Segmentation of HER2 protein overexpression in immunohistochemically stained breast cancer images using Support Vector Machines

Raquel Pezoa\textsuperscript{1,2}, Luis Salinas\textsuperscript{1,2}, Claudio Torres\textsuperscript{1,2}, Steffen Härte\textsuperscript{3}, Cristián Maureira-Fredes\textsuperscript{4} and Paola Arce\textsuperscript{1}

\textsuperscript{1} Department of Informatics, Universidad Técnica Federico Santa María, Valparaíso, Chile
\textsuperscript{2} Centro Científico Tecnológico de Valparaíso, Universidad Técnica Federico Santa María, Valparaíso, Chile
\textsuperscript{3} SCIAN-Lab, ICBM, BNI, University of Chile, Chile
\textsuperscript{4} Max Planck Institut für Gravitationsphysik (Albert-Einstein-Institut), D-14476 Potsdam, Germany

E-mail: raquel.pezoa@usm.cl

Abstract. Breast cancer is one of the most common cancers in women worldwide. Patient therapy is widely supported by analysis of immunohistochemically (IHC) stained tissue sections. In particular, the analysis of HER2 overexpression by immunohistochemistry helps to determine when patients are suitable to HER2-targeted treatment. Computational HER2 overexpression analysis is still an open problem and a challenging task principally because of the variability of immunohistochemistry tissue samples and the subjectivity of the specialists to assess the samples. In addition, the immunohistochemistry process can produce diverse artifacts that difficult the HER2 overexpression assessment.

In this paper we study the segmentation of HER2 overexpression in IHC stained breast cancer tissue images using a support vector machine (SVM) classifier. We assess the SVM performance using diverse color and texture pixel-level features including the RGB, CMYK, HSV, CIE \textit{L*a*b*} color spaces, color deconvolution filter and Haralick features. We measure classification performance for three datasets containing a total of 153 IHC images that were previously labeled by a pathologist.

1. Introduction

A common practice for evaluating the progression of diverse types of cancers is the analysis of immunohistochemistry-stained tissue sections. Immunohistochemistry (IHC) is a method for detecting specific antigens in tissues or cells based on antigen-antibody recognition. This technique takes advantage of the specificity provided by the binding of an antibody with its antigen shown at a microscopic level [1]. The antigen-antibody interaction is visualized through a color-producing reaction in various cellular structures such as nucleus, cytoplasm or membrane, for instance.

A typical biomarker observed on the cell membrane is the Human Epidermal growth factor Receptor 2 (HER2). This is a protein normally expressed on the cell membrane surface of various organs epithelia such as lung, bladder, pancreas, breast, stomach and prostate [2]. When HER2 is present in normal amounts, the protein plays an important role in normal cell growth...
development. However, when the HER2 protein is overexpressed, cells multiply and grow more rapidly than normal, promoting carcinogenesis in diverse types of cancer [3].

The HER2 protein has become an important biomarker for invasive breast cancer. Around 15%-25% of breast cancer overexpress HER2 protein. Its measurement helps to guide the therapeutic decision process.

![IHC-stained breast tissue images with score (a) 1+, (b) 2+ and (c) 3+ (40x magnification and 1000 × 1000 pixels). The cell structures of interest are cell membranes with HER2 overexpression which are stained in brown color. Nuclei are stained in a blue color, to give contrast. (d) and (e) show two examples of cells with different levels of HER2 overexpression.](image)

**Figure 1.** IHC-stained breast tissue images with score (a) 1+, (b) 2+ and (c) 3+ (40x magnification and 1000 × 1000 pixels). The cell structures of interest are cell membranes with HER2 overexpression which are stained in brown color. Nuclei are stained in a blue color, to give contrast. (d) and (e) show two examples of cells with different levels of HER2 overexpression.

When IHC is used to assess the amount of HER2 overexpression, two main chemical compounds (commonly referred to as stains) are used to visualize cell nuclei and membranes: hematoxylin and diaminobenzidine (DAB). The hematoxylin is used as a counterstain and produces a blue-violet staining of nuclear elements. The diaminobenzidine produces a dark brown reaction on the cell membrane (only when there is HER2 overexpression).

