Feature Analysis of Biomarker Descriptors for HER2 Classification of Histology Slides

Ramakrishnan Mukundan[1][0000-0003-4578-1931]

[1]Department of Computer Science and Software Engineering
University of Canterbury, Christchurch, New Zealand.
mukundan@canterbury.ac.nz

Abstract. Whole slide images (WSI) of histology slides are increasingly being used for computer assisted evaluations, automated grading and classification. In this rapidly evolving research field, several classification algorithms and feature descriptors have been reported for histopathological analysis. While some algorithms use pixel values of entire images as features, other methods try to use specific biomarker related features. This paper analyses in detail feature descriptors that have been found to be efficient in classifying ImmunoHistoChemistry (IHC) stained slides. These features are directly related to the Human Epidermal Growth Factor Receptor 2 (HER2) biomarkers that are commonly used for grading such slides. Characteristic curves are intensity features that encode information about the variation of the percentage of stained membrane regions with saturation levels. The uniform Local Binary Patterns (ULBP) are texture features extracted from stained regions. ULBP contains several components and generates a high dimensional feature vector that needs to be compressed. Fisher Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are used to select feature components important in classification. The paper proposes a method to combine different types of features (eg., intensity and texture) after dimensionality reduction, and to improve classification accuracy by maximizing inter-class separability. The paper also discusses methods to visualize class-wise distribution of the computed feature vectors. Experimental analysis performed using a WSI dataset of IHC stained slides and aforementioned features are also presented.

Keywords: Uniform Local Binary Patterns, Characteristic Curves, Whole Slide Image Processing, Feature Analysis, Linear Discriminant Analysis, Principal Component Analysis, Cosine Measure.

1 Introduction

The field of digital pathology has recently seen an exponential growth in image analysis applications resulting from the availability of powerful scanners that can produce very high resolution Whole Slide Images (WSIs) of entire biopsy slides [1]. The images contain billions of pixels (typically of size 50,000 × 50,000 pixels), at high magnifications (up to 40×). This technology allows physical slides to be

transformed into digital resources that can be processed by computer software for extensive analysis of complex cellular and protein features [2]. The digital images, also known as "virtual slides" can be easily stored and almost instantaneously transmitted to a different location in a reliable and secure manner. A digital pathology platform thus allows instant sharing of WSIs where multiple pathologists can review the same case in parallel. WSIs are increasingly being used in clinical applications for pathological diagnosis [3].

The need for developing image analysis methods that can measure biomarker specific features is emphasized in recent studies [4][5]. In the past few years, WSI instrumentation has rapidly advanced to the level where scanners can automatically load up to 300 slides without user intervention, with considerably faster scanning speeds [5]. This has further led to the need for fast image processing algorithms that can detect and analyse various cytological features, biomarkers and texture characteristics, and accurately extract information that are relevant to histopathological studies and diagnosis. In this context, algorithms for automatic classification of breast cancer histology slides have received significant attention in the recent past [6][7]. Online contests and programming challenges have also been recently organized by research groups with the aim of accelerating the development of computational techniques for automated analysis, classification and scoring of breast cancer histology slides [8][9].

The Human Epidermal Growth factor Receptor 2 (ERBB2 or HER2) protein is an important cell membrane biomarker commonly used for breast cancer diagnosis. Patients with breast tumors that overexpress HER2 have aggressive disease and poor prognosis [10]. Breast tissue samples are assigned HER2 scores 0, 1+ (Negative), 2+ (equivocal) or 3+ (positive) depending on the intensity, percentage and pattern of membrane staining observed in Immunohistochemistry (IHC) stained slides [11]. Using biomarker specific image features in classification algorithms can lead to better accuracy and diagnostic concordance with pathologist's assessments [12][13]. For this, the feature representation capabilities and the discriminating power of the selected descriptors will need to be evaluated in detail. Further, a thorough discriminant analysis will help in reducing the dimensionality of the features to a near-optimal value. To the author's knowledge, not much work has been reported in the area of analysis of features for automated classification of histology slides. This paper deals with two recently introduced image features associated with HER2 over-expression in IHC stained slides: (i) characteristic curves [14], and (ii) rotation invariant uniform local binary patterns (ULBP) [15]. Characteristic curves encode minimal information about the variation of observed percentage of staining with respect to saturation levels, and therefore has a low dimension. On the other hand, higher order texture descriptors such as the ULBP contain a large number of components. These features have important geometrical characteristics that are useful for both feature reduction and visualization. When feature points fall along a smooth curve, we could make use of the information redundancy in the set to reduce the dimension of the feature vector. Similarly, some of the common geometrical characteristics of the shapes of the feature curves could be used to get a two or three

