International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 32, No. 1 (2018) 1860004 (20 pages)

DOI: 10.1142/S0218001418600042

World Scientific
www.worldscientific.com

Semi-Automatic Analysis for Unidimensional Immunoblot Images to Discriminate Breast Cancer Cases Using Time Series Data Mining

Diana M. Sánchez-Silva[*,†], Héctor G. Acosta-Mesa[*]
and Tania Romo-González[†,‡]

[*]*Centro de Investigación en Inteligencia Artificial*
Universidad Veracruzana, Xalapa Veracruz, México
[†]*Área de Biología y Salud Integral*
Instituto de Investigaciones Biológicas
Universidad Veracruzana
Xalapa Veracruz, México
[‡]*tromogonzalez@uv.mx*

Received 28 October 2016
Accepted 16 May 2017
Published 26 July 2017

Breast cancer (BC) is one of the leading causes of death in adult women worldwide and the best way to reduce mortality and improve prognosis is through early diagnosis. Thus, it is necessary to optimize diagnostic methods; one option could be the automatic detection of patterns in 1D-II. In that respect, through recent analysis of unidimensional Immunoblot Images (1D-II), it was possible to distinguish between women with and without breast disease using as a discrimination criterion the presence of autoantibodies (bands) in their blood. However, the analysis of 1D-II is a difficult task even for an expert, generating great subjectivity and complexity in the process of interpretation.

In the present study, a semi-automatic methodology for the bands' analysis contained in the 1D-II's was implemented and evaluated, the bands were extracted using digital image processing techniques. This was possible through the recognition of banding patterns represented as time series to distinguish between three classes: women with breast cancer (BC), women with benign breast pathology (BBP) and women without breast pathology (H). The classification was performed using the machine learning algorithm k-nearest neighbors (KNN) with different parameters over the time series representation.

The semi-automatic method here presented was able to reduce the time, complexity and subjectivity of the image analysis with the performance metrics compared, obtaining similar percentages for both representations. With the traditional analysis, binary representation [Accuracy 72.8%, Precision 73.42% for three classes (BC, BBP and H) and Accuracy 90.91% Accuracy 92.55% Sensitivity 93.57% and Specificity 92.99% for two classes (BC and H)], versus Time series representation [Accuracy 66.4%, Precision 67.07% for three classes (BC, BBP and H) and Accuracy 86.36% Accuracy 87.31% Sensitivity 95.86% and Specificity 85.56% for two classes (BC and H)].

‡Corresponding author.

Keywords: Breast cancer; unidimensional immunoblot; protein bands; time series data mining; semi-automatic method; digital image processing.

1. Introduction

Breast cancer (BC) is the most common pathology in women worldwide and one of the leading causes of death in adult women. According to the World Health Organization (WHO), each year 1.38 million new cases are detected and 548 thousand people die from this cause.[1] In Mexico, the BC's mortality rate has increased significantly, particularly in women between 45–54 years old, followed by women ages 35–44,[2,3] becoming the leading cause of death in women over 35 years old.[4] This occurs since the detection usually takes place in advanced stages of the disease. In this regard, it is necessary to optimize and improve diagnostic methods.

It has recently been proposed that autoantibodies are useful in diagnosis, prognosis and follow-up of patients with several diseases.[5–7] Interest surrounding the role of autoantibodies has steadily increased during the past decade, resulting in a more intense focus on their development as early biomarkers of cancer.[8–20]

In this regard, our research group noted that through the analysis of unidimensional immunoblot images (1D-II) it was possible to distinguish between women with and without breast disease as reported in Esquivel-Velazquez *et al.*[21] Our results show that IgG 1D-II with sera from women with BC, women with benign breast pathology (BBP) and without breast disease (H), when reacted with whole protein extracts of T47D cells, display a huge diversity within and between groups. None of the banding patterns analyzed were the same in two or more individuals. However, despite this huge diversity, 1D-II from BC and BBP women may be confidently distinguished from those of healthy women, reaching sensitivity values of 46–100% and specificity of 74–98%, depending on whether the immunoblots detected as few as one *High Risk Band* or more.[21] However, even with the benefits of this method, the analysis of 1D-II's was very complex since the individual strips from 150 women must be aligned in groups of about fifteen or seventeen strips, in which the bands of each strip are compared with each other, in order to decide if its identity is equal or not between them and with a control strip (a strip included in all the images). In addition, since one image includes about fifteen or seventeen strips, the bands' analysis should consider the comparison of the bands' identity between images (in our case the comparison between fifteen images). Thus, the analysis with traditional software's requires a very expert and trained eye and the analysis of a single experiment (an image) can take up to a month (Fig. 1).

Thus, Western blot is widely used in proteomics[6] and is often used in research to separate and identify proteins.[22] The inception of the protocol for protein transfer from an electrophoresed gel to a membrane by Towbin (1979) has evolved greatly[23,24] however, image analysis is still rudimentary. This is because software programs that exist today for western blot image analysis, only have the capacity to propose to the researcher potential bands found on a strip and to align the bands

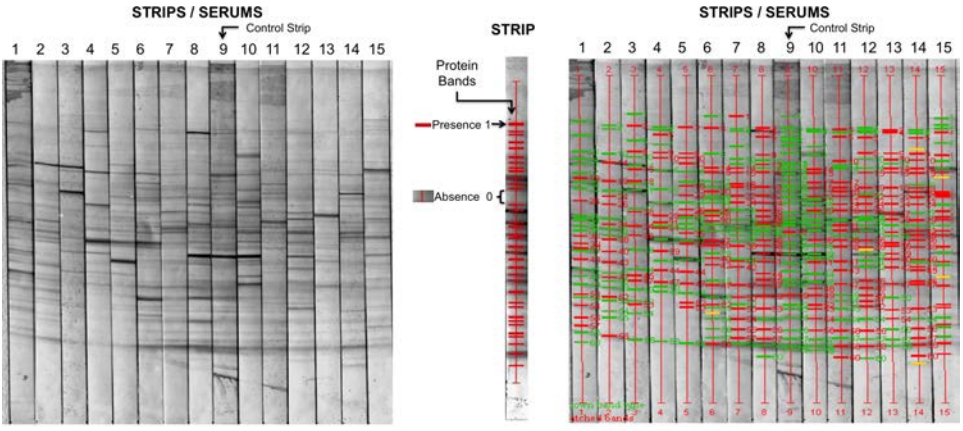

Fig. 1. (a) Example of an image includes fifteen strips, where strip 9 represents the control strip (H111). (b) Example of a strip/serum Immunoblot in which the bands are identified, reflecting the presence (1) or absence (0) of proteins. (c) The bands in each lane were selected using quantity one (green bands are known band types, red bands are bands that have been automatically matched with the known band types, yellow bands are bands that have not been matched and are unclassified).

between strips, but in the end the researcher is the one who decides whether the software has made a good detection or alignment of the bands or not. Additionally, the existing software does not allow the comparison between different images, which makes the analysis even more complicated, because it depends on an expert eye. Therefore, it is necessary to have automatic methods of analysis in order to simplify the procedure and improve the results by eliminating subjectivity, making this task easier and quantifiable.[25]

In the present study we proposed a semi-automatic method of analysis in which the 1D-I digital image is processed and classified by its banding patterns using a data mining time series approach. We explored the performance of the semi-automatic methodology in a sample of 150 female patients categorized into three classes: patients with BC patients with a BBP and healthy patients (H); and compared its accuracy, precision, sensitivity and specificity with a traditional 1D-I image device running by an expert human eye.

