Automatic classification of acetowhite temporal patterns to identify precursor lesions of cervical cancer

K Gutiérrez-Fragoso[1], H G Acosta-Mesa[2], N Cruz-Ramírez[2] and R Hernández-Jiménez[3]

[1] Biomedical Research Center, Universidad Veracruzana. México
[2] School of Physics and Artificial Intelligence, Department of Artificial Intelligence, Universidad Veracruzana. México
[3] Obstetrician and Gynaecologist, Private Practice. México

E-mail: kagutierrez@uv.mx, heacosta@uv.mx, ncruz@uv.mx, roheji@msn.com

Abstract. Cervical cancer has remained, until now, as a serious public health problem in developing countries. The most common method of screening is the Pap test or cytology. When abnormalities are reported in the result, the patient is referred to a dysplasia clinic for colposcopy. During this test, a solution of acetic acid is applied, which produces a color change in the tissue and is known as acetowhitening phenomenon. This reaction aims to obtaining a sample of tissue and its histological analysis let to establish a final diagnosis. During the colposcopy test, digital images can be acquired to analyze the behavior of the acetowhitening reaction from a temporal approach. In this way, we try to identify precursor lesions of cervical cancer through a process of automatic classification of acetowhite temporal patterns. In this paper, we present the performance analysis of three classification methods: kNN, Naïve Bayes and C4.5. The results showed that there is similarity between some acetowhite temporal patterns of normal and abnormal tissues. Therefore we conclude that it is not sufficient to only consider the temporal dynamic of the acetowhitening reaction to establish a diagnosis by an automatic method. Information from cytologic, colposcopic and histopathologic disciplines should be integrated as well.

1. Introduction

The worldwide estimates of incidence, mortality and prevalence of cancer in 2008 presented by the World Health Organization (WHO) reported that cervical cancer remains the second leading cause of death in women. Early diagnosis of precursor lesions of this cancer has set itself as a challenge. The most common screening test is the Pap smear; if the presence of abnormalities are reported in the result, the next step is a colposcopy test [15]. The colposcopy allows us to visualize the uterine cervix using a microscope fitted with a light source. Colposcopic appearance of a tissue is constituted by three main factors: the appearance of the epithelium, the composition of the underlying stroma and the configuration of the surface [25]. During the colposcopy test, the appearance of the cervix is observed and a solution of 3% acetic acid is spread on the epithelium, which produces a change from the usual pink tissue to a whitish color. This reaction is called acetowhitening [13]. The effect of the acetic acid is more evident and disappears more slowly in high-grade lesions (HSIL) and invasive cancer in early stages compared with low-grade lesions (LSIL) and subclinical changes associated with human papillomavirus (HPV). However, acetowhite appearance is not exclusive to high-grade lesions and cancer in early stages; it is also observed in other conditions such as immature squamous metaplasia, congenital transformation zone, epithelial regeneration and healing process (associated with inflammatory processes), leukoplakia (hyperkeratosis) and condyloma [22].

In order to establish the diagnosis of cervical cancer, presence of malignant cells in cytology, appearance of malignant images and histopathological confirmation are required [19]. That is, the final diagnosis depends on the visual sampling of the cervix to obtain a representative biopsy. This procedure is based on the appearance of the tissue and the reaction produced by acetowhitening. Nevertheless, there are discrepancies on how each specialist proceeds and this is associated with the subjectivity of different observers. Because of this issue, some methods for automatically classifying precursor lesions of cervical cancer using colposcopic images have been developed to try to establish an appropriate diagnosis. However, the work done so far by means of a time series approach do not have an integral methodology to acquire, process and analyze the images. In general, they have a reduced amount of cases [6, 7, 20, 23, 24].

In this work, we present the performance analysis of three automatic classification methods using time series to describe the dynamics of the acetowhitening reaction for different types of cervical tissue. The methods used are: kNN, Naïve Bayes, and C4.5. The temporal representation of the acetowhitening phenomenon was obtained from a sequence of digital images acquired during the colposcopy test. This paper is structured in three sections. In the first section, the subject preparation, the data acquisition process, the preprocessing techniques and the automatic classification methods are described in detail. In the second section, the results and a discussion of the findings are explained. Finally, in the third section, the conclusions and some proposals for future analysis are presented.