dimensional representation of the feature vector for visualizing how the points are clustered within each class, and how they are separated between classes.

This paper uses Fisher Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) to evaluate the effectiveness of the features, reduce feature dimension, combine features of different types, maximize inter-class separability and to improve the overall classification accuracy. Such discriminant analysis approaches are commonly used to study the feature transformations used in multi-class classification algorithms [16]. The importance of both PCA and LDA is also highlighted in [17]. We illustrate our approach using experimental results obtained with a large WSI dataset containing IHC stained slides.

This paper is organized as follows. The next section gives a description of the dataset, the methods used for classification and experimental analysis. Section 3 and 4 give an overview of characteristic curves and uniform local binary patterns respectively, and ways to visualize their distribution using low-dimensional data. Section 5 gives the results of principal component analysis of the data matrices. The linear disriminant analysis performed on the outputs of the PCA is outlined in Section 6. This section also gives an overview of changes in classification accuracy and feature dimension at each processing stage. Section 7 provides a summary of the work presented in the paper and outlines future research directions.

2 Materials and Methods

The dataset used in this research work was provided by the University of Warwick as part of the online HER2 scoring contest [8]. Permission was granted by the contest organizers to participating teams for the use of the dataset for research and academic purposes. The dataset consisted of a total of 172 whole slide images in Nano-zoomer Digital Pathology (NDPI) format. These WSIs were extracted from 86 cases of patients with invasive breast carcinomas [7]. For each case, WSIs of both Hematoxylin and Eosin (H&E) stained and Immunohistochemical (IHC) stained slides were provided. The training data included ground truth provided by expert pathologists and consisted of the HER2 scores assigned for each case and also the observed percentage of membrane staining in the tissue sample.

For the experimental analysis presented in this paper, we used 52 WSIs of IHC stained images from the training dataset, with 13 WSIs belonging to each of the four HER2 classes. The WSIs had varying sizes, with the average size of about 60,000 × 55,000 pixels. Each image was further subdivided into approximately 26 small tiles (image patches) of size 512×512 pixels containing at least 80% region of interest. A total of 1271 image patches from this set were used in our experimental analysis. This set was further subdivided into a training set consisting of 900 image patches, and a cross-validation set consisting of 371 image patches. As features we used two independent sets: characteristic curves (dimension = 20) and uniform local binary patterns (dimension = 168). PCA was used to find the optimal dimension for each set and then LDA used to maximize class separability (Fig. 1). For the purpose of

evaluating the variation of accuracy of the feature descriptors, we used one-vs-rest logistic regression for training and cross-validation.

