Automatic Classification of Cell Patterns for Triple Negative Breast Cancer Identification

Juan Luis Fernández-Martínez*, Ana Cernea*, Enrique J. deAndrés-Galiana*,
Primitiva Menéndez-Rodríguez†,
José A. Galván‡ and Carmen García-Pravia†
*Mathematics Department, Oviedo University,
Email: jlfm@uniovi.es, cerneadoina@uniovi.es, ej.deandres@gmail.com
†Hospital Universitario Central de Asturias, Oviedo
Email: tiva@hca.es, carmen.garciapravia@gmail.com
‡Institute of Pathology, University of Bern
Email: josegalvan6@hotmail.com

***Abstract*—This paper is devoted to presenting a methodology in Artificial Intelligence and Cognition for the optimization of basal cell patterns classification. Different unsupervised and supervised learning techniques are applied to the analysis, diagnosis and prognosis of cell patterns classification for Triple Negative Breast Cancers (TNBC), a group of cancers that, together with basal-like breast cancers, have a very bad prognosis. For that purpose, different machine learning algorithms are performed on histological images, and on a list of pathological and immunohistochemical variables currently used in medical practice. The main objective is to design a biomedical robot able to assist physicians with the kind of histological grade of different subgroups of TNBC samples in order to optimize the treatment protocol. The proposed methodology is performed on a database of 116 patients. The results show that pathological and immunohistochemical variables and histological images provide complementary information to improve the classification of TNBC samples.**

Keywords–Artificial Intelligence; Cognition; Cell Patterns; Triple Negative Breast Cancers (TNBC); Machine Learning.

I. INTRODUCTION

Breast cancer (or neoplasia) is a very heterogeneous disease. This term encompasses a variety of entities with distinct morphological features and clinical behaviors. For a long time, breast tumors have been classified according to their morphologic features (histologic type and grade) to ascertain prognostic outcome in patients. Subsequently, molecular markers (the expression of estrogen and progesterone receptor and human epidermal growth factor 2 receptor) were used to provide additional predictive power. Therefore, Triple Negative Breast Cancers (TNBC) refers to any breast cancer characterized by the absence of Estrogen Receptors (ER), Progesterone Receptors (PR) and Human Epidermal Growth Factor 2 Receptors (HER2). This classification is important from a clinical and therapeutic point of view, since TNBC are resistant to targeted therapies, because they do not express these receptors [13]. Statistics showed that TNBC account for approximately $15\% - 25\%$ of all breast cancer cases. Recently, a molecular classification based on gene expression profiles classified tumors into five groups that were not detected using traditional histopathologic methods. This classification includes the basal-like tumors group [17]. These tumors are defined by: (1) the lack of ER, PR, and HER2 expressions; (2) the expression of one or more high-molecular-weight/basal cytokeratins (CK5/6, CK14); (3) the lack of expression of ER and HER2 in conjunction with expression of CK5/6; and (4) the lack of expression of ER, PR, and HER2 in conjunction with expression of CK5/6. Also, from a morphological point of view both basal-like and triple negative breast cancers share a predominance of high histologic grades. The analysis of gene expression profiles showed that 77% of basal-like tumors were of TNBC phenotype [17].

The treatment for TNBC is adjuvant chemotherapy and radiotherapy. Unfortunately, response to chemotherapy does not correlate with overall survival. In addition, most recurrences are observed in TNBC during the first and third years after therapy, and most deaths take place in the first five years. The survival decreases after the first metastatic event. Therefore, in this heterogeneous group of tumors, new identification and classification techniques are necessary to establish a better diagnosis and prognosis, and to outline appropriate therapies. [6].

The main objective of this research is to design a biomedical robot able to help physicians with the diagnosis of different subgroups of TNBC in order to optimize their treatment protocol. The first aim of this research is to analyze the possibility of performing an automatic histological grade prediction using different biometric attributes of TNBC images and also a list of currently-used pathological and immunohistochemical variables. The methodology used in this paper is inspired by previous research works [7][8]. The preliminary conclusion of this study is that the use of both pieces of information (immunohistochemical markers and histological images) might improve the accuracy of TNBC histological grades classification and survival.

The structure of this paper is as follows: Section II describes the database of histological images and pathological, clinical and immunohistochemical variables; Sections III and IV describe the machine learning methods used in both cases. Section V presents the histological image attributes, and finally, Section VI draws the conclusion and future research work.

II. Database Description

A. Histological Images

A cohort of 105 Caucasians women diagnosed in the Hospital Universitario Central de Asturias (Spain) with TNBC and ages between 30 and 94 years were enrolled in this study, which was developed in accordance with the last version of the Helsinki Declaration of 1975 [21]. Tumor samples were obtained from surgical resection. They were fixed in 10% formaldehyde and paraffin embedded, then cut $4\mu m$ thick, mounted on treated slides, and stained with H&E stain (Hematoxylin and eosin stain). Finally, the sections were studied and photographed at two different resolutions (100X and 400X) using an Olympus light microscope. Most of the cancers in this cohort were classified in histological degrees 2 (20 samples) and 3 (89 samples), and only two samples were in degree 1. Also, a few samples have a histological degree, which is unknown. This methodology will be used in the future to asses the histological grade of the TNBC samples, and to analyze the possibility of predicting the patients survival.

B. Pathological and Clinical Variables

The pathological and clinical variables description is important to understand the different classification problems involved in this analysis, and the way the histological grade of the TNBC is established in medical practice. The Nottingham Histologic Score system (the Elston-Ellis modification of Scarff-Bloom-Richardson grading system [2], [5]) has been applied to establish the histological grades of the TNBC cancers. This system is based on the ability of the tumor to form structures similar to the ducts where the tumor has originated, on the similarity between the cancer cells and the original benign cells and finally on its proliferative activity. The histological grade will be used as the class for the different machine learning classification problems presented in this paper.