2. Materials and Methods

2.1. *Grouping of participants and sample sizes*

For this study we recruited BC and BBP patients at their first consultation at the Hospital General de México "Dr. Eduardo Liceaga" as reported in Romo-González *et al.*[26] The molecular profiles of the breast tumors were: 58% hormone receptor-positive, 30% HER2-positive and 12% triple-negative. We also recruited 50 women without breast pathology who volunteered to participate in the study and had blood

drawn by trained personnel in the Instituto de Investigaciones Biomédicas, UNAM. All participants were informed of the details of the study (scientific and technical basis, exclusive use of their blood sample for immunodiagnostic of BC, anonymity, confidentiality) and signed a letter of informed consent. The protocol was reviewed and approved by the Committee of Ethical Research of the Hospital General de México "Dr. Eduardo Liceaga" (DI/12/111/03/064). The study conforms to The Code of Ethics of the World Medical Association (Declaration of Helsinki), printed in the British Medical Journal (18 July 1964).

2.2. *Sample's preparation, unidimensional electrophoresis and immunoblots of the protein extracts from the T47D cell-line*

Venous blood was collected as in Romo-González *et al.*[26] Briefly, ten milliliters of blood were drawn from each participant using BD Vacutainer® kits. Sera were collected and aliquoted before storage at −80°C.

The human BC cell line T47D was cultured and harvested as established in Romo-González *et al.*[26] Briefly, cells were grown on plastic tissue culture plates in 95% humidity and 5% CO_2 at 37°C and harvested by treatment with PBS with EDTA, before pelleting and freezing at −80°C. Cells from other BC cell-lines, namely MCF7 and MDA-MB-23, were cultured and prepared in a similar manner.

Cell pellets in a denaturing lysis buffer (M urea, 4% (w/v) CHAPS, 65 mM DTT and Halt protease inhibitor cocktail), were centrifuged and supernatants were recuperated. The cell lysates were pooled, the protein concentration was measured, aliquoted and kept at −80°C until further use.

Western Blot optimization: Prior to performing the experiments, several controls were executed and the optimum parameters determined. To determine both the optimum serum and secondary antibody dilution, a series of dilutions were performed and two secondary antibodies were tested (goat anti-human IgG (H+L) and goat anti-human IgG (FC); THERMO) by Western Blot. The optimum serum dilution was determined to be 1:300 and of the secondary antibody 1:2,500. No difference was found between secondary antibodies and none detected any bands on the separated proteins from the extract when incubated alone. Likewise, a group of randomly-selected sera were probed at different dilutions against the separated proteins of the fetal calf serum used for the culture of T47D, and none of the sera recognized any bands. Additionally, two reagents for blocking the membranes (5% Svelty skimmed milk, and 2% and 5% Albumin) and two systems for detecting the presence of the bound secondary antibody (HRP-Diaminobenzidine (SIGMA) and AP-NBT/BCIP (THERMO) were probed, with no differences found among them.

The protein extract (PE) from the T47D Cell-Line was subjected to electrophoresis using polyacrilamide gels (4–20% TGX Bio-Rad) before transferring the proteins onto nitrocellulose membranes (High Bond, Amersham Biosciences). Western Blot was then performed. Before blocking the membranes with 5% skimmed milk, they were marked with two horizontal pencil lines at their upper and lower limits

which were then reversibly-stained using copper phthalocyaninetetrasulphonic acid (Sigma-Aldrich, St. Louis, MO), scanned and destained. Each membrane was cut vertically into seventeen or eighteen 4 mm wide strips. Each strip was then probed individually with the serum of a different participant. As an internal control, the same identical serum from a healthy woman (H111) was included in each set of 17–18 strips (17 sets). Bound antibodies were detected by incubation of the strips with HRP-conjugated secondary antibody.

2.3. *Traditional 1D-II analysis*

As reported in Romo-González *et al.*,[26] the strips of the Immunoblots were scanned at 300 dpi in TIF format. The digitalized images of the strips were aligned with their corresponding image of reversible-stained membrane and the "smiling effect" was corrected with Photoshop CC 2014 (Perspective Warp tool).

Figure 1(A) shows one image including fifteen strips of the immunoblots, that were compared with each other using the control strip in order to identify the total number of different bands. Banding patterns between strips form different membranes were compared using the control strip (H111) in order to identify the total number of different bands and create a binary database with the presence (1) or absence (0) of each band in each strip according to the expert criteria (Fig. 1(B)). Bands were numbered from 1 to 228 as their molecular weight (MW) decreased.

Afterwards, bands in each image were detected with Quantity-One Software (Bio-Rad).[27] Figure 1(C) shows the bands in each lane were selected, where green bands are known band types, red bands are bands that have been automatically matched with the known band types and yellow bands are bands that have not been matched and are unclassified.

2.4. *Semi-automatic 1D-II analysis using time series representation*

In order to explore the classification of banding patterns using time series data mining, a continuous representation by time series was obtained from 1D-II. In addition the results obtained with the traditional analysis (binary representation manually defined) were compared with the semi-automatic method (continuous representation). Figure 2 shows a graphic comparison between the processes carried out with the traditional analysis against the proposed semi-automatic method by time series.

The process for the continuous representation by time series is explained below:

2.4.1. *Preprocessing and transformation*

The study was conducted from 15 color images in TIF format known as unidimensional immunoblot images (1D-II) previously used in the traditional analysis. 1D-II images were represented in RGB color model, in each of the color channels or HIS (Hue, Saturation, Intensity) and in scales of gray in order to explore in which of the