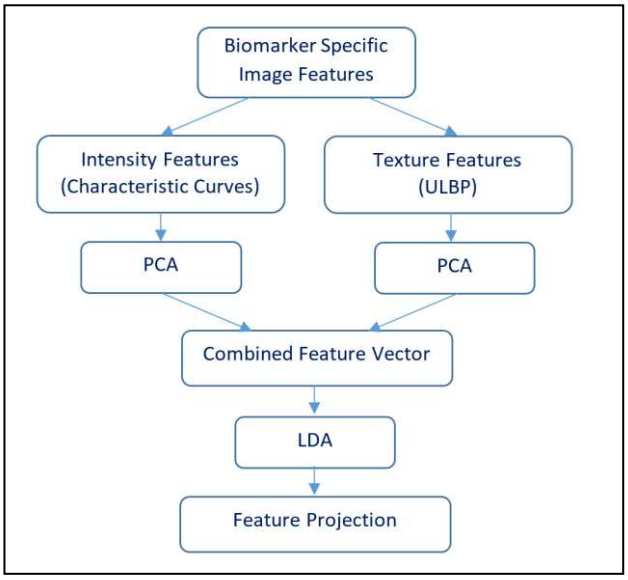

Fig. 1. A flow diagram of the discriminant analysis performed using biomarker features.

Characteristic curves and ULBP are completely independent features. They are therefore analysed independently using PCA to get the reduced set of features based on the variation of classification accuracy with dimension of the principal components. The optimal set of features are then combined and linear discriminant analysis performed to get the projected set of features with maximum class separability. These feature are used in the final classification algorithm using one-vs-all logistic regression.

3 Characteristic Curves

Characteristic curves (or percentage-saturation curves) represent the variation of the percentage of stained pixels in an image (within predefined experimentally determined hue thresholds) with changes in the saturation threshold [14] from 0.1 to 0.5. As the number of pixels above the threshold reduces when the saturation threshold is increased, the graph takes the shape of a cumulative histogram plotted in the opposite direction, with a monotonically decreasing trend. Fig. 2 shows the formation of the characteristic curve using a sample image.

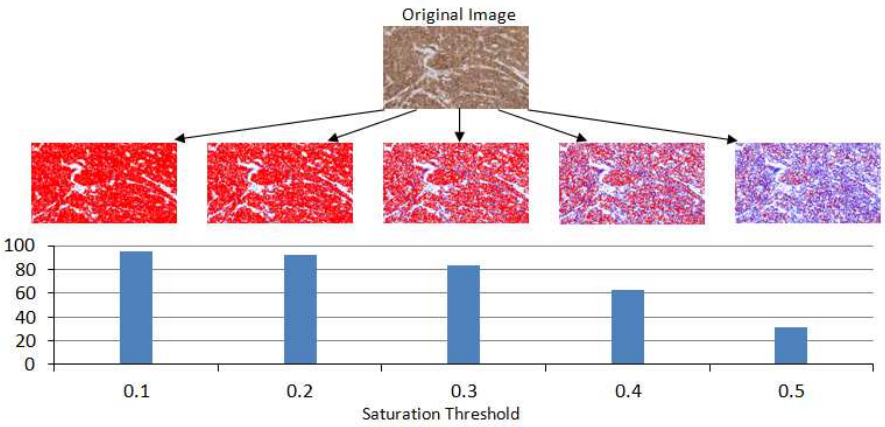

Fig. 2. Values of the characteristic curve for a sample image. Top row: IHC stained image, middle row: thresholded images with saturation value increased, bottom row: corresponding charts showing the percentage of stained pixels.

The characteristic curves assume distinct shapes for each HER2 score as shown in Fig. 3, and therefore are good candidate features for classification algorithms. The shapes of the characteristic curves can be directly correlated with the staining levels required for HER2 scores as per the assessment guidelines [11]. For example, the characteristic curve always lies below the 10% threshold when the score is 0, and only a small initial segment of the curve lies above the 10% mark when the score is 1. If the score is 3+, the curve lies completely above the 30% mark showing a strong and complete membrane staining. As seen in Fig. 3, the curve passes through a much wider range of values of percentage staining when the score is 2+.