2. Materials and Methods

2.1. Subjects preparation

In this study, 200 women were included. Within the total quota of patients, in 100 cases a tissue sample or biopsy was obtained because some changes in the appearance of the cervical epithelium were observed by the colposcopist and these alterations led to suspicion of a lesion. In the other 100 patients, the specialist did not find changes that suggest the presence of a lesion and it was not necessary to obtain a biopsy. Of the total quota, 93 cases were positive for precursor lesions of cervical cancer and 107 negative. All of the patients were referred to the Medical Specialties Center of the State of Veracruz (CEMEV) "Dr. Rafael Lucio" because in the analysis of Pap smear, test abnormalities were reported. The age average was 34 years (SD = 9), 88% claimed not to smoke, 54% reported having one sexual partner, the age average of first sexual intercourse was 18 years (SD = 4), 40% had bilateral tubal obstruction (OTB) as a method of family planning and only 2% used oral hormonal contraceptives. The patients signed an informed consent. Subsequently, the colposcopic test was performed.

2.2. Colposcopy

The technique of colposcopy began by explaining the procedure, the patient was placed in the gynecological position on the examination table. After this, a speculum was placed and the cervical mucus was cleaned by cotton swabs impregnated with saline solution. The appearance of cervical tissue was observed and approximately three milliliters of 3% acetic acid solution were spread in the cervical area. A cotton swab was placed in the lower part of the cervix to absorb the excess solution. In cases where a biopsy was obtained, solution of Monsel was used to coagulate the bleeding at the site where the tissue sample was taken.

2.3. Data acquisition

A set of digital images was obtained during colposcopic test. The acquisition was performed using a colposcope Vasconcellos CP-M1225 with a camera STC-N63BJ. The green filter was used to acquire the images because a previous study had shown that in this wavelength range the whitened areas were better highlighted [7]. The dimension of the images was 352 x 240 pixels with a sampling frequency of 1 frame/second. The images were stored as separate files in BMP format. A tool was developed for the acquisition and it was implemented in MATLAB 7.0. Before the application of acetic acid, 10 images were obtained as a reference to calculate the percentage of color change of the tissue. Then, 180 images were acquired during a period of 3 minutes. When the acquisition process was completed, the region where the biopsy was obtained was selected by the colposcopist on one of the images previously acquired. In the cases where it was not necessary to obtain a biopsy, a representative region of the kind of tissue was selected in the image. The computer processing of the images was done in grayscale.

2.4. Preprocessing

During the image acquisition process, slight movements occur commonly, which are attributed to nervousness, muscle tone and breathing of the patient. Therefore, a technique was used to align the sequence of images and achieve anatomical correspondence among them. This process is called registration and it is essential for the analysis. There are various registration methods but, basically, they can be based directly on the image intensity values (area-based methods) or can be accomplished using some features calculated from the images (feature-based methods). Because colposcopic images do not contain many differences through the sequence, an area-based method was chosen. The classical method in this category is the normalized cross-correlation (CC). This method takes advantage of the image intensities and this similarity metric is calculated by pairs of windows from the input and reference images until its maximum is reached [27]. The input and the reference images are updated starting with the first and second images of the sequence respectively, then the input and the reference images are redefined by the second and the third images and so on.

2.5. Time series extraction

The set of sequential images acquired at time t and represented in two dimensions (x,y) with an intensity value I according to the levels in the gray scale, allows us to construct a time series based on the intensity value of each pixel over the time (Acetowhite response function, *Awrf*). In order to compare the time series taken from different subjects, a standardization method was applied, which calculates the percentage of change of the signal with respect to the reference period. The values of the time series are divided by the mean value of its basal period [2]. In this study, three representations of the time series were considered: standardized data, data adjusted to a polynomial model and polynomial model parameters. The first representation refers to the standardized data by calculating the percentage of change. The last two correspond to the following polynomial model:

$$Awrf = \theta_0 + \theta_1 t + \frac{\theta_2}{t} + \frac{\theta_3}{t^2} + \frac{\theta_4}{t^3} \tag{1}$$

where:
$Awrf$ = acetowhite response function
θ = explanatory variables
t = time series

The polynomial model was obtained experimentally by analyzing the behavior of the time series [2]. The parameters θ were also used as a compact representation of the time