In general, pathologists perform a qualitative assessment of HER2 overexpression through a visual inspection and classify IHC tissues according to the ASCO/CAP guidelines [4]. They should estimate two main parameters associated to HER2 overexpression: intensity and completeness. The former corresponds to the intensity of brown staining (from weak to intense) and the latter corresponds to the percentage of staining completeness with respect to the membrane length.

According to the two estimated parameters the observed specimen is classified into four categories: 0, 1+, 2+ and 3+. The classification system proposed in [4] indicates: 0 means no staining was observed, 1+ barely perceptible, 2+ weak to moderate and 3+ means strong intensity was observed in more than 10% of tumor cells (see Fig. 1).

Thus, the classification (and the whole analysis) is mainly based on the strength of the staining, whose evaluation clearly depends on the experience of the pathologist. This subjective analysis can produce inconsistent results. Hence, the importance of a quantitative and objective analysis is recognized by researchers in both, the image analysis and pathology fields [5].

Several cell segmentation methods have been proposed. However HER2 overexpression segmentation is still and open problem. This is due to the non-standardization of tissue samples and also due to the undesired brown staining over the cytoplasm. Regarding non-standardization, immunohistochemical tissue samples can be quite variable due to the use of different whole slide imaging (WSI) scanners, immunohistochemical kit systems or staining procedures [6], [7], [8]. This tissue variability makes difficult the finding of a general tissue pattern. On the other hand, HER2 overexpression is visualized through the brown color reaction produced on the membrane, but it is quite frequent that brown color spreads over the cytoplasm, which is an artefact and should not be considered as HER2 overexpression.
In this study we address the problem of segmentation of cell membranes with HER2 overexpression. This is a previous but crucial step before being able to estimate intensity and completeness parameters of the HER2 overexpression. We proposed a method for the segmentation of HER2 overexpression using a support vector machine (SVM). We assessed the SVM performance using diverse color and texture pixel-level features including the RGB, CMYK, HSV, CIE L*a*b* color spaces, color deconvolution filter and Haralick features.

The outline of this paper is described as follows. Section 2 describes the proposed method, including the computation of color and texture features and SVM pixel classification. In section 3 we measure the performance of SVM classifiers on three different sets of features and three datasets including images 1+ in data set one, images 2+ in data set two and images 3+ in data set three. Finally, Section 4 presents the conclusions and perspectives of this work.

2. Segmentation using SVM

The proposed method addresses the segmentation of HER2 overexpression as a binary classification problem and each pixel belongs to one of the two classes: overexpressed and non-overexpressed. The method uses a SVM as binary pixel classifier. As mentioned before, IHC images can be highly variable which makes difficult to find a common tissue pattern. Color variation is one of the most important issues, and this is produced for a number of reasons such as: the variable histochemical stains kits coming from different manufacturers, the staining procedure (which in many cases is performed in a manual way), or the use of different whole slide imaging (WSI) scanners. In fact, there is a lack of standardization between IHC clinical laboratories. Current developments in this field strongly suggest the need for the standardization of this process [9, 8, 7, 6]. Thus, considering this variable scenario, we decide to build one SVM classifier per IHC image.

The proposed method includes three main stages: pixel-level feature computation, training data generation, and SVM pixel classification (see Fig. 2). The aim of feature computation is to represent the IHC image pixels by means of a descriptive set of features that captures the complex characteristics of HER2 overexpression. In this paper, we compute a set of 210 color and texture features. Training data generation is a fundamental step of any supervised machine learning method. In this study, training pixels are selected from a small sub-image of the IHC image. Then, we build the SVM classifier in order to predict the label of the remaining pixels. Notice that the selection of the sub-image is an important issue. The proposed method needs that the expert provides training data to build the SVM classifier. This training data should be a small but representative set of the whole IHC image. The point of doing this is that we want to include the pathologist knowledge in order to guide the segmentation, due to the complexity of the IHC images, and pathologist are able to label small portions of the whole IHC image. In this paper, we have the gold standard of the whole IHC image, thus, we are able to simulate when the pathologist generates the training data.

2.1. Pixel-level Feature Computation

In this work, we calculate a total of 210 color and texture pixel-level features (see Table 1). We use color features that correspond to different space color channels and a color deconvolution filter. We use four color spaces: RGB, CMYK, HSV and CIEL*a*b*, together with the grayscale image. In addition, we use a color deconvolution filter for stain separation in IHC stained images. This deconvolution is achieved by means of a transformation of the original RGB IHC image, which is based on the Lambert-Beer law and the optical density. In particular, we use the DAB channel obtained with the color deconvolution filter [10].