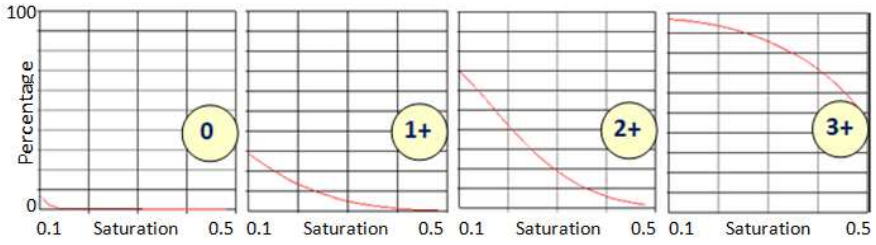

Fig. 3. The shapes of the characteristic curves for images with different HER2 scores. The *x*-axis represents saturation values from 0.1 to 0.5, and the *y*-axis represents percentage of staining from 0 to 100%.

Each feature vector in our classification algorithm (one-vs-all logistic regression) used 20 points sampled along the characteristic curves. Since all characteristic curves are non-increasing, and lie between fixed saturation thresholds along the *x*-axis, one global characteristic of the shape is the area under the characteristic curve. A box-plot

showing the distribution of area in the dataset containing 1271 samples is shown in Fig. 4. The single metric itself shows a good inter-class separation of the feature vectors. Another method to visualize the inter-class variance and intra-class similarities of the feature vectors is to project them to two or three principal components and plot the values as two or three dimensional set of points. However, the visualization we get with such minimal set of values may not give an accurate representation of either the discriminating power or the information redundancy in the feature vectors. This is particularly true when the feature vector has a large dimension as in the case of the ULBP vector discussed in the next section.

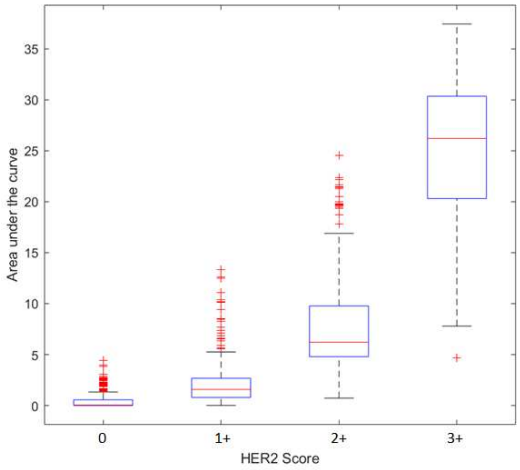

Fig. 4. Box plot showing the distribution of the area under characteristic curves for an input dataset containing 1271 samples.

4 Uniform Local Binary Patterns

Uniform local binary patterns (ULBP) are rotation invariant texture descriptors useful for characterising changes in the local texture in pixel neighbourhoods. The computation of nine ULBP components $U_0 ... U_8$ are detailed in [15]. We use only the first eight components since U_8 mainly represents background regions of constant intensity. Each ULBP feature curve consisted of 21 sampled points, and therefore the whole feature vector with eight components had a dimension of 168. Fig. 5 shows the variation of U_0, U_1, U_2 with the saturation threshold. As seen in the figure, ULBP feature curves also exhibit good inter class variance between classes with HER2 scores 1+, 2+ and 3+. However, the variance is found to be small between classes 0 and 1+ because between those two classes, there is no significant difference in the texture of staining patterns. Similarly, when the saturation threshold is increased, regions become more uniform in colour values, and hence the LBP values all tend to zero.