In order to assign the grade score, several factors are taken into account: **(1) Tubular structure formation**: the score increases with the percentage of tumor area forming glandular/tubular structures, as follows: score 1: $> 75\%$ of tumor area forming glandular/tubular structures; score 2: 10% to 75% of tumor area forming glandular/tubular structures; score 3: $< 10\%$ of tumor area forming glandular/tubular **(2) Nuclear pleomorphism**: the score increases with variation of size and shape of cells, as follows: score 1: Small nuclei with little increase in size in comparison with normal breast epithelial cells, regular outlines, uniform nuclear chromatin, little variation in size; score 2: Cells larger than normal with open vesicular nuclei, visible nucleoli, and moderate variability in both size and shape; score 3: Cells with vesicular nuclei, often with prominent nucleoli, exhibiting marked variation in size and shape. **(3) Mitotic rate**: The mitotic count score depends on the field diameter of the microscope used by the pathologist. The pathologist will count how many mitotic figures are seen in 10 high power fields. Using a high power field diameter of 0.50 mm, the criteria are as follows: score 1: less than or equal to 7 mitoses per 10 high power fields; score 2: 8-14 mitoses per 10 high power fields; score 3: equal to or greater than 15 mitoses per 10 high power fields.

The final total score is used to determine the histological grade of a TNBC sample, as follows:

- Histological grade 1: tumors with a total score between 3 and 5;
- Histological grade 2: tumors with a total score between 6 and 7;
- Histological grade 3: tumors with a total score between 8 and 9.

In the present case, most TNBC samples were classified with grades 2 (20 samples) and mainly 3 (89 samples). Higher scores are usually correlated with the worse prognostic.

Other currently used variables include: **(1) TNM stage**: The pathologic stage of breast cancer takes into consideration the tumor size (T) and the presence of any lymph nodes metastases (N) or distant organ metastases (M). **(2) Differentiation**: it is a combined measure of the tubular formation and the pleomorphism. It is a descriptor provided by the medical experts based on visual inspection of the histological images. This histological variable is expected to be highly correlated in medical practice to the histological grade of the different TNBC samples. **(4) Vascular and perineural invasion**: binary variable indicating the presence or absence of tumor cells inside the vessels and nerves, respectively. **(5) Necrosis**: binary variable indicating the presence of death cells. This variable is correlated with the TNBC aggressiveness. **(6) In situ component**: binary variable indicating the absence of invasion of tumor cells into the surrounding tissue. Most of these variables are provided by the pathologist by visual inspection of the TNBC images.

C. Immunohistochemical variables

The following immunohistochemical variables are also currently monitored: **(1) Estrogen receptors** (ER), Progesterone receptors (PR) and Androgen receptors (AR) nuclear expression: Hormone receptor status is important because these variables serve to decide whether the cancer is likely to respond to hormonal therapy or other treatments. **(2) Human epidermal growth factor receptor 2** (HER-2): HER2 testing is performed to assess prognosis and to determine suitability for trastuzumab therapy. **(3) Ki67 expression**: The percentage of Ki-67 ($< 2\%$, $2 - 20\%$ and $> 20\%$) can be used to aid in assessing the proliferative activity of normal and neoplastic tissue. **(4) Bcl-2 expression**: Bcl-2 is specifically considered as an important anti-apoptotic protein and classified as an oncogene. Bcl-2 expression is associated with a better prognosis [4]. **(5) E-cadherin expression**: Reduction or loss of E-cadherin expression is associated with invasive carcinoma and possibly metastasis in a variety of carcinomas [16]. **(6) P53 expression:** is also a maker used in breast cancer, but its significance in predicting clinical outcome remains controversial. **(7) CK-5/6** and **CK-14 expression**: are helpful markers in the identification of breast cancer with a basal phenotype. **(8) COL11A1**: is a stromal marker of invasion [10].

All these variables but COL11A1 have been described in the literature as useful for TNBC description.

It is important to note that the histological grade estimation of the TNBC samples is established in medical practice as it has been shown in this section, using the pathological, clinical and immunohistochemical variables that are measured for each patient, and also by visual inspection of the histological images after surgical resection. Therefore, and up to our knowledge, no biomedical robot exists to automatize these approaches, to integrate both types of information, and to assist physicians with the diagnostic. This decision has very important consequences on the prescribed treatment for the patient.

III. Machine learning using pathological and immunohistochemical variables

The aim of this section is to analyze the most discriminatory pathological and immunohistochemical variables of the histological grade. Data preprocessing includes in this case the imputation of the clinical variables that have not been measured for some patients in real practice and the normalization of these variables in the interval $[0, 1]$ according to their own empirical cumulative distribution. The imputation algorithm that is used is based on the nearest neighbor estimator. Its workflow is as follows: (1) Finding the subset S_{fi} of samples (patients) that are fully-informed for all the control variables. (2) For each patient k that is not fully-informed, finding the set of variables $\mathbf{m}_k(var_1 : var_q)$ that are missed. These variables are interpolated using the values of the same variables corresponding to the nearest fully-informed patient f_k in S_{fi}:

$$\mathbf{m}_k^*(var_1 : var_q) = \mathbf{m}_{f_k}(var_1 : var_q). \tag{1}$$

In order to measure the similarity between patients we use the cosine criterion induced by the Euclidean scalar product defined over the set of fully-informed variables in the current sample (patient):

$$cos(\mathbf{m}_k, \mathbf{m}_j) = \frac{\mathbf{m}_k \cdot \mathbf{m}_j}{\|\mathbf{m}_k\|_2 \|\mathbf{m}_j\|_2}, \tag{2}$$

where $\mathbf{m}_k$ and $\mathbf{m}_j$ stand for the vectors of fully-informed variables in patients k and j. Once the variables are imputed, the normalization of variable $\mathbf{var_j}$ is based on the p-percentile concept: $P(\mathbf{var_j} \leq c_p) = p$, by assigning the probability p to the value c_p of the attribute $\mathbf{var_j}$. Figure 1 shows the normalized pathological and immunohistochemical variables. The samples are arranged by their histological grades (2 to 3) beginning by the top of the image. It can be observed the high variability of these variables within the different classes (histological grades).