Texture features includes the calculation of Gray Level Co-occurrence Matrices (GLCMs) in order to obtain GLCM mean, GLCM standard deviation and eleven Haralick texture features [11] (a total of thirteen features). This calculation has a high computational cost, because we
Figure 2. Overview of the main stages of the proposed method. The input is an RGB image obtained from a WSI scanner. The output is a binary image, where white pixels indicate overexpressed class, and black pixels indicate non-overexpressed class.

| Type            | Feature                                      |
|-----------------|----------------------------------------------|
| Color           | RGB                                          | 3 channels, R, G and B                      |
|                 | CMYK                                         | 4 channels, C, M, Y and K                   |
|                 | HSV                                          | 3 channels, H, S and V                      |
|                 | CIEL*a*b*                                    | 3 channels, L*, a* and b*                   |
|                 | Grayscale image                              | 1 channel                                   |
|                 | Color deconvolution                          | 1 DAB channel                               |
| Texture         | Haralick features                            | 11 Haralick features calculated for each the 15 color channels (which give a total of 165) |
|                 | GLCM mean                                    | Calculated for each the 15 color channels    |
|                 | GLCM standar deviation                       | Calculated for each the 15 color channels    |

Table 1. Color and texture pixel-level features.

are computing GLCM matrices in a pixel-level. For this calculation we define a neighborhood for each pixel and a GLCM matrix is calculated for each neighborhood. Then, from each GLCM matrix, we calculate thirteen texture features for each pixel. Due to the nature of the algorithmic process, we implemented a set of parallel routines in C++ code, using the MPI library, in order to speed up the whole process.

2.2. SVM pixel classification
Support Vector Machines (SVM) are supervised learning algorithms widely used for binary classification. SVM were introduced by Vladimir Vapnik et al. [12] and the general idea is to find an optimal separating hyperplane that maximizes the minimum distance between any data point in order to separate data into two classes. We use SVM as a binary pixel classifier, thus,
the segmentation of HER2 overexpression is handled as a classification problem. Each pixel is classified into one of the following classes: overexpressed or non-overexpressed. We build a SVM model for each IHC image using the pixels of a sub-image (training data). Then we predict the class labels of the remaining pixels (testing data).

3. Experiments

3.1. Data

A total of 153 images were used for the experiments, including around 750 cells. The images are grouped into three datasets. Dataset 1, dataset 2 and dataset 3 contain images of category 1+, 2+ and 3+, respectively. The images are regions of interest (ROIs) extracted from paraffin embedded breast cancer tissue slides obtained from the archives of the Department of Pathology at the Hospital Clínico Universidad de Chile. The tissues were scanned using the NanoZoomer-XR C12000 whole slide imaging scanner (Hamamatsu Photonics K.K.) at a magnification of 40x and a resolution of 0.23 µm/pixel. All images were manually labeled by a pathologist in order to generate training data and gold standard for the evaluation of the proposed method.

3.2. Experiment Setup

The goal of this experiment is to analyze the SVM performance using three set of features: the whole set of 210 features; (ii) a subset of 4 features corresponding to the channels of CMYK color space; (iii) and another subset of 14 features corresponding to the texture features applied to the gray channel (see Table 1).

For building the SVM classifiers, we select the Gaussian Radial Basis Function (RBF) as the SVM kernel function and we perform 5-fold cross-validation. The F1 metric [13] is used to measure performance in the cross-validation stage. We select the F1 metric because we are dealing with an unbalanced classification problem (there are more pixels belonging to the non-overexpressed class than the overexpressed class) [14]. Notice that the most commonly metric used in the machine learning literature is accuracy [13], however it is not well suited in this context.

In this experiment, for each IHC image \( I \), we generate \( N \) sub-images \( s_k \), \( k = 1, \ldots, N \). All the pixels of the sub-image \( s_k \) are used to train SVM classifiers \( \text{svm}_k \), \( k = 1, \ldots, N \). For each \( k \), the \( \text{svm}_k \) classifier is then used to classify the remaining pixels of the IHC image. Since we have the gold standard of the IHC image, we can calculate the performance \( \eta_k \) of the \( \text{svm}_k \) classifier. From the \( N \) sub-images \( s_k \), we select the sub-image which gives the best SVM performance. We call this sub-image the “best sub-image”, \( s^* \) (see Fig. 3).