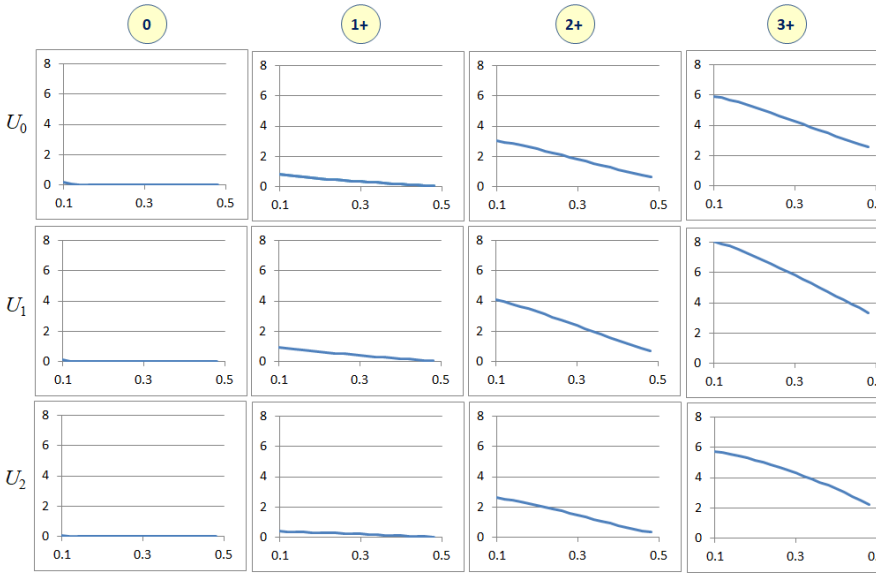

Fig. 5. Variations of ULBP components with saturation thresholds for images with different HER2 scores. The x-axis represents saturation values from 0.1 to 0.5, and the y-axis represents the ULBP feature values.

In Fig. 5, we notice that the ULBP feature curves have very low curvature compared to the characteristic curves. They can therefore by approximated by linear segments and parameterized into slope and y-intercepts to visualize their distribution in a training set. Fig. 6 shows such a distribution where each point on the slope-height plot represents a ULBP feature curve. This figure clearly shows how the points are clustered together in each class, and how much inter-class separation exists between them.

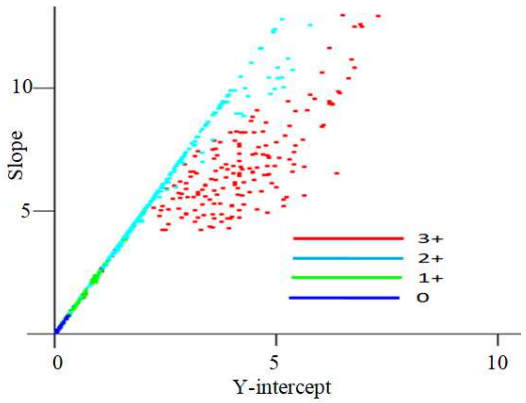

Fig. 6. Two-dimensional parameterization of ULBP feature vectors.

Fig. 6 also shows an important aspect of ULBP features that they have much wider range of variation in slopes with height value for HER2 class 3+, while for other classes, the slope varies nearly linearly with height. This is expected since when the intensity and percentage of staining are low, the variations in texture will be nearly uniform, whereas for cases where the HER2 score is 3+, we observe significant variations in texture patterns.

5 Principal Component Analysis

The principal component analysis (PCA) is a well-known technique for dimensionality reduction. It allows us to discover and retain important feature components while discarding irrelevant components and maintaining the desired level of classification accuracy. The ULBP feature vector consists of eight ULBP components, each comprising of 21 points. Such high dimensional feature vectors will contain redundant information and strongly correlated data which could be minimized using PCA. Several methods perform feature normalization before using PCA. In our case, the inter class separation of features is primarily in the magnitude of feature values, particularly for ULBP. As seen in Fig. 5 above, the features for classes 0 and 1+ are separated by magnitude, and not slope. Similarly, features for classes 2+ and 3+ also have similar slopes but different magnitudes. Feature normalization will bring vectors representing different classes closer together, reducing inter-class separability. Therefore, we skipped the process of feature normalization and used the original set of features as inputs to the PCA.