To perform machine learning, we have first used feature selection methods to finding the minimum-size list of most discriminatory variables. For that purpose, we defined the Generalized Fisher's ratio of the attribute j, for a binary classification problem, as follows [9]:

$$FR_j = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2}, \tag{3}$$

where is μ_{j1} a measure of the center of the distribution of the attribute j in class i, and σ_{ji} is a measure of its dispersion within the class i. The Fisher's ratio (GFR) can be also generalized for multiclass classification as follows:

$$GFR_j = \sum_{i=1}^{N_c} \sum_{k=i+1}^{N_c} \frac{(\mu_{ji} - \mu_{jk})^2}{\sigma_{ji}^2 + \sigma_{jk}^2}, \tag{4}$$

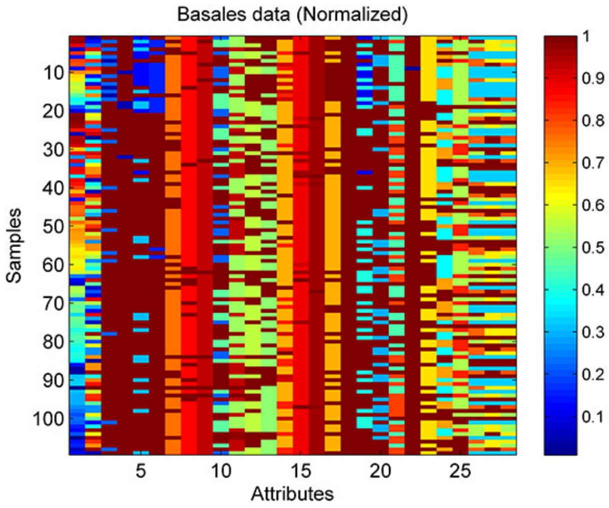

Figure 1. Normalized pathological and immunohistochemical array. The samples are ordered by their histological grade (first 20 samples with degree 2, followed by 89 samples with histological degree 3.

where N_c is the number of classes, j is the attribute index and i, k the classes indices. This feature selection method looks for attributes that are homogenous within each class (low intra-class dispersion) and show a high separation between the center of the corresponding distributions (inter-class distance). Most discriminatory attributes correspond to higher Fisher's ratios. The algorithm to find the minimum-size list of features is based on Recursive Feature Elimination, that is:

1) Attributes are ranked by the decreasing value of their Fisher's ratio.
2) Beginning by the tail of the list we calculate the accuracy of the different set of attributes, that are formed by dropping one attribute at each time. The set with the optimum accuracy and minimum size is therefore selected.
3) Finally, the accuracy of this reduced set of the attributes in the class prediction is based on Leave-One-Out method, using the average distance on the reduced set of attributes. The class with the minimum distance is assigned to the sample test. The average accuracy is calculated by iterating over all the samples.

The classification problem is linearly separable when this simple algorithm provides high accuracies. In the case where these accuracies decrease, other nonlinear classification algorithms should be used instead. If despite all these modifications, there is no improvement in the accuracy, this would mean that the data set (data and class) is noisy.

Table 1 shows the list of attributes selected by the Fisher's ratio (FR) analysis using the median and the mean to describe the centers of the distribution of the corresponding classes. These attributes are ranked by decreasing discriminatory power (Fisher's ratio) in the case of the median. In the case of the mean, the Ki67 expression should be in the third position. Five of the eight attributes in these lists are in common: *Mitotic count, Differentiation, AR expression, Ki67 expression and Tubular Formation*. The reduced base of features with the highest accuracy $(96, 4\%)$ is composed by the four first

Table I. List of most discriminatory attributes and their corresponding Fisher's ratios using the median and the mean.

Attributes	FR median	FR mean
Mitotic count (10HPF)	4.56	2.00
Differentiation	4.55	2.41
AR expression	2.60	0.64
Tubular Formation	2.46	0.56
Insitu	2.06	-
T	2.05	-
N	1.99	-
Ki67 expression	1.78	0.94
pro-CollA1 intensity	-	0.24
Bcl2 expression	-	0.22
pro-Coll1A1 Score	-	0.21

markers (*Mitotic count, Differentiation, AR expression, and Tubular Formation*). Only four patients are wrongly predicted using these attributes. Also, using the mean the two first markers (*Differentiation and Mitotic count*) provide an accuracy of 94.4%. Other list of features with high accuracy (93.6%) includes HER2, PR expression, nipple and/or skin invasion, ER, AR and Ki67 expressions. Interesting, the main attributes in these lists coincide with those used by medical experts to asses the histological grade of TNBC samples, nevertheless, these lists also show other attributes that are important for this automatic classification and were not directly used by the pathologists.

We have also analyzed the possibility of predicting the median survival of the different patients. This analysis has shown that the best markers to predict survival (with 78% of accuracy) are: E-cad expression, tumor size, perineural invasion, tubular formation, differentiation, and TNBC subtype. The accuracy of this prediction is lower than in the former case (histological grade), showing that this problem is not very well linearly separable using these attributes. Other additional variables, such as the kind of treatment followed by the patient, should be also used. The use of a nonlinear neural network classifier (extreme learning machine) [12] improved the accuracy of the prediction till 84%.