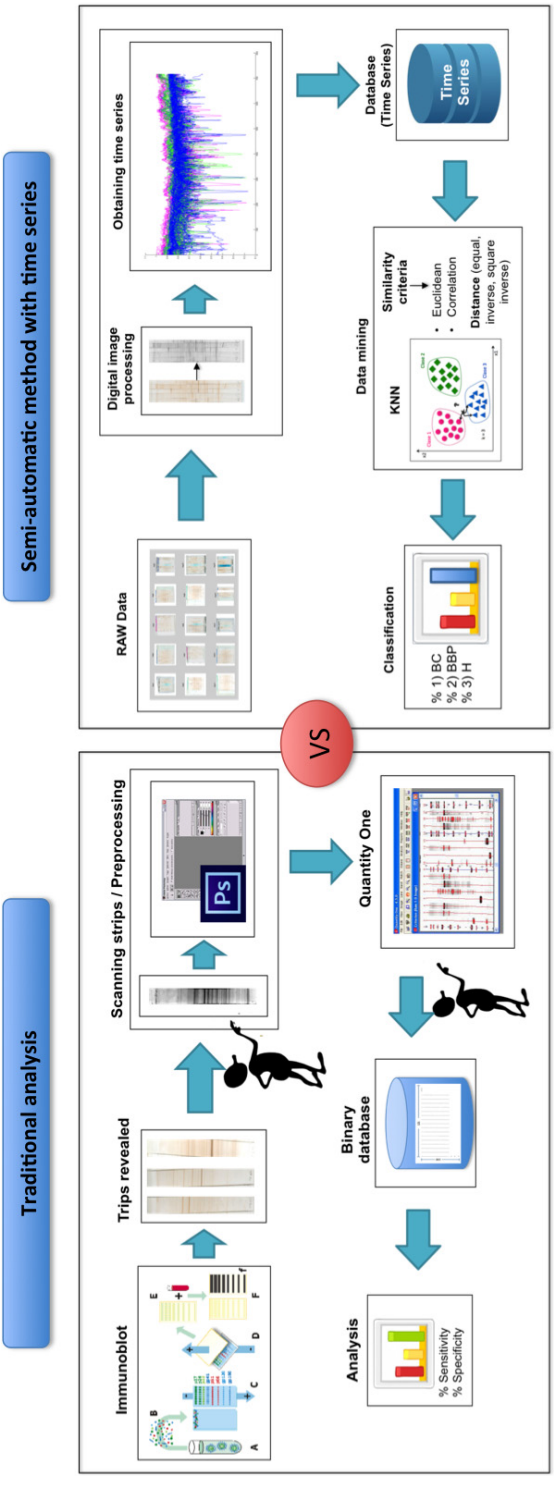

Fig. 2. Comparative image between the traditional analysis versus the semi-automatic method by time series.

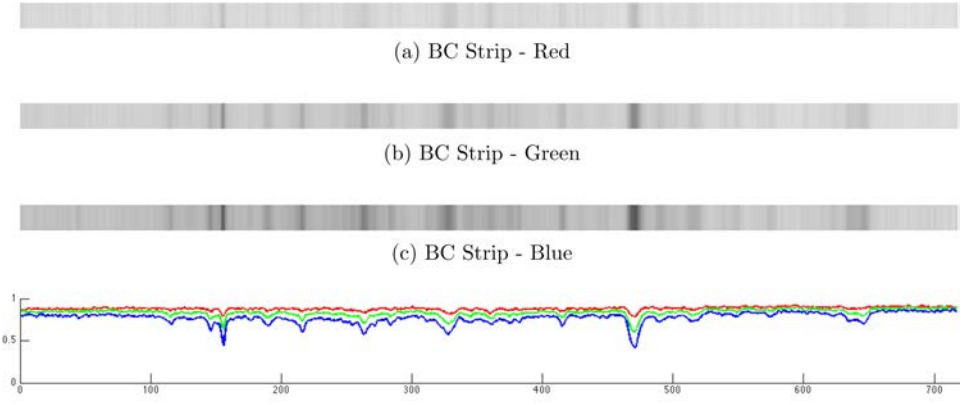

Fig. 3. Strip from serum of a BC patient seen in the three channels of RGB color model. (a) Red channel filter. (b) Green channel filter. (c) Blue channel filter. (d) Time series graph for the three channels of RGB color model. Note that in the blue channel the amplitude of the signal has the highest amplitude.

color spaces did the bands show the best definition. All the 1D-II were transformed to a blue channel of RGB color scale, since it was the channel in which the bands were appreciated with more amplitude (Fig. 3).

2.4.2. *Time series extraction*

A time series represents a continuous set of numerical data in time that is arranged chronologically, which helps to describe, explain, predict and control the processes that somehow occur over time.[28] That is, if we considered the sequence presented in the strip by western blot running as a key change over time for each pixel, the patterns can be extracted directly from the 1D-II image (Fig. 4). Each western blot strip can be thought of as a time series. Where strip (t) represents a vector of size (t).

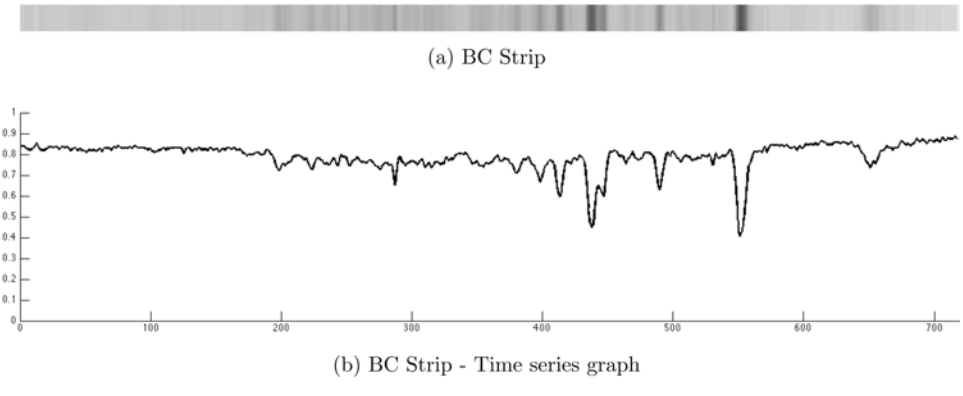

Fig. 4. Graph showing time series of a strip corresponding to a BC patient's serum.

Thus there are (t) pixels. Where $p(i)$ represents the color of the pixel (i), $i = 1, \ldots, m$. The intensity value of each pixel over time is used to construct a time series. Hence the strip dataset can be thought of as a data matrix. Where dataset (t, n) represents a stack of n strips of size (t).

The data of each strip can be extracted directly from the image marking a straight line over its central part. In order to define the region of interest (ROI) on the image strip, it was necessary that the researcher define the start and end point of the strip band on the image. Due to the difference in size of the ROIs on each 1D-I, the length of the time series was different. Thus, a cubic interpolation scaling was performed in order to standardize its length.

2.4.3. *Creation of the times series database*

With the same size time series of all the strips, a database was created with the values of the continuous data.[28,29] Figure 5 shows the time series from the 150 strips/ patients separated by class (BC, BBP and H).

2.4.4. *Classification*

The classification is a task undertaken to find common cases from a set of features within a database.[30,32] The learning process was performed using the classification algorithm *K-nearest neighbor* (KNN). This algorithm was chosen since it is a simple and efficient algorithm than works with continuous data. The main idea behind this