This selection is repeated for each IHC image of our dataset. The performance presented in this paper correspond to the performance of the best sub-images of each IHC image. In addition, we group them into the 1+, 2+ and 3+ categories and into the three different set of features (see box plots of Fig. 4).

Regarding the computational resources, we used a DELL Studio server to run the experiments. This server has 8 cores (Intel(R) Core(TM) i7 CPU 960 @3.20GHz) 24 GB RAM, and the Archlinux operating system. The framework was mainly programmed in Python, using the scikit-learn library [15] for machine learning algorithms. In particular, the Haralick features were calculated in the datura cluster (Max-Planck-Institut für Gravitationsphysik, Albert-Einstein-Institut, Germany), which has a total of 165 nodes (we used 128), and each node has 12 cores (CPU: Intel(R) Xeon(R) CPU X5650 @ 2.67GHz), 24 GB RAM, Scientific Linux 6.0 operating system, and InfiniBand for nodes communication.

3.3. Performance Evaluation

The SVM pixel classification performance is evaluated using five different metrics: ROC AUC, accuracy, precision, recall, and F1 [13]. The results can be seen in the box plots of Figure 4 and
Figure 3. Selection of the best sub-image. We divide the image into $N$ sub-images ($N = 9$ in this example) and we rank the sub-images according to the performance that is measured on the testing data (the whole IHC without the training data). The sub-image with the best performance, is called the best sub-image.

A summary of these is shown in Table 2.

In most of the cases, the CMYK feature set produced the better SVM performance. We may expect that using the full feature set (210 features) should produce the best performance, but the experiments with high dimensionality do not show this trend. Recall that the performance is measured using a gold standard manually labeled by one expert pathologist. This is a critical point, because the intrinsic subjectivity and variability of the gold standard can affect our results. Ficarra et al [16] clearly describe this situation.

In addition, the experiments show that the best sub-image is not always the same. We observed that when we change the feature set and the performance metric, the best sub-image also changes.

| Dataset-Features | ROC AUC     | Acc.     | Prec. | Rec.     | F1     |
|------------------|-------------|----------|-------|----------|--------|
| 1-210            | 0.75(0.08)  | 0.75(0.08) | 0.97(0.03) | 0.51(0.16) | 0.65(0.13) |
| 1-CMYK           | **0.79(0.06)** | **0.79(0.06)** | 0.95(0.03) | **0.61(0.14)** | **0.73(0.12)** |
| 1-Haralick       | 0.73(0.06)  | 0.73(0.06) | 0.95(0.03) | 0.49(0.13) | 0.64(0.12) |
| 2-210            | 0.72(0.07)  | 0.72(0.07) | **0.98(0.01)** | 0.44(0.14) | 0.6(0.12) |
| 2-CMYK           | **0.79(0.07)** | **0.79(0.07)** | 0.97(0.02) | **0.59(0.14)** | **0.72(0.12)** |
| 2-Haralick       | 0.73(0.08)  | 0.73(0.08) | **0.98(0.01)** | 0.47(0.16) | 0.62(0.15) |
| 3-210            | 0.79(0.08)  | 0.79(0.08) | **0.94(0.04)** | 0.61(0.17) | 0.73(0.13) |
| 3-CMYK           | **0.84(0.07)** | **0.84(0.07)** | 0.93(0.03) | **0.73(0.14)** | **0.81(0.1)** |
| 3-Haralick       | 0.80(0.07)  | 0.80(0.07) | 0.93(0.04) | 0.65(0.16) | 0.75(0.11) |

Table 2. Mean values (and standard deviation) of performance metrics obtained in each data set considering different sets of features.
Figure 4. Box plots of SVM performance. We measured 5 performance metrics (ROC AUC, accuracy, precision, recall and F1). The SVM classifiers were built using the pixels of the best sub-image as training data. From left to right, columns correspond to 210, CMYK and Haralick features sets. From top to down, rows represent dataset 1 (42, 1+), dataset 2 (50, 2+) and dataset 3 (61 images, 3+).