IV. Machine learning using histological images

The second aim of this research is to analyze the possibility of performing an automatic histological grade prediction using different biometric attributes of TNBC images corresponding to different histological grades taken at two different resolutions. These pattern images have been chosen by expert pathologists in this field. Figure 2 shows different histological images at two different resolutions for cancers in degrees 1, 2 and 3. It can be observed that the main differences reside on the histological variables that have been described in the previous section, that are visually assessed by medical experts. The question resides in the possibility of capturing these characteristics using image processing techniques and machine learning.

A. The automatic image classification problem

The automatic image recognition problem consists in assigning a class to a new incoming image $I \notin B_d$, given a database of TNBC color training images

$$B_d = \left\{ I_k \in S_{(n,m,3)} \left(\mathbf{N} \right) : k = 1, \ldots, N \right\}, \qquad (5)$$

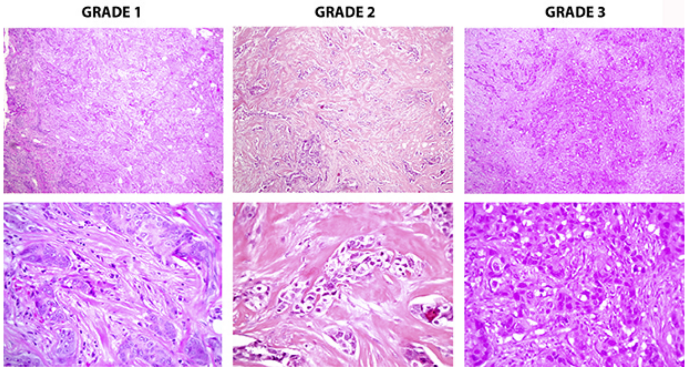

Figure 2. Basal images at two resolutions (100x and 400x magnification) for TNBC with three different histological grades. In the present case the histological grade 1 is very bad represented (only two samples).

that are characterized by a set of histological grades (labels) annotated by medical experts

$$C_d = \left\{ c_k \in \{1, 2, ..., N_c\}, k = 1, \ldots, N \right\}. \qquad (6)$$

In this definition, $S_{(n,m,3)}$ is the space of color images of size $m \times n$, and N_c is the number of classes (3 in this particular case). To perform the classification it is necessary to construct a learning algorithm $C^* : S_{(n,m,3)} \rightarrow C_d$ for the class prediction:

The classification is based on a nearest neighbor algorithm [15]:

1) First, finding the image $I_k \in B_d$ such as:

$$d\left(I, I_k\right) = \min d\left(I, I_j\right), I_j \in B_d, \qquad (7)$$

where d is a suitable distance (or norm) criterium defined over $S_{(n,m,3)}$.

2) Once this image has been found, assigning the class as follows: $C_I^* = C_{I_k} = C_k$.

The images are represented by a feature vectors calculated for each individual method of analysis (or attribute). Naming $\mathbf{v}_i^k \in \mathbf{R}^{s_k}$ the feature vector of image I_i according to the attribute k, the distance between two images I_i and I_j is defined as follows:

$$d\left(I_i, I_j\right) = \|\mathbf{v}_i^k - \mathbf{v}_j^k\|_p, \qquad (8)$$

where p is a certain norm defined over the k-attribute space $(\mathbf{R}^{s_k})$.

The final classification will be performed by consensus [14]:

1) From every individual non-supervised classifier built using the different attributes, we retain the first N_f images that are closer to I. Based on this classification a matrix $M \in M_{Nf \times Na}$ is built, containing the N_f image candidates for each of the N_a attributes and their corresponding histological grades. The score of image I according to the attribute j $(j = 1, N_a)$ to belong to the class k $(k = 1, N_c)$ is established as follows:

$$s_{jk} = \frac{1}{f_k} \frac{N_{jk}}{N_f}, \qquad (9)$$

where f_k is the sampling frequency of class k in the training database (examples) and N_{jk} is the number of images belonging to class k within the N_f candidates found for attribute j.

2) The final score for a new incoming image I to belong to class k is calculated as follows:

$$S_k = \sum_{j=1}^{N_a} s_{jk} w_j = \mathbf{s}_{jk} \cdot \mathbf{w}, \ k = 1, N_c \qquad (10)$$

where $\mathbf{s}_{jk}$ is the score assigned by attribute j to class k and $\mathbf{w}$ is a vector of weights corresponding to the trust factors assigned to any individual classifier (attribute). After calculating the scores for all the classes, the final classification of the test image I is performed by selecting the class with the major score. Eventually, the number of candidate images (N_f), the sampling frequencies (f_k), and the trust factors ($\mathbf{w}$), can be optimized (supervised learning) using global algorithms, such as PSO.

V. Image Attributes

In this paper, we have used the following list of attributes, statistical based (histogram and variogram), spectral (discrete cosine transform), and image segmentation/regional descriptors (edges, texture and Zernike Moments). In this case all attributes will be calculated as global descriptors since TNBC image comparison should not be pixel-based.