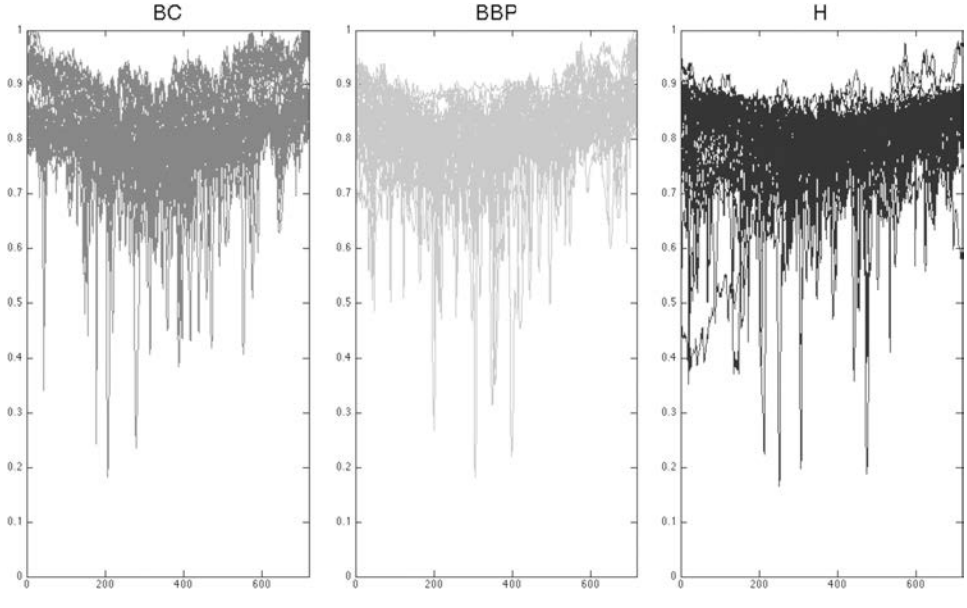

Fig. 5. Representation of time series obtained from the 150 strips separated by class: 50 time series from the BC class; 50 time series from BBP class; and 50 time series from H class.

algorithm is that a new case is classified as the closest related class under its KNN.[30,33] Due to the fact that a similarity measure can be determinant on the accuracy, precision, sensitivity and specificity results, different metrics were evaluated on this study: similarity criteria (Correlation, Euclidean), distance measure (equal, weighted inverse, weighted square inverse).

In order to evaluate the performance of the classifier, the *k-fold* cross-validation technique was used, which segments the data into k partitions of equal size.[31] During execution, one of the partitions was chosen to test while the rest was used for training. This procedure was repeated k times, where the value of $k\text{-}fold = 10$.

2.5. *Evaluation of the performance of the binary and continuous representations*

We analyzed both representations (binary and continuous). For the binary representation, since its representation is discrete, the following classification algorithms were used: J48, Naïve Bayes, KNN, Linear Discriminant, Support Vector Machine, and Multilayer Perceptron. In the case of the time series proposed approach, only KNN was used because of the continuous nature of this representation. The performance of the algorithm KNN was explored assigning values $k = 3$, $k = 5$ and $k = 10$. Similarity criteria (SC) used were: Correlation and Euclidean, and values used for the distance measure (D) were: equal (no weighting), inverse (weight is 1/distance) and square inverse (weight is $1/\text{distance}^2$).[34,35]

Correlation: Measure of statistical dependence between two variables. It is zero if and only if the variables are statistically independent. However, if the result is one, the variables are similar in shape (Eq. (1)). Here X and Y represent the registers being compared.

$$\text{Correlation} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2) * (n\sum y^2 - (y)^2)}}. \tag{1}$$

Euclidean: The Euclidean distance between any two instances is the length of the line segment connecting them (Eq. (2)). Where Xi and Yi represent the vectors of the items being compared.

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2} \tag{2}$$

2.5.1. *Classification of performance metrics*

Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset (Eq. (3)).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3}$$

Precision is calculated as the number of correct positive predictions divided by the total number of positive predictions (Eq. (4)).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{4}$$

Sensitivity is calculated as the number of correct positive predictions divided by the total number of positives (Eq. (5)).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

Specificity is calculated as the number of correct negative predictions divided by the total number of negatives (Eq. (6)).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{6}$$

where TP (true positive) denotes the number of correct positive predictions (class BC).

FP (false positive) denotes the number of incorrect positive predictions (class BC).

TN (true negative) denotes the number of correct negative predictions (class H).

FN (true negative) denotes the number of incorrect negative predictions (class H).

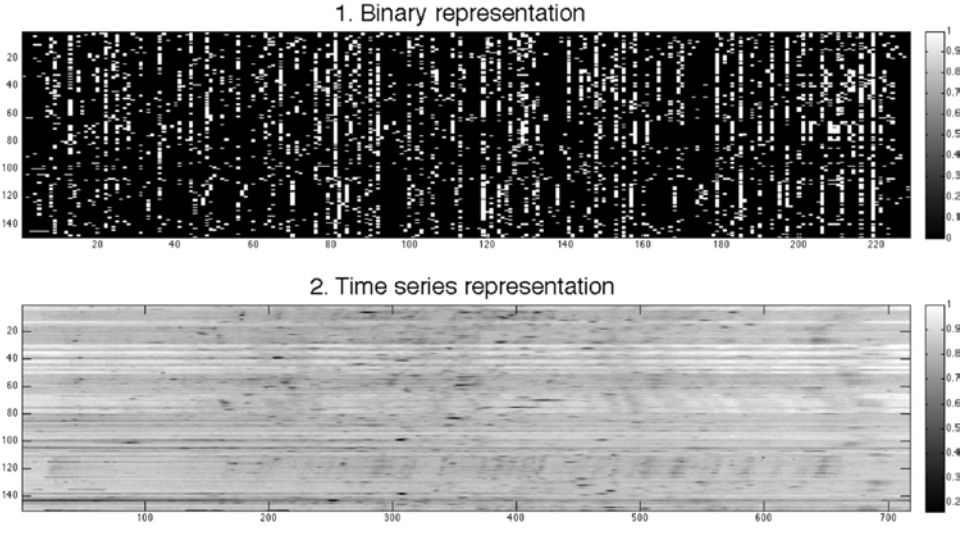

Fig. 6. Graphic representations of the data. (1) Binary representation with 228 attributes, 150 instances an 3 classes (BC, BBP, H); 2) Time series representation of with 717 attributes, 150 instances and 3 classes (BC, BBP and H).

2.5.2. *Statistical analysis*

Because data was not normally distributed, Kruskal–Wallis test followed by Tukey post-hoc tests for multiple comparisons were made. Data was analyzed using MATLAB R2009a scientific software (the MathWorksTM). Classification analysis was made using WEKA Data Mining Software.[39]

2.6. *Comparing a semi-automatic method based on time series against the traditional method*

As was mentioned above, two representations were used to evaluate and compare results: the binary representation obtained by the traditional method (helped by Quantity one software), in which the values are discrete, forming a vector of 228 positions (features); and the representation of time series (semi-automatically extracted from the images) composed for a continuous data represented as a vector of 717 positions. In both cases 150 instances and three classes (BC, BBP and H) were used. The representations of both sets of data are shown in Fig. 6.

3. Results

3.1. *Classification for binary representation*

In order to obtain the performance metrics for the binary representation, the classification algorithms J48, Naïve Bayes, KNN (SC = Correlation | Euclidean, D = equal | inverse | square inverse, $k = 3$), Linear Discriminant, Support Vector Machine and Multilayer Perceptron were used. The classifiers comparison was made considering three (BC, BBP and H) and two classes (BC and H). Table 1 shows the accuracy, precision, sensitivity and specificity percentages obtained by each classifier. The highest percentage of accuracy and precision for the evaluation of the three classes (BC, BBP, H) was obtained with KNN ($k = 3$, SC = Correlation, D = square inverse) accuracy 72.8%, precision 73.42% with the same parameters for both metrics, but the differences in the percentages were not statistically significant. Similarly, for the representation of the binary two classes (BC, H) the highest accuracy, precision and specificity percentages was obtained with KNN ($k = 3$, SC = Euclidean, D = equal), reporting accuracy 90.91%, precision 92.55% and specificity 92.99%, while the highest sensitivity percentage obtained was 93.57% with KNN ($k = 3$, SC = Correlation, D = square inverse). However, in the case of the two classes analysis the highest percentages with these classifiers were statistically significant (Table 1). It is noteworthy that all classifiers increase their accuracy and precision when only the BC and H classes were considered.