The PCA is computed using the left singular vector L of the covariance matrix given by

$$\Sigma = \frac{1}{m} X^T X \tag{1}$$

where X is the $m \times n$ data matrix containing feature values. The number of training samples m in our case is 900, and n is the feature dimension. For characteristic curves, n = 20, and for ULBP curves, n = 168. Given the required number k of principal component vectors, we extract the first k eigen vectors in L to form a $n \times k$ matrix L_k. We then project X onto the new space of k vectors:

$$Z = X L_k \tag{2}$$

where Z is the new data matrix of size $m \times k$.

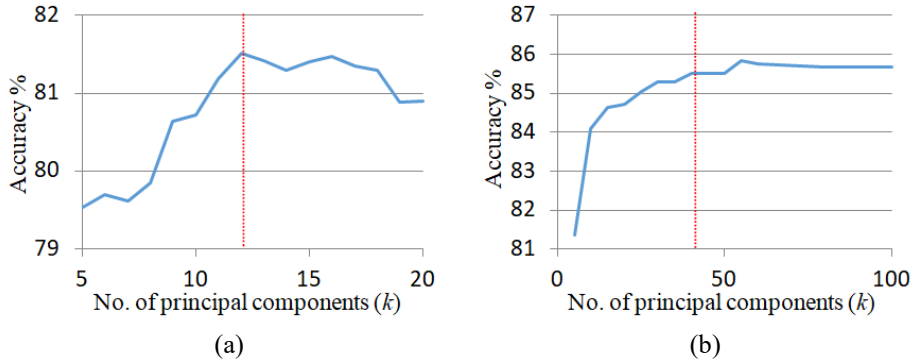

Fig. 7. Variation of classification accuracy with the number of principal components for (a) characteristic curves and (b) ULBP curves. The red line indicates the number of principal components selected based on an acceptable accuracy level.

Using PCA, the number of principal components k was varied over a range of values as shown in Fig. 7, and the classification accuracy using the projected feature vector obtained for both characteristic curve features and ULBP features as inputs. The analysis showed that the principal component dimension of the characteristic curve feature could be reduced from 20 to 12, and the ULBP feature from 160 to 40 without significantly affecting the classification accuracy. These principal component vectors are combined to form a feature vector of dimension 52. The classification accuracy with this combined feature vector showed an improvement over the accuracy obtained using individual features (Table 1).

6 Linear Discriminant Analysis

Linear discriminant analysis allows us to project the dataset Z (Eq. 2) obtained from PCA onto even lower dimensional space with increased feature separability between classes. Unlike PCA which ignores class labels and projects data to directions of maximal variance, LDA uses class labels to analyse the variations of each component of the feature vector within each class and also in-between classes. LDA has been recently used for reducing the feature dimension and improving the classification accuracy of an expert system for stomach cancer images [18]. Similar to PCA, LDA also solves the generalized eigen value problem for the matrix $S_W^{-1}S_B$ where S_W is intra-class scatter matrix and S_B is the inter-class scatter matrix [16]:

$$S_W = \sum_{c=1}^{C} \sum_{x \in c} (x - \mu_c)(x - \mu_c)^T \quad (3)$$

$$S_B = \sum_{c=1}^{C} N_c(\mu_c - \mu)(\mu_c - \mu)^T \quad (4)$$

where, μ_c denotes the mean vector for class c, N_c the number of samples in class c, and μ the overall mean vector. As an example, the values of vector μ_c for the ULBP feature vectors are shown in Fig. 8. Each ULBP feature vector contains 40 values.

The average of the ULBP vectors over images belonging to the four HER2 classes are taken, providing four average ULBP vectors each of length 40. Fig. 8 shows that there is a good inter-class separability between classes 1+, 2+, and 3+, while significant overlap exists between classes 0 and 1+.