A. PCA analysis using attributes of the histological images

In this section, we analyze the possibility of discriminating the different histological grades of the TBNC samples by means of unsupervised classification using the Principal Component Analysis (PCA). PCA aims at finding the orthogonal basis by diagonalizing the experimental covariance matrix of training images [18]:

$$S = \sum_{k=1}^{N} (X_k - \mu)(X_k - \mu)^T, \qquad (11)$$

where $X_k \in \mathbf{R}^{Npixels}$ transformed into $1-D$ column vectors, $\mu = \frac{1}{N}\sum_{k=1}^{N} X_k$ is the images sample mean, N is the number of sample images contained in the learning database, and N_{pixels} is the number of pixels of each image. The eigenfaces u_k are the eigenvectors of S, corresponding to the largest eigenvalues. The dimensionality reduction from N_{pixels} to q parameters, is obtained by retaining the q first eigenfaces $\mathbf{u}_k$, spanning most of the database variability. Figure 3 shows the PCA plot in two dimensions (two first PCA coordinates) of the different TNBC images at 10X resolution. We also show the TNBC samples that have positive androgen receptors (AR). This information is important since it implies a different type of TNBC (apocrine carcinoma). Apocrine carcinoma is a subtype of TNBC that expresses androgen receptor (AR), but often lacks estrogen receptor (ER) and progesterone receptor (PR). It is possible to observe that TNBC samples with $HG = 2$ are mainly located on three different clusters, surrounded by samples with $HG = 3$. Also, most of the $HG2$ samples correspond to apocrine type. Taking this fact into account, it seems that non-apocrine $HG2$ samples are only located in very restricted areas of the PCA diagram. Figure 4 shows the 10 first mean PCA coefficients for TNBC images with histological grades 2 and 3. It can be observed that biggest differences occur high-order harmonics (4th). This attribute is expected to have a medium discriminatory power on TNBC images.

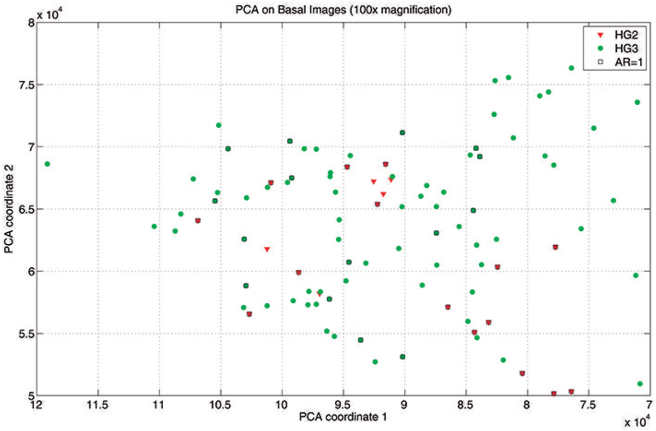

Figure 3. PCA plot (two first PCA coordinates) of the basal images corresponding to the histological grades 2 (HG2) and 3 (HG3), at 100x magnification. We also show the samples with positive androgen receptor ($AR = 1$).

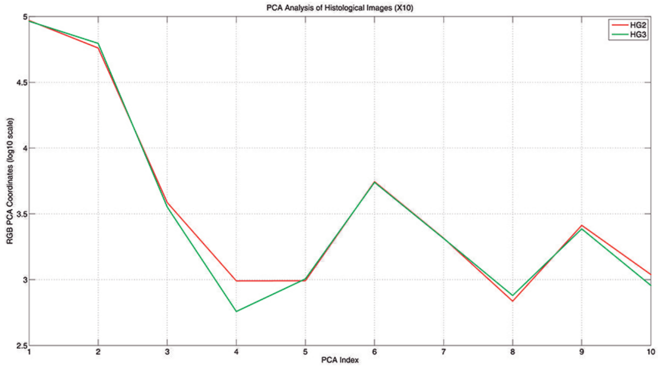

Figure 4. Mean 10 first PCA coefficients for HG2 and HG3 images. It can be observed that the biggest difference occurs for PCA number 4. PCA with higher indexes correspond to increasing high frequency harmonics in the images.

B. Color Histograms

An image histogram describes the frequency of the brightness in the image. The shape of the histogram provides information about the nature of the image [20]. For a gray-scale digital image I the histogram represents the discrete probability distribution of the gray-levels in the image. For this purpose the gray-scale space ($[0, 255]$ for an $8-\mathrm{bit}$ image) is divided into L bins, and the number of pixels in each class n_i, $(i = 1, L)$ is calculated. In this case the attribute vector has dimension L:

$$H_I = (n_1, ..., n_L). \qquad (12)$$

Relative frequencies can be also used by dividing the absolute frequencies n_i by the total number of pixels in the image. In the case of RGB images, the histogram is calculated for each color channel I_R, I_G and I_B, and then all the channels histograms are merged together, as follows:

$$H_I = \left(H\left(I_R\right), H\left(I_G\right), H\left(I_B\right)\right). \qquad (13)$$

Figure 5 shows the relative histograms of the color channels for TNBC images with HG2 and HG3. It can be observed that the major differences occur in the green channel, being its relative frequency lower than those of the red and blue channels. This attribute is expected to have a medium discriminatory power for TNBC images.

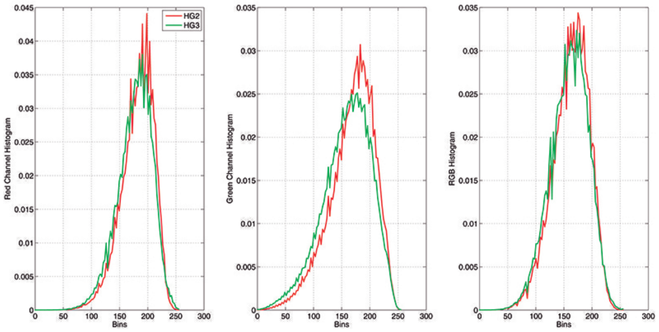

Figure 5. Mean histograms for HG2 and HG3 images. We show a different histogram for each color channel. The biggest differences between classes HG2 and HG3 occur in the green color channel.