3.2. *Classification for the time series representation*

For the continuous representation by time series the KNN classification algorithm was used with both Correlation and Euclidean similarity criteria. Table 2 shows

Table 1. Percentages of the performance metrics evaluated of the binary representation considering the three classes (BC, BBP and H) and only two (BC and H).

Classification algorithm	Accuracy (BC, BBP, H)	Precision (BC, BBP, H)	Accuracy (BC, H)	Precision (BC, H)	Sensitivity (BC, H)	Specificity (BC, H)
J48	63.31 ± 16.02[1]	59.68 ± 9.29	73.84 ± 9.77	74.56 ± 9.5	72.15 ± 12.02	76.5 ± 8.8
Naïve Bayes	72.65 ± 8.41	72.31 ± 8.58	84.43 ± 4.59	85.08 ± 4.59	79.94 ± 8.02	87.82 ± 6.52
KNN ($k=3$) SC = Euclidean D = equal	68.6 ± 5.25	69.19 ± 5.76	$\mathbf{90.91 \pm 5.34}$	$\mathbf{92.55 \pm 6.9*}$	88.13 ± 10.81	$\mathbf{92.99 \pm 6.47*}$
KNN ($k=3$) SC = Euclidean D = inverse	69.4 ± 1.9	69.25 ± 2.46	86.97 ± 4.96	90.74 ± 9.28	83.06 ± 8.91	91.48 ± 8.62
KNN ($k=3$) SC = Euclidean D = square inverse	69.4 ± 8.49	70.53 ± 8.33	88.18 ± 4.62	86.75 ± 8.55	89.02 ± 5.14	87.97 ± 6.66
KNN ($k=3$) SC = Correlation D = equal	69.4 ± 4.62	69.04 ± 4.88	86.97 ± 5.9	82.76 ± 9.32	91.8 ± 7.96	82.05 ± 10
KNN ($k=3$) SC = Correlation D = inverse	70 ± 4.71	70.64 ± 4.37	84.55 ± 5.61	81.98 ± 7.98	88.37 ± 7.94	81.35 ± 9.14
KNN ($k=3$) SC = Correlation D = square inverse	$\mathbf{72.8 \pm 6.41}$	$\mathbf{73.42 \pm 5.87}$	86.06 ± 6.09	80.09 ± 9.63	$\mathbf{93.57 \pm 5.8*}$	79.19 ± 8.28
Linear Discriminant	60 ± 9.84	60.8 ± 9.7	75.76 ± 4.04	81.56 ± 11.46	71.76 ± 14.95	81.02 ± 13.13
Support Vector Machine	59.4 ± 6.26	60.55 ± 6.12	85.76 ± 7.96	87.25 ± 10.09	83.92 ± 12.54	88.15 ± 8.3
Multilayer Perceptron	69.78 ± 13.21	70.63 ± 11.72	86.78 ± 5.22	87.45 ± 4.92	89.33 ± 4.56	84.13 ± 10.29

[1]Means and standard deviation.
(*p-$value \leq 0.05$).

Table 2. Percentages of the performance metrics evaluated of the continuous representation considering the three classes (BC, BBP and H) and only two (BC and H).

Similarity Criteria	k value	Distance	Accuracy (BC, BBP, H)	Precision (BC, BBP, H)	Accuracy (BC, H)	Precision (BC, H)	Sensitivity (BC, H)	Specificity (BC, H)
Euclidean	3	Equal	64.2 ± 6.63	64.6 ± 6.72	$\mathbf{86.36 \pm 2.95}$*	$\mathbf{87.31 \pm 7.56}$	88.16 ± 8.26	$\mathbf{85.56 \pm 7.06}$*
		Inverse	64.2 ± 4.66	64.64 ± 4.61	82.12 ± 6.62	84.38 ± 8.34	79.94 ± 9.77	84.47 ± 10.42
		Square inverse	63.8 ± 4.37	63.9 ± 4.47	83.03 ± 4.09	81.19 ± 6.85	84.3 ± 5.24	81.71 ± 4.51
	5	Equal	58.4 ± 8.15	60.23 ± 6.99	85.45 ± 5.5	83.9 ± 3.46	87.15 ± 8.12	83.46 ± 5.96
		Inverse	62 ± 4.42	62.29 ± 4.69	80 ± 5.38	79.24 ± 9.51	81.86 ± 11.32	79.08 ± 9.91
		Square inverse	$\mathbf{66.4 \pm 4.2}$	66.38 ± 4.3	82.42 ± 7.67	79.68 ± 7.25	88.93 ± 12.26	76.97 ± 9.95
	10	Equal	62.8 ± 7.67	63.81 ± 7.89	77.7 ± 12.93	77.7 ± 12.93	89.32 ± 8.96	78.35 ± 9.95
		Inverse	58.8 ± 6.94	59.5 ± 7.54	80 ± 6.42	81.38 ± 6.68	79.85 ± 8.76	80.28 ± 6.98
		Square inverse	63 ± 8.29	63.45 ± 7.8	75.76 ± 6.23	75.8 ± 10.88	78.61 ± 10.45	73.64 ± 12.96
Correlation	3	Equal	63.4 ± 3.27	64.1 ± 3.29	81.82 ± 4.29	80.58 ± 7.05	87.84 ± 8.02	74.96 ± 9.56
		Inverse	64.6 ± 6.26	65.34 ± 6.3	83.64 ± 4.56	81.2 ± 7.72	87.99 ± 9	79.44 ± 8.83
		Square inverse	$\mathbf{66.4 \pm 6.02}$	$\mathbf{67.07 \pm 5.17}$	85.45 ± 4.24	81.15 ± 8.08	92.92 ± 5.87	78.52 ± 10.62
	5	Equal	60.6 ± 7.12	61.86 ± 6.66	84.85 ± 4.04	78.41 ± 5.26	$\mathbf{95.86 \pm 4.77}$*	74.24 ± 7.54
		Inverse	62 ± 4.42	62.29 ± 4.69	80.91 ± 4.05	76.47 ± 7.92	90.01 ± 6.07	72.93 ± 8.79
		Square inverse	65.4 ± 7	65.89 ± 6.69	81.82 ± 3.78	76.29 ± 5.98	93.27 ± 7.34	71.11 ± 6.77
	10	Equal	58 ± 5.66	60.18 ± 3.65	78.18 ± 10.67	70.78 ± 14.78	95.31 ± 6.99	65.18 ± 15.63
		Inverse	65.8 ± 6.7	66.25 ± 6.56	85.15 ± 3.63	81.83 ± 5.84	91.66 ± 7.05	77.81 ± 7.35
		Square inverse	63.4 ± 8.64	64.13 ± 8.32	82.73 ± 6.4	77.22 ± 10.05	94.09 ± 5.68	72.12 ± 12

[1]Means and standard deviation. (*p-$value \leq 0.05$)

accuracy, precision, sensitivity and specificity percentages obtained by each classifier and its parameters.