4. Conclusions

We have presented a method to segment HER2 overexpression in IHC stained breast cancer tissue images. We have described the main stages of the method, including the computation of 210 color and texture features, training data generation and SVM pixel classification; and performed experiments to evaluate the contribution of different sets of features to the SVM performance.

The experimental results have shown that using the best sub-image for training data selection and increasing the dimensionality of the feature set do not necessarily produce a better performance (considering our SVM configuration). The results showed that the set of 4 features corresponding to the CMYK color space produced the better SVM performance in most of the cases. This suggests that the aggregation of additional features, i.e., increasing dimensionality, does not contribute to a better segmentation. However, this is not enough evidence to remove those extra dimensions.

Two important aspects of our study have to be highlighted. First, we used a gold standard that was manually labeled by a pathologist. Thus, the interpretation of our results must consider the intrinsic subjectivity of the gold standard. Second, we were able to rank the sub-images
because we have a gold standard, however this is precisely what is missing for a real case. Therefore, our next goals will be to continue analyzing the best set of features (using a gold standard generated for more than one pathologist), and developing methods able to detect the best sub-image, but now, from unlabeled set of images.

4.1. Acknowledgments

The authors acknowledge the assistance in tissue digitization performed at SCIAN Laboratory (University of Chile) and Hamamatsu TIGA Center (Heidelberg University). The authors thanks the Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Potsdam-Golm, Germany, for providing access to datura cluster for feature calculation.

Partial financial support for this research is acknowledged from CCTVAL-FB0821, FONDEF D1111096 and FONDECYT 1150810.

5. References

[1] Taylor C R, Shi S R and Barr N J 2011 Diagnostic Immunohistochemistry (THIRD EDITION) ed Dabbs D J (Philadelphia: W.B. Saunders) pp 1 – 41
[2] Alvarez R, Cortés J, Mattos-Arruda L, Falzon M, Fasolo A, Gandy M, Gianni L, Harbeck N, Piccart M, Zambelli S et al. 2013 Handbook of HER2-targeted agents in breast cancer (Springer Healthcare Limited)
[3] Di Fiore P, Pierce J, Kraus M, Segatto O, King C and Aaronsen S 1987 Science 237 178–182
[4] Wolff A C, Hammond M E, Hicks D G, Dowsett M, McShane L M, Allison K H, Allred D C, Bartlett J M, Bilous M, Fitzgibbons P, Hanna W, Jenkins R B, Mangu P B, Paik S, Perez E A, Press M F, Spears P A, Vance G H, Viale G and Hayes D 2013 Journal of Clinical Oncology 31 3997–4013
[5] Gurcan M, Boucheron L, Can A, Madabhushi A, Rajpoot N and Yener B 2009 Biomedical Engineering, IEEE Reviews in 2 147–171
[6] Khan A, Rajpoot N, Treanor D and Magee D 2014 Biomedical Engineering, IEEE Transactions on 61 1729–1738 ISSN 0018-9294
[7] Li X and Platoniotis K N 2015 IEEE Transactions on Biomedical Engineering 62 1862–1873
[8] Bejnordi B E, Litjens G, Timofeeva N, Otte-Hiller I, Homeyer A, Karssemeijer N and van der Laak J A 2016 IEEE Transactions on Medical Imaging 35 404–415
[9] Lin F and Chen Z 2014 Archives of Pathology & Laboratory Medicine 138 1564–1577
[10] Ruijbroek A C and Johnston D A 2001 Analytical and quantitative cytology and histology / the International Academy of Cytology [and] American Society of Cytology 23 291–299
[11] Haralick R M, Shanmugam K and Dinstein I 1973 Systems, Man and Cybernetics, IEEE Transactions on SMC-3 610–621
[12] Bossu D, Guyon I M and Vapnik V N 1992 Proceedings of the Fifth Annual Workshop on Computational Learning Theory COLT ’92 (New York, NY, USA: ACM) pp 144–152
[13] Powers D M W 2011 Journal of Machine Learning Technologies 2 37–63
[14] He H and Garcia E A 2009 IEEE Transactions on Knowledge and Data Engineering 21 1263–1284 ISSN 1041-4347
[15] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E 2011 Journal of Machine Learning Research 12 2825–2830
[16] Ficarra E, Cataldo S D, Acquaviva A and Macii E 2011 IEEE Trans. Biomed. Engineering 58 1421–1429