C. Variogram

The variogram of an image describes the spatial distribution in each color channel. In spatial statistics the variogram describes the degree of spatial dependence of a spatial random field or stochastic process, the gray-scale in this case. For a given value of vector h, defined by a modulus and direction, the variogram is an index of dissimilarity between all pairs of values separated by vector h. The omnidirectional variogram is the mean of the p-absolute difference between the color values of the $N(h)$ pairs of pixels that are located at the same distance h:

$$\gamma_i(h) = \frac{1}{N(h)} \sum_{k=1}^{N(h)} |c_i(x_k) - c_i(x_k + h)|^2. \qquad (14)$$

To compute the variogram each color channel (matrix) is transformed into the corresponding color vector $c_i(x)$. Typically $N(h)$ is limited to one third of the total number of pixels. The number of classes that have been considered in this case was $N(h) = 100$. Variograms are usually used to analyze spatial continuity and anisotropies. The sill of the variogram is related to the color channel variability, its range to the spatial continuity, and its nugget (origin value) to the image low scale variabilities. Figure 6 shows the omnidirectional variograms of the three color channels for TNBC images with HG2 and HG3. It can be observed that the major differences occur in all the channels, being the blue and green the most discriminatory with respect to this attribute. The green channel also shows the biggest nugget. This attribute is expected to have a high discriminatory power for TNBC images.

D. Texture Analysis

Texture analysis of an image consists in analyzing regular repetitions of a pattern. In this paper, we use the spatial gray level co-occurrence matrix to describe an image texture. The gray level co-occurrence matrix (GLCM), or spatial dependence matrix of an image I is an estimate of the second-order joint probability function $P_{d,\theta}(i,j)$ of the intensity values of two pixels i and j located at a distance d apart (measured in number of pixels) along a given direction θ. Typically the GLCM is calculated for different pairs of d and θ. Different statistical moments can be calculated from the GLCM matrix, such as contrast, homogeneity, squared energy, correlation and entropy [1]. In the present case, we have used a lag $d = 1$ for the directions 0, 45, 90 and 135. Figure 7 shows the texture moments of the three color channels for TNBC images with HG2 and HG3. The conclusions are similar than in the case of variogram. This attribute is expected to have a high discriminatory power for TNBC images.

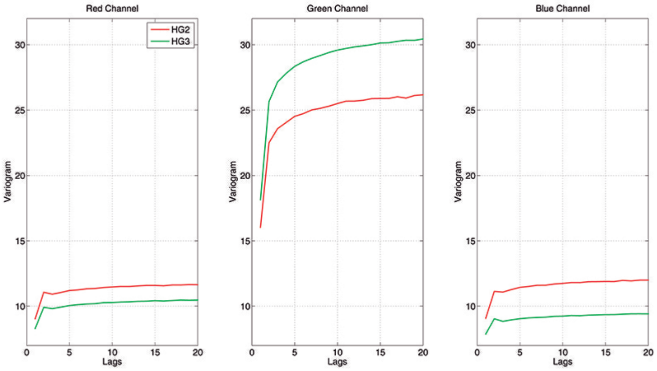

Figure 6. Mean variograms for HG2 and HG3 images. The biggest differences in the sill occur in the green and blue channels.

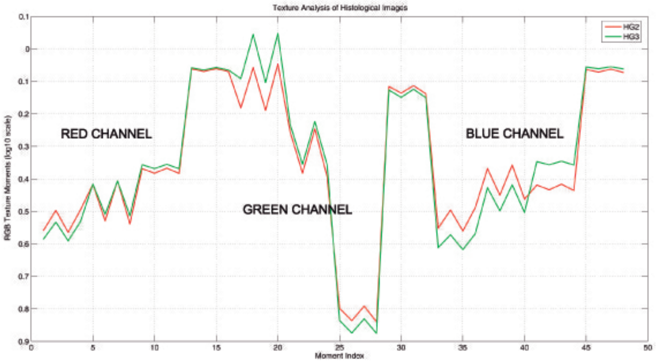

Figure 7. Mean texture coefficients for HG2 and HG3 images. The biggest differences also occur in the green and blue channels.

E. Edges Detection

Edges are determined by sets of pixels where there is an abrupt change in intensity. If a pixel's gray level value is similar to those around it, there is probably not an edge at that point. However, if a pixel has neighbors with widely varying gray levels, it may represent an edge. Thus, an edge is defined by a discontinuity in the gray-level values. More precisely, we can consider an edge as a property associated to a pixel where the image function $f(x, y)$ changes rapidly in the neighborhood of that pixel. Related to f, an edge is a vector variable with two components: magnitude and direction. The edge magnitude is given by the gradient of the pixels intensities function, and its direction is perpendicular to the gradient's direction:

$$|\nabla f(x,y)| = \sqrt{\frac{\partial f}{\partial x}^2 + \frac{\partial f}{\partial y}^2}, \tag{15}$$

$$\theta(x,y) = arctg\left(\frac{\partial f}{\partial y}, \frac{\partial f}{\partial x}\right) \pm \frac{\pi}{2}. \tag{16}$$

To compute the partial derivatives of $f(x,y)$ we have used the Canny edge detection operator [3], which is one of the most commonly used in image processing, due to its property of detecting edges in a very robust manner in the case of noisy images. The edge detection algorithm provides an image of the same size than the original image on which this analysis is performed. To produce the edge attributes we use a compression of edge image using the DCT. In the present case this analysis provides an attribute vector of dimension 48 for each image. Figure 8 shows the DCT-edges moments of the three color channels for TNBC images with HG2 and HG3. The main differences occur for the first coefficients in each color channel. This attribute is expected to have a low/medium discriminatory power for TNBC images.