The highest percentage of precision for the evaluation of three classes (BC, BBP, H) was obtained with KNN using values $k = 5$, SC = Euclidean, D=square and using values $k = 3$ SC = Correlation, D = square inverse; obtaining 66.4% in both cases. For the metric of precision the highest percentage obtained was with the parameters of $k = 3$, SC = Correlation and D = square inverse, in the case of the three classes the differences in the percentages were not statistically significant. For the representation of the time series two classes (BC, H), the highest percentages of accuracy (86.36%), precision (87.31) and specificity (85.56%) was obtained with KNN with values of $k = 3$, SC = Euclidean and D = equal. Finally, the highest sensitivity percentage was 95.86% obtained with $k = 5$, SC=correlation and D = equal. However, in the case of the two classes analysis the highest percentages of accuracy, sensitivity and specificity with these classifiers were statistically significant (Table 2).

3.3. *Evaluation of the semi-automatic analysis method, binary and continuous representation comparison*

As can be seen in Tables 1 and 2 the accuracy, precision, sensibility and specificity percentages of classification obtained by the two representations or methods were very similar in both cases. As expected, there was no statistically significant difference (3 classes; $p - value = 0.7658$ and 2 classes (BC and H; p-$value = 0.8338$).

4. Discussion

Breast cancer is one of the leading causes of death in adult women worldwide and the best way to reduce its mortality and improve prognosis is through early diagnosis. Recently our group found that through the analysis of unidimensional Immunoblot Images (1D-II), it was possible to distinguish between women with and without breast disease using as a discrimination criterion the presence of autoantibodies (bands) in their blood. However, the analysis of 1D-II is a difficult task even for an expert, generating great subjectivity and complexity in the process of interpretation.

Although some commercial software to analyze unidimensional Western blot images do exist, most of these tools are designed to analyze the identity of bands in a single image. Thus, the researcher has to align the bands from multiple strips in order to decide if its identity is equal or not between them and with a control strip. In addition, since one image includes about fifteen or seventeen strips, the bands' analysis should consider the comparison of the bands' identity between images. Thus the analysis with traditional software requires a very expert and trained eye and the analysis of a single experiment (an image) can take up to a month.

Taken this complex task, we developed a semi-automatic tool to optimize the procedure and make it less subjective. In this semi-automatic method of analysis, the

1D-I digital image is processed and classified by its banding patterns using a data mining time series approach. We explored the performance of the semi-automatic methodology in a sample of 150 female patients categorized into three classes: BC, BBP and women without pathology (H); and compared its accuracy, specificity, sensibility and precision with a traditional 1D-I image software running by an expert human eye.

Our results show that all classifiers in both methods increase their accuracy when only the BC and H classes were considered. So it seems that the BBP group, being a group with characteristics of both health and illness (in-between group),[21,26,36] generates errors in the classification, thus it is better only to consider H and BC women in the analysis procedure.

In addition, there was no statistically significant difference between the traditional 1D-I Image Software running by an expert human eye and the semi-automatic method of analysis here proposed; that is, it can be concluded that the two methods are equivalent, however the continuous representation by times series is semi-automatic and avoids the subjectivity and complexity of the manual binarization process.

Thus the image analysis method using time series data mining proposed here, not only reduces the subjectivity of the human eye by semi-automating the process, but also facilitates the comparison of multiple images at the same time, which reduces time and effort for the user. That is, the analysis running by commercial software can take even a month for one image, which implies several hours for the researcher in front of a computer; while the semi-automatic method could take 15 min without any effort from the user and without the subjectivity of the human eye. The latter is of great importance since there are studies that show that the human eye can change its perception with the time of exposition to an image.[37,38]

In addition to the reduction of time and subjectivity, the proposed method can be used without previous biological, medical or informatics knowledge, background or technical training in western blot or programing, since the tasks that are performed with the semi-automatic method are simple, such as selecting the area of 1D-II and each of the strips representing the time series to subsequently form the database.

Even when this protocol was implemented in breast cancer western blot images', since Western blot is widely used by many medical and biological scientists in order to separate and identify proteins,[6,22] the semi-automatic method can be applied to any 1D-Western blot image, providing an easy tool to reduce time and subjectivity. Particularly, when the purpose of the Western blot experiment is to find banding patterns and discern this complex patterns when a comparison between strips and groups is made. Therefore, the method of time series analysis proposed here could very well be useful for research in different fields. However, it is still necessary to improve the method and make it automatic through the development of software in which any western blot image can be analyzed in an easier and faster way.

5. Conclusions and Future Work

The present work presents a semi-automatic method for unidimensional immunoblot images to discriminate breast cancer cases using time series data mining. Our results suggest that it is possible to reach similar performance than those obtained by human experts using manual methods. Our method significantly reduces not only the time processing from days to minutes, but also the subjectivity involved in the manual process. Additionally this method could be generalized to analyze any set of Western blot 1D-II patterns.

Although this method improved the traditional process for classification and identification of banding patterns, more resources could be used to extend the work carried. Particularly, we plan to continue working on the following points:

As part of the time series extraction process, due to the difference in size of the ROIs on each 1D-II, it was necessary to manually resize the time series in order to standardize their length. This process can be improved using DTW as an automatic method to realign the 1D-II sequences. Although the DTW algorithm is proposed in time series data mining as a similarity measure method, it can be used as well for registration, because it searches for the best match between two temporal sequences. This algorithm has been successfully applied in different time series data mining applications.[56,57]

Finally, among the classification algorithms used to evaluate the accuracy of the proposed method, in the scope of the probabilistic approaches, only the Naïve Bayes algorithm was included. The advantage of this algorithm is that it is very simple, fast and efficient in general terms, this is because it assumes conditional independence among attributes; however, in practice this is not always true. A better alternative is to use Bayesian networks (BN), under this approach all the conditional probabilities between attributes are tested in order find relationships.[58,59] These relationships are shown in a graphical model in which the attributes represent the nodes and the arches represent the conditional dependences between them. A disadvantage of this method is that it needs many more training examples than Naïve–Bayes in order to create a solid model.

Acknowledgments

Authors would like to thank Reynaldo Domínguez Castillo for his support in statistical analysis. Diana M. Sánchez-Silva was recipient of a master fellowship from the Consejo Nacional de Ciencia y Tecnología (CONACYT). The manuscript was proofread by Sara Robledo Waters.

References

1. World Health Organization (WHO), Breast cancer: prevention and control, WHO 2006. Available at: http://www.who.int/cancer/detection/breastcancer/en/ (accessed on February 11, 2016).