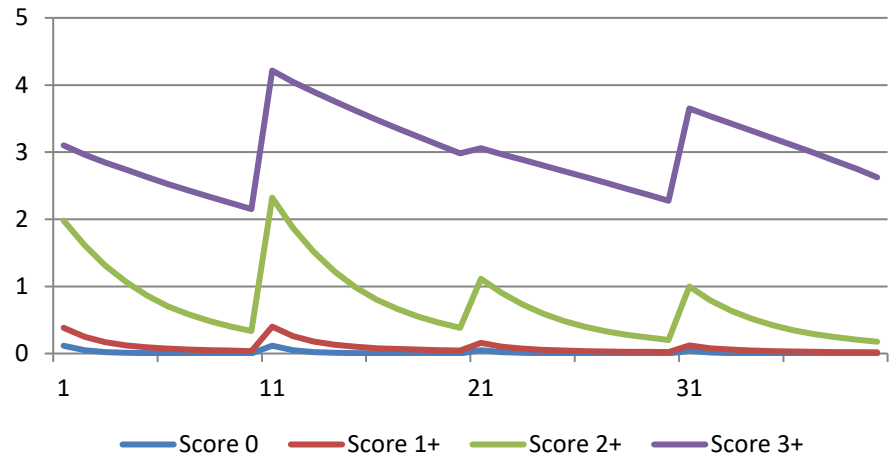

Fig. 8. The mean values μ_c of the ULBP vector for each of the four classes.

For our analysis, the input dataset Z contained 900 training samples and each vector had a dimension 52 as described in the previous section. Using LDA, we could further reduce the dimension to 49, and the accuracy improved due to increased class separation. For comparison, the accuracy values obtained from different processing stages are shown in Table 1.

Table 1. Variations in feature dimensions and classification accuracy after PCA and LDA stages.

Feature vector	Dimension	Classification Accuracy
Characteristic curves	20	80.72 %
ULBP feature curves	168	85.60%
Characteristic curves (CC) after PCA	12	81.51 %
ULBP feature curves after PCA	40	85.52 %
Combined feature vector (CC + ULBP)	52	86.46%
Combined feature vector after LDA	49	87.2 %

7 Conclusions and Future Work

This paper has demonstrated the importance of discriminant analysis of biomarker specific features used for classification of breast cancer histology slides. Specifically,

two types of features, characteristic curves and rotation invariant uniform local binary pattern curves were considered and their properties discussed. The feature vectors belong to completely different categories: one is intensity based while the other is texture based, and they also have a large difference in their dimensions. Methods based on the geometrical characteristics of the features to visualize their distribution in the training set have been presented. The paper also presented a method to independently reduce the dimensionality of the two feature sets and to combine them using principal component analysis and linear discriminant analysis. The improvement in the classification accuracy was also demonstrated using experimental data. Even though the results presented in the paper are specific to the two types of feature vectors mentioned above, the proposed method can be applied to other types of biomarker features useful for classifying histology slides where different types of vectors representing completely different image attributes need to be combined and reduced.

Further work is directed towards improving the accuracy of features and also refining the training set by analysing the outputs of the proposed method to identify regions where classes overlap. We have observed significant overlap between regions corresponding to classes 1+ and 2+, and also some overlap between classes 2+ and 3+. Minimising such overlaps will significantly improve the overall accuracy. It should also be noted that due to inaccuracies present in the process of IHC staining of slides, there will always be some level of uncertainty in the stain intensity that will correspond to inaccuracies in the assessment of slides [19].

References

1. Farahani, N., Parwani, A. V., & Pantanowitz, L. (2015). Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathology and Laboratory Medicine International, 7, 23-33. doi:10.2147/Plmi.S59826.
2. Ghaznavi, F., Evans, A., Madabhushi, A., & Feldman, M. (2013). Digital imaging in pathology: whole-slide imaging and beyond. Annu Rev Pathol, 8, 331-359. doi:10.1146/annurev-pathol-011811-120902
3. Volynskaya, Z., Evans, A. J., & Asa, S. L. (2017). Clinical Applications of Whole-slide Imaging in Anatomic Pathology. Adv Anat Pathol, 24(4), 215-221. doi:10.1097/PAP.0000000000000153.
4. Williams, B. J., Hanby, A., Millican-Slater, R., Nijhawan, A., Verghese, E., & Treanor, D. (2017). Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. Histopathology. doi:10.1111/his.13403
5. Hamilton, P. W., Bankhead, P., Wang, Y., Hutchinson, R., Kieran, D., McArt, D. G., . . . Salto-Tellez, M. (2014). Digital pathology and image analysis in tissue biomarker research. Methods, 70(1), 59-73. doi:https://doi.org/10.1016/j.ymeth.2014.06.015
6. Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., & Polónia, A. (2017). Classification of breast cancer histology images using Convolutional Neural Networks. PLoS ONE, 12(6). doi:10.1371/journal.pone.0177544