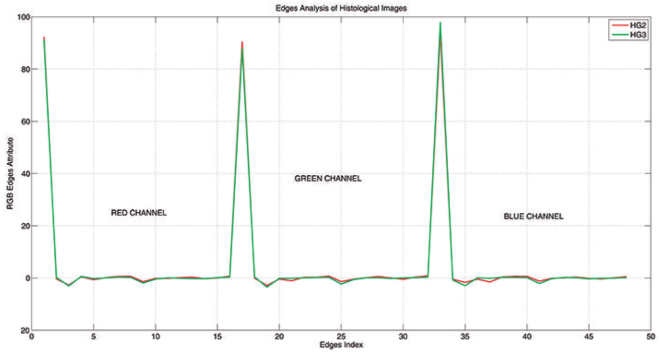

Figure 8. Mean edges vectors (after compression by the DCT) for HG2 and HG3 images. In this case we plot consecutively the first 16 DCT-edges coefficients of each color channel. No big differences are visually observed.

F. Discrete Cosine Transform (DCT)

DCT is a free-covariance model reduction technique that attempts to decorrelate 2D images by projecting the rows and columns of the incoming image into cosines of increasing frequency [11]. DCT is a discrete Fourier transform that expresses a signal in terms of a sum of sinusoids with different frequencies and amplitudes. For an image I_k the DCT is defined as follows:

$$D(u,v) = c(u)c(v)\sum_{i=0}^{s-1}\sum_{j=0}^{n-1} D_{(i,j)} \tag{17}$$

$$D_{(i,j)} = I_k(i,j) \cdot \cos\frac{\pi(2i+1)u}{2s}\cos\frac{\pi(2j+1)v}{2n}, \tag{18}$$

with $u = 0, ..., s-1$, $v = 0, ..., n-1$, and

$$c(\alpha) = \begin{cases} \frac{1}{\sqrt{N}}, & if \quad \alpha = 0, \\ \sqrt{\frac{2}{N}}, & if \quad \alpha \neq 0. \end{cases} \tag{19}$$

N is either the number of rows (s) or columns (n) of the image. The DCT can be expressed in matrix form as an orthogonal transformation [7].

$$D_{CT} = U_{DC} I_k V_{DC}^T, \tag{20}$$

where matrices U_{DC} and V_{DC} are orthogonal. This transformation is separable and can be defined in higher dimensions. The feature vector of an image I_k is constituted by the $q_1 - q_2$ block of D_{CT}, $D_{CT}(1:q_1, 1:q_2)$, where q_1, q_2 are determined by energy reconstruction considerations using the Frobenius norm of the image I_k. Figure 9 shows the DCT coefficients of the three color channels for TNBC images with HG2 and HG3. As in the previous case the main differences occur for the first coefficients in each color channel. This attribute is expected to have a low/medium discriminatory power for TNBC images.

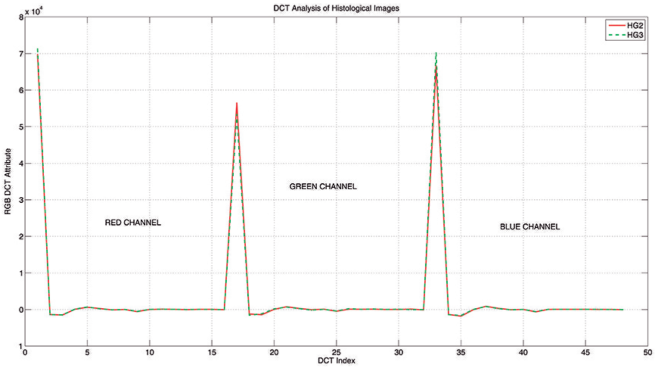

Figure 9. Mean DCT coefficients for HG2 and HG3 images. In this case we plot consecutively the first 16 DCT coefficients of each color channel. No big differences are visually observed.

G. Zernike Moments

Zernike polynomials are a sequence of polynomials that are orthogonal on the unit disk, and are widely used as basis functions for image analysis. Due to the orthogonality of Zernike polynomials, Zernike moments are image descriptors used in many applications due to their properties of orthogonality and rotation invariance. In biomedical applications, Zernike moments have been used as shape descriptors to classify benign and malignant breast masses [19]. Figure 10 shows the Zernike moments for the TNBC images of degree 2 and 3 for polynomials of order 10. This analysis provided an attribute of dimension 36x3 for each color image in this case. Although Zernike moments has been previously applied as shape descriptors to classify benign and malignant breast tissues [19], the differences do not seem very important in this particular case and occur mainly for higher order polynomials in the green and blue channels. This attribute is expected to have a medium discriminatory power.

Finally, as a compendium of all these analysis, Figure 11 shows the PCA plots in 2D (similar to figure 3) of all the attributes that have been commented. It can be observed that the HG2 and HG3 samples are located differently in each of these diagrams. Using the above mentioned image attributes, the unsupervised machine learning algorithm commented in section IV-A provided an accuracy of 86.8%, which is slightly higher than the majority voting algorithm (80%). Future research will be devoted to this important problem.

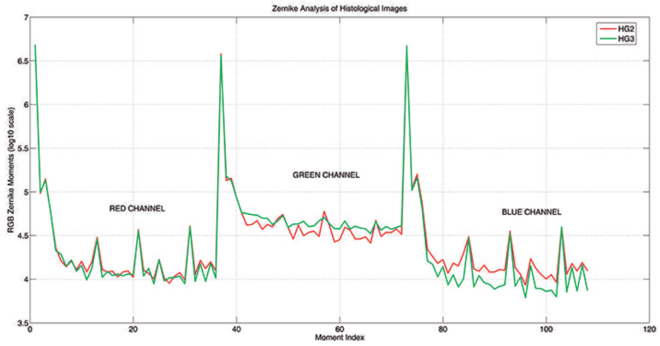

Figure 10. Mean Zernike moments for polynomials of order 10, for HG2 and HG3 images. The biggest differences occur in the green and blue channels.