2. Instituto Nacional de Estadística y Geografía (INEGI). Available at: http://www.inegi.org.mx [accessed on April 7, 2016].
3. ISSEMYM —Instituto de Seguridad Social del Estado de México y Municipios, Cáncer de mama. ISSEMYM, México, Available at: http://www.issemym.gob.mx/index.php?page=cancer-de-mama.
4. E. Azuara, L. E. Álvarez and P. Gariglio, SALUD DE LAS MUJERES. Cáncer, Biología Molecular Genómia y Proteómica. Instituto de Ciencia y Tecnología del DF. Universidad Autónoma de la Ciudad de México. 1st edn., 2010.
5. E. P. Nagele, M. Han, N. K. Acharya, C. DeMarshall, M. C. Kosciuk and R. G. Nagele, Natural IgG autoantibodies are abundant and ubiquitous in human sera, and their number is influenced by age, gender, and disease, *Plos One* **8**(4) (2013), doi: https://doi.org/10.1371/journal.pone.0060726.
6. A. Aggarwal, Role of autoantibody testing, *Best Pract. Res. Clin. Rheumato.,* doi: http://dx.doi.org/10.1016/j.berh.2015.04.010.
7. C. DeMarshall, A. Sarkar, E. P. Nagele, E. Goldwaser, G. Godsey and N. K. Acharya, Utility of autoantibodies as biomarkers for diagnosis and staging of neurodegenerative diseases, *Int. Rev. Neurobiol.* **122** (2015), doi: http://dx.doi.org/10.1016/bs.irn.2015.05.005.
8. S. Gnjatic, C. Wheeler, M. Ebner, E. Ritter, A. Murray, N. K. Altorki *et al.*, Seromic analysis of antibody responses in non-small cell lung cancer patients and healthy donors using conformational protein arrays, *J. Immunol. Methods* **341**(1–2) (2009) 50–58. Availabel at: http://dx.doi.org/10.1016/j.jim.2008.10.016.
9. E. M. Tan and J. Zhang, Autoantibodies to tumor-associated antigens: Reporters from the immune system, *Immunol. Rev.* **222** (2008) 328–340. Available at: http://dx.doi.org/10.1111/j.1600-065X.2008.00611.x.
10. M. Nesterova, N. Johnson, C. Cheadle and Y. S. Cho-Chung, Autoantibody bio- marker opens a new gateway for cancer diagnosis, *Biochimica et Biophysica Acta* **1762**(4) (2006) 398–403. Available at: http://dx.doi.org/10.1016/j.bbadis.2005.12.010.
11. C. J. Chapman, A. J. Thorpe, A. Murray, C. B. Parsy-Kowalska, J. Allen, K. M. Stafford *et al.*, Immunobiomarkers in small cell lung cancer: Potential early cancer signals, *Clin. Cancer Res.* **17**(6) (2011) 1474–1480. Available at: http://dx.doi.org/10.1158/1078-0432.CCR-10-1363.
12. I. Diesinger, C. Bauer, N. Brass, H. J. Schaefers, N. Comtesse, G. Sybrecht *et al.*, Toward a more complete recognition of immunoreactive antigens in squamous cell lung carcinoma, *Int. J. Cancer* **102**(4) (2002) 372–378. Available at: http://dx.doi.org/10.1002/ijc.10714.
13. M. J. Scanlan, Y. T. Chen, B. Williamson, A. O. Gure, E. Stockert, J. D. Gordan *et al.*, Characterization of human colon cancer antigens recognized by autologous anti-bodies, *Int. J. Cancer* **76**(5) (1998) 652–658.
14. M. L. Disis, E. Calenoff, G. McLaughlin, A. E. Murphy, W. Chen, B. Groner *et al.*, Existent T-cell and antibody immunity to HER-2/neu protein in patients with breast cancer, *Cancer Res.* **54**(1), 1994 16–20.
15. X. Wang, J. Yu, A. Sreekumar, S. Varambally, R. Shen, D. Giacherio *et al.*, Autoantibody signatures in prostate cancer, *New England J. Med.* **353**(12) (2005) 1224–1235. Available at: http://dx.doi.org/10.1056/NEJMoa051931.
16. M. Chatterjee, S. Mohapatra, A. Ionan, G. Bawa, R. Ali-Fehmi, X. Wang *et al.*, Diagnostic markers of ovarian cancer by high-throughput antigen cloning and detection on arrays, *Cancer Res.* **66**(2) (2006) 1181–1190. Available at http://dx.doi.org/10.1158/0008-5472. CAN-04-2962.
17. T. E. Carey, K. A. Kimmel, D. R. Schwartz, D. E. Richter, S. R. Baker and C. J. Krause, Antibodies to human squamous cell carcinoma, *Otolaryngo. Head Neck Surg.* **91**(5), (1983) 482–491.

18. E. M. Smith, L. M. Rubenstein, J. M. Ritchie, J. H. Lee, T. H. Haugen, E. Hamsikova *et al.*, Does pretreatment seropositivity to human papillomavirus have prognostic significance for head and neck cancers? *Cancer Epidemiol. Biomarkers Prevention* **17**(8) (2008) 2087–2096. Available at: http://dx.doi.org/10.1158/1055-9965.EPI-08-0054.
19. C. Xie, H. J. Kim, J. G. Haw, A. Kalbasi, B. K. Gardner, G. Li *et al.*, A novel multiplex assay combining autoantibodies plus PSA has potential implications for classification of prostate cancer from non-malignant cases, *J. Translational Med.* **9** (2011) 43. Available at: http://dx.doi.org/10.1186/1479-5876-9-43.
20. C. Chapman, A. Murray, J. Chakrabarti, A. Thorpe, C. Woolston, U. Sahin *et al.*, Autoantibodies in breast cancer: Their use as an aid to early diagnosis, *Ann. Oncol.* **18**(5) (2007) 868–873. Available at: http://dx.doi.org/10.1093/annonc/mdm007.
21. M. Esquivel-Velázquez, T. Romo-González, R. León-Díaz, J. Pérez-H, A. Zentella, R. Pérez-Tamayo *et al.*, Autoantibodies increase in number while their inter-individual variation decreases in women diagnosed with Breast Cancer: The first steps towards developing a Prognostic/Diagnostic tool. In preparation.
22. P. C. Yang and T. Mahmood, Western blot: Technique, theory, and trouble shooting, *North Am. J. Med. Sci.* **4**(9) (2012).
23. Y. Yan and M. Hongbao, *Western Blot. ELISA Tech.* **1**(2) (2009) 67–86.
24. B. T. Kurien and R. H. Scofield, *Western Blott. Methods* 38(4) (2006) 283–293.
25. M. Gassmann, B. Grenacher, B. Rohde and J. Vogel, Quantifying Western blots: Pitfalls of densitometry, *Electrophoresis* **30** (2009) 1845–1855, doi: 10.1002/elps.20 0800720.
26. T. Romo-González, M. Esquivel-Velázquez, P. Ostoa-Saloma, C. Lara, A. Zentella, R. León-Díaz *et al.*, The network of antigen-antibody reactions in adult women with breast cancer or benign breast pathology or without breast pathology. Plos One. (2015). Available at: https://doi.org/10.1371/journal.pone.0119014 [Accessed on March 1, 2016].
27. Bio-Rad Laboratories, Quantity One 1-D Analysis Software 2015. Available at: http:// www.bio-rad.com/en-mx/product/quantity-one-1-d-analysis-software.
28. H. G. Acosta Mesa, R. F. Ramirez, E. Mezura Montez, N. Cruz Ramírez, R. Hernandez Jiménez, Application of time series discretization using evolutionary programming for classification of precancerous cervical lesions., *J. Biomed. Informat.* (2014). Volume 49, 73–83, doi: http://dx.doi.org/10.1016/j.jbi.2014.03.004 (accessed on February 23, 2016).
29. M. Last, A. Kandel and H. Bunke, *Data Mining in Time Series Databases* (World Scientific Press, Singapore, 2004).
30. P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining* (2005) Addison-Wesley Publishing Co, Boston, MA, USA.
31. E. Refaeilzadeh, L. Tang and H. Liu, *Cross-Validation.* Arizona State University. Available at: http://leitang.net/papers/ency-cross-validation.pdf
32. T. M. Mitchell Tom, *Machine Learning* (McGraw-Hill, 1997).
33. J. Han and M. Kamber, *Data Mining: Concepts and Techniques* (Morgan Kaufmann Publishers, 2001).
34. G. Batista and D. Furtado Silva, How K-Nearest Neighbor Parameters Affect its Performance. 28° JAIIO-Simposio Argentino de Inteligencia Artificial, 2009.
35. MATLAB. Matriz Laboratory. 1994–2016. The MathWorks, Inc. Available at: http:// www.mathworks.com [accessed on February 23, 2016].
36. I. Montes-Nogueira, Y. Campos-Uscanga, G. Gutiérrez-Ospina, C. Larralde and T. Romo-González, Psychological Features and Breast Cancer in Mexican Women II: The Psychological Network. Advances in Neuroimmune Biology. Submitted.
37. D. H. Hubel, *Eye, Brain, and Vision* (W.H. Freeman, 1995).

