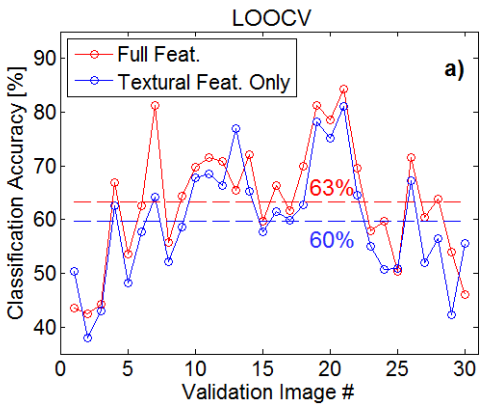



Figure 2. a) Classification accuracy of classifier produced by Method A and b) illustration of convergence criterion for logistic regression algorithm. c) Classification accuracy during leave-one-out cross validation (LOOCV) using the classifier produced by Method B. d) Confusion matrix for the classifier produced by Method B.

tends to be mislabeled as stromal tissue by ModA. The images used to test ModA have more stromal tissue than the training images (based on histograms of labels in each set), so the penalty for this error is less severe. The classification accuracy in either case is quite low, which may be due to poor algorithm convergence or poor features. Algorithm convergence for ModA is observed in Figure 2b, which is defined as when the change in the norm of the parameter vector reduces below a defined threshold. The learning rate is reduced until reliable convergence is achieved. High bias despite algorithmic convergence suggests that the model is not complex enough, i.e. we need more or better features. This conclusion is further supported by the fact that the bias does not improve as the number of training examples increases (Fig. 2a).

Intuitively, we should expect high bias out of ModA since the majority of the features used to construct ModA are intensity based. To enhance the feature set, Haralick textural features [7] were added for training ModB. In addition, an SVM algorithm was used for faster training. The training error and generalization error during LOOCV is shown in Figure 2c. The average training accuracy is 67%, while the average generalization accuracy is 63%. This is a significant improvement over the classification accuracy of ModA. It is interesting to investigate the impact of the different features. In Figure 3a, the generalization error of models trained using textural and intensity features is compared to the generalization error of models trained using only textural features. The full feature set only improves the average error by 3%, suggesting the intensity features do not contribute significantly to the overall classifier. The ratio of the confusion matrices produced by the models of the two feature sets is shown in Figure 3b. A number greater than one indicates the full feature set improves over the textural feature set alone. The addition if intensity features does increase the overall accuracy of the classifier, but the dominant features appear to be textural. As a future investigation, it may be interesting to recalculate these features on filtered versions of the image. For instance, derivatives could be used for edge detection and integrals could be used for smoothing.

**Method B - Feature Analysis**




Full Feat. Conf Matrix / Textural Feat. Conf. Matrix

b)

| | E | S | B |
|---|---|---|---|
| E | 1.06 | 0.81 | 0.95 |
| S | 0.72 | 1.02 | 1.06 |
| B | 0.95 | 0.85 | 1.09 |

Figure 3. a) Generalization accuracy comparison for models trained using different feature sets. b) Ratio of confusion matrices for model using full feature set and model using only the textural features.

**References**

1. D.H. Patey, R.W. Scarff, The position of histology in the prognosis of carcinoma of the breast. *Lancet* 211, 801-804 (1928)
2. C.W. Elston, I.O. Ellis, Pathological prognostic factors in breast cancer I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up. *Histopathology* 19, 403-410 (1991)
3. A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, D. Koller. "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival." *Science translational medicine* 3.108 (2011)
4. D. Hanahan, R. A. Weinberg, The hallmarks of cancer. *Cell* 100, 57–70 (2000)
5. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell* 144, 646–674 (2011).
6. T. R. Fanshawe, A. G. Lynch, I. O. Ellis, A. R. Green, R. Hanka, Assessing agreement between multiple raters with missing rating information, applied to breast cancer tumour grading. *PLos One 3*, e2925 (2008).
7. R. M. Haralick, K. Shanmugam, I. Dinstein, Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* vol. smc-3, No. 6, 613-621 (1973).
8. R.-E. Fa, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, LIBLINEAR: A library for large linear classification *Journal of Machine Learning Research* 9(2008), 1871-1874.