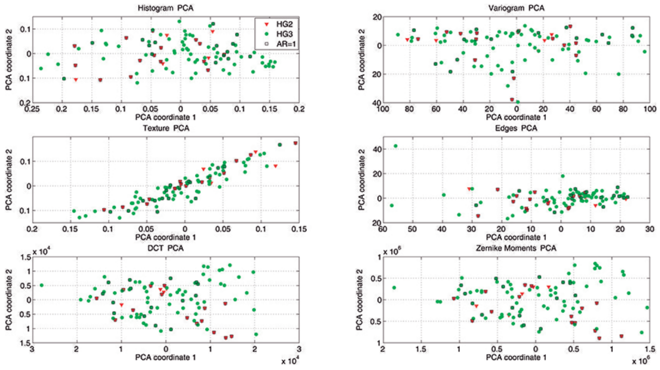

Figure 11. PCA plots (two first PCA coordinates) for all the attributes used in the analysis of the histological images.

VI. Conclusions and Future Research

In this paper, we have explored the possibility to design a biomedical robot able to assist physicians on the kind of histological grade/survival of different subgroups of TNBC samples in order to optimize their diagnosis/treatment and prognosis. Very promising preliminary results are shown using pathological and immunohistochemical variables and histological images of a cohort with 105 patients. Future research will include the possibility of using other supervised learning techniques and global optimization methods (PSO) to optimize the machine learning parameters and to improve the accuracy of the classification. Also, it is expected that the use of both pieces of information (pathological and immunohistochemical variables and histological images) will provide complementary information, improving the accuracy in the classification of TNBC samples (histological grade and survival).

Acknowledgment

The authors want to acknowledge the *Hospital Central de Asturias (Oviedo, Spain)* for providing all the facilities to acquire and treat the TBNC samples.

References

[1] M. H. Bharati, J. J. Liu, and J. F. MacGregor, Image texture analysis: methods and comparisons. *Chemometrics and Intelligent Laboratory Systems*, 72(1), 2004, pp. 57-71.

[2] H. J. Bloom and W. W. Richardson, Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years, *British journal of cancer* 11 (3), 1957, pp. 359377.

[3] J. Canny, A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6), 1986, pp. 679-698.

[4] S. J. Dawson et al., BCL2 in breast cancer: a favourable prognostic marker across molecular subtypes and independent of adjuvant therapy received, *British Journal of Cancer.* 103(5), 2010, pp. 668-75.

[5] C. W. Elston and 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. In *Histopathology*, 19, 1991, pp. 403-10.

[6] O. Fadare and F. A. Tavassoli, Clinical and pathologic aspects of basal-like breast cancers. In *Nat Clin Pract Oncol.* 5(3), 2008, pp. 149-59.

[7] J. L. Fernández-Martínez and A. Cernea, Numerical analysis and comparison of spectral decomposition methods in biometric applications. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 28(1), 2014, pp.14560-14593.

[8] J. L. Fernández-Martínez, A. Cernea, E. García-Gonzalo, J. Velasco and B. Ketan Panigrahi, Aligned PSO for Optimization of Image Processing Methods Applied to the Face Recognition Problem, In *Swarm, Evolutionary, and Memetic Computing (SEMCCO), Springer Berlin Heidelberg*, Lecture Notes in Computer Science, 8297, 2013, pp 642-651.

[9] R. A. Fisher, The use of multiple measurements in taxonomic problems. *Annals Eugen*, 7, 1936, pp. 179-188.

[10] C. García-Pravia et al., Overexpression of COL11A1 by cancer-associated fibroblasts: clinical relevance of a stromal marker in pancreatic cancer. In *PLoS One.* 8(10), 2013, e78327.

[11] Z. M. Hafed and M. D. Levine, Face recognition using the discrete cosine transform. *Int. J. Comput. Vision*, 43(3), 2001, pp. 167-188.

[12] G. B. Huang et al., Extreme learning machine: Theory and applications. *Neurocomputing*, 70, 2006, pp. 489501.

[13] E. A. Rakha et al., Breast cancer prognostic classification in the molecular era: the role of histological grade. In *Breast Cancer Research,* 12(14), 2010, pp. 12-207.

[14] L. Rokach, Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2), 2010, pp. 1-39.

[15] G. Shakhnarovich, T. Darrell, and P. Indyk, *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing)*, The MIT Press, 2006.

[16] S. M. Siitonen et al., Reduced E-cadherin expression is associated with invasiveness and unfavorable prognosis in breast cancer. *American Journal of Clinical Pathology,* 105, 1996, pp. 394-402.

[17] T. Sørlie et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. In *Proc Natl Acad Sci USA* 98(19), 2001, pp. 10869-10874.

[18] M. Turk and A. Pentland, Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991, pp. 71-86.

[19] A. Tahmasbi, F. Saki, H. Aghapanah, and S.B. Shokouhi, A Novel Breast Mass Diagnosis System based on Zernike Moments as Shape and Density Descriptors, *Proceeding of 18th Iranian Conference of Biomedical Engineering (ICBME)* , 2011, pp.100-104.

[20] S. E. Umbaugh. *Computer Vision and Image Processing: A Practical Approach Using CVIPtools*. Number ISBN 0-13-264599-8. Prentice Hall Professional Technical Reference, 1998.

[21] World Medical Association, Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects, *The Journal Of the American Medical Association (JAMA)*, 2013, Volume 310(20), pp.2191-2194.