38. D. Chabot and M. Chabot, Emotional Pedagogy, To feel in order to learn Incorporating Emotional Intelligence in your teaching strategies, Trafford, 2004.
39. WEKA, *Waikato Environment for Knowledge Analysis* (University of Waikato, New Zealand). Available at: http://www.cs.waikato.ac.nz/ml/weka/.
40. A. Herrera Gomez and M. Grandos Garcia, *Manual de Oncología. Procedimientos médico quirúrgicos* (McGraw Hill, Quinta Edición, México, 2015).
41. IARC (2016). CANCER Mondial. International Agency for Research on Cancer. Available at: http://www-dep.iarc.fr/ [accessed on April 7, 2016].
42. Y. Chávarri-Guerra, C. Villarreal-Garza, P. E. R. Liedke, F. Knaul, A. Mohar, D. M. Finkelstein and P. E. Goss, Breast cancer in Mexico: A growing challenge to health and the health system, *Lancet Oncol.* **13**(8) (2012) e335–43, doi: 10.1016/S1470-2045(12) 70246-2.
43. T. J. Kindt and A. Goldsby Richard, *Inmunología de Kuby* (McGraw Hill, Sexta edición, 2007).
44. C. Desmetz, J. Lacombe, A. Mange, T. Maudelonde and J. Solassol, Autoanticorps et diagnostic précoce des cancers. Médicine/sciences 2011.
45. G. L. Kumar and L. Rudbeck, *Immunohistochemical Staining Methods Pathology* (Education Guide, Dako North America, 2009).
46. V. Ameyalli, Desarrollan nanobiosensor para detectar cáncer de mama con saliva. México, DF. 28 de Junio 2015. Available at: http://www.conacytprensa.mx/index.php/tecnologia/nanotecnologia/2027-nanobiosensor-para-detectar-cancer-de-mama?utm_source=newsletter_816&utm_medium=email&utm_campaign=conacyt-newsletter-31-2015.
47. Collaborative Group on Hormonal Factors in Breast Cancer. Breast Cancer and Hormonal Contraceptives: Collaborative reanalysis of individual data on 53297 women with breast cancer and 100239 women without breast cancer from 54 epidemiological studies, *Lancet* **347** (1996).
48. INEGI — Instituto Nacional de Estadística y Geografía ESTADÍSTICAS A PROPÓSITO DEL DÍA INTERNACIONAL CONTRA EL CÁNCER DE MAMA. INEGI, México, Available at: http://www.inegi.org.mx/inegi/contenidos/espanol/prensa/Contenidos/estadisticas/2013/mama0.pdf.
49. Tak-Chung Fu, A review on time series data mining, *Eng. App. Artif. Intell.* (2010), doi: 10.1016/j.engappai.2010.09.007.
50. P. Esling and C. Agon, Time-Series data mining. *ACM Comput. Surv.* **45**(1) Article 12, 2012, doi: http://doi.acm.org/10.1145/2379776.2379788.
51. A. Roche, Árboles de decisión y Series de Tiempo. Tesis de Maestría en Ingeniería Matemática, UDELAR. 2009.
52. E. Morales Manzanares, *Machine Learning. Instituto Nacional de Astrofísica* (Óptica y Electrónica (INAOE), 2005).
53. C. González Rafael and E. Woods Richard, *Digital Image Processing*, 2nd edn., University of Tennessee, MedData Interactive (Prentice Hall, 2002).
54. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining* (Wiley-Interscience, 2004).
55. M. Kantardzic, *Data Mining: Concepts, Models, Methods and Algorithms* (Wiley-Interscience, 2003).
56. H. Li, On-line and dynamic time warping for time series data mining, *Int. J. Mach. Learn. Cyber.* **6** (2015) 145, doi: 10.1007/s13042-014-0254-0.
57. P. Singh, Rainfall and financial forecasting using fuzzy time series and neural networks based model, *Int. J. Mach. Learn. Cyber.* (2016), doi: 10.1007/s13042-016-0548-5.

58. M. Van der Heijden, M. Velikova and P. J. F. Lucas, Learning Bayesian networks for clinical time series analysis, *J. Biomed. Inform.* **48** (2014) 94–105, https://doi.org/10.1016/j.jbi.2013.12.007.
59. E. Philippot, K. C. Santosh, A. Belaïd *et al.*, Bayesian networks for incomplete data analysis in form processing, *Int. J. Mach. Learn. Cyber.* **6** (2015) 347, doi: 10.1007/s13042-014-0234-4

Diana M. Sánchez-Silva received her Master's degree in Artificial Intelligence from Universidad Veracruzana, México. She is currently a research assistant at the Instituto de Investigaciones Biológicas, Universidad Veracruzana. Her areas of research are psychoneuroimmunology and wellness, lifestyles and health (diagnosis, prevention and treatment).

Tania Romo-González received her Ph.D. in Biomedical Sciences (Universidad Nacional Autónoma de México (UNAM), México). She is currently a researcher at the Instituto de Investigaciones Biológicas, Universidad Veracruzana. Her areas of research are psychoneuroimmunology and wellness, lifestyles and health (diagnosis, prevention and treatment).

Héctor G. Acosta-Mesa received his Ph.D. in Artificial Intelligence from the University of Sheffield, United Kingdom. He is currently working in the Artificial Intelligence Vision Research Unit, Universidad Veracruzana. His area of research is on applications of computer vision algorithms in medical imaging.