7. Qaiser, T., Mukherjee, A., Reddy Pb, C., Munugoti, S. D., Tallam, V., Pitkaaho, T., . . . Rajpoot, N. (2018). HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology, 72(2), 227-238. doi:10.1111/his.13333
8. University of Warwick, (2016). "HER2 Scoring Contest." Retrieved 4 April, 2016, from https://warwick.ac.uk/fac/sci/dcs/research/tia/her2contest/
9. BACH (2018). "ICIAR-2018 Grand Challenge on Breast Cancer Histology Images." Retrieved 10 Jan, 2018, from https://iciar2018-challenge.grand-challenge.org/home/
10. Ross, J. S., Slodkowska, E. A., Symmans, W. F., Pusztai, L., Ravdin, P. M., & Hortobagyi, G. N. (2009). The HER-2 receptor and breast cancer: ten years of targeted anti-HER-2 therapy and personalized medicine. Oncologist, 14(4), 320-368. doi:10.1634/theoncologist.2008-0230
11. Hicks, D. G., & Schiffhauer, L. (2011). Standardized Assessment of the HER2 Status in Breast Cancer by Immunohistochemistry. Laboratory Medicine, 42(8), 459-467. doi:10.1309/LMGZZ58CTS0DBGTW
12. Akbar, S., Jordan, L. B., Purdie, C. A., Thompson, A. M., & McKenna, S. J. (2015). Comparing computer-generated and pathologist-generated tumour segmentations for immunohistochemical scoring of breast tissue microarrays. British Journal of Cancer, 113(7), 1075-1080. doi:10.1038/bjc.2015.309
13. Goacher, E., Randell, R., Williams, B., & Treanor, D. (2017). The Diagnostic Concordance of Whole Slide Imaging and Light Microscopy: A Systematic Review. Arch Pathol Lab Med, 141(1), 151-161. doi:10.5858/arpa.2016-0025-RA
14. Mukundan, R. (2017). A Robust Algorithm for Automated HER2 Scoring in Breast Cancer Histology Slides Using Characteristic Curves. Medical Image Understanding and Analysis, MIUA-2017: Univ of Edinburgh. Springer International Publishing: 386-397. doi:10.1007/978-3-319-60964-5_34
15. Mukundan, R. (2018). Image Features Based on Characteristic Curves and Local Binary Patterns for Automated HER2 Scoring. Journal of Imaging, 4(2), 35. doi:10.3390/jimaging4020035
16. Li, T., Zhu, S., & Ogihara, M. (2006). Using discriminant analysis for multi-class classification: an experimental investigation. Knowledge and Information Systems, 10(4), 453-472. doi:10.1007/s10115-006-0013-y
17. Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), 228-233. doi:10.1109/34.908974
18. Korkmaz, S. A., Bínol, H., Akçiçek, A., & Korkmaz, M. F. (2017, 14-16 Sept. 2017). A expert system for stomach cancer images with artificial neural network by using HOG features and linear discriminant analysis: HOG_LDA_ANN. Paper presented at the 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY).
19. Gavrielides, M. A., Gallas, B. D., & Hewitt, S. M. (2015). Uncertainty in the assessment of immunohistochemical staining with optical and digital microscopy: lessons from a reader study. SPIE Medical Imaging.