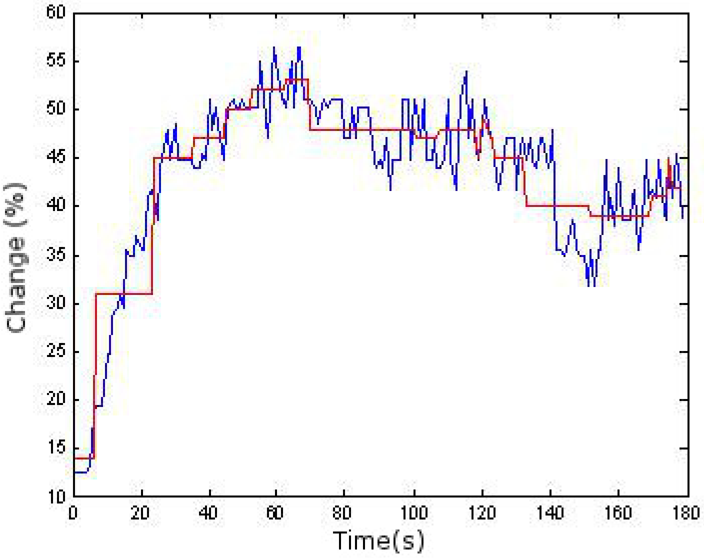

Figure 1. Example of representation PLA

series. Because time series databases are often extremely large, some methods to reduce the dimensionality have been developed. There are methods to produce approximations of time series through discretization schemes. The discretization process is focused on mapping variables with continuous values into discrete values. This mechanism has been widely used to compress data and facilitate their computational treatment. The reduction of dimensionality in the x-axis is obtained by dividing the total length of the time series into fragments of a certain size (*word size*). It is also necessary to establish a number of intervals in the y-axis to compress the values of the time series (*alphabet size*) [12]. In this work, the discretization scheme was obtained by the method developed in [21], in which the word size and alphabet is optimized as a single parameter. The approach used by the authors was evolutionary programming. The discretization scheme obtained allowed us to establish the size of the segments in order to divide the length of the time series. Two criteria to discretize the time series were used:

- Piecewise Linear Approximation (PLA). This method computes the average of the values in each segment on the x-axis. Then the average value is mapped to a discrete value, searching the interval in y-axis in which it is included [5]. The segment size (*word size*) and the intervals (*alphabet size*) were established according to the discretization scheme.
- Piecewise Slope Approximation (PSA). This algorithm is similar to the previous one but in this case, the slope is calculated for each segment and this value is mapped among 7 possible values: 3 negative values, 3 positive values and the number 0 represents no change.

We decided to use these discretization methods to be able to compare our results with those obtained previously [4]. Although both of the studies utilized the discrete representations mentioned above (PLA and PSA), in [4] the word and alphabet parameters were assigned directly considering all segments of the same size whereas in this work the length of these parameters was variable and it was established by the discretization scheme obtained through the method developed in [21].

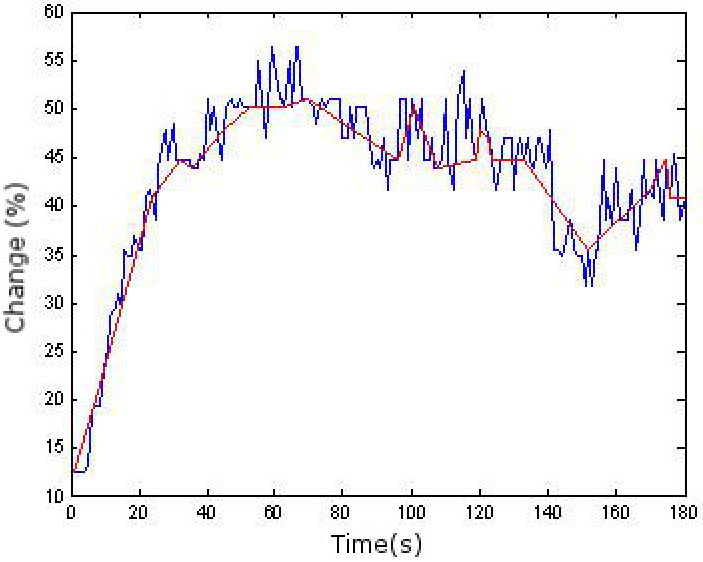

Figure 2. Example of representation PSA

2.6. Databases

It was mentioned in *section 2.3* that after the acquisition of the images, the colposcopist selected a region where a biopsy was obtained or a representative region of the tissue type. Evaluating the automatic classification methods, 5 time series databases were incorporated, they were constructed as follows: one database included a time series of a pixel within the region from which a biopsy was obtained. The remaining databases were constructed using the average of the time series of windows of different sizes within the region. In this way, we obtained the average time series from windows of 7x7, 11x11, 15x15 and 19x19 pixels. Every database included one time series obtained from the colposcopic image sequence of each patient.

2.7. Supervised Learning

Supervised learning is an area of machine learning in which models are constructed from a set of observations presented as examples. The goal of these methods is to predict the label of the class of a new observation given the model constructed from a set of training data, which can be represented by a database containing observations and corresponding labels of the classes. When the label of the class is categorical, learning is called classification [11]. There are different models for the classification process such as k-nearest neigbors (kNN), Naïve Bayes (NB) and decision trees (ID3, C4.5 and 4.8) among others [14]. We used the classification methods kNN, NB and C4.5 in this work and each is described below.

2.7.1. k Nearest Neighbors (kNN). In this method, given an unclassified example the class is assigned based on similar observations with the new example from a training dataset. When the value of k is equal to 1, the class of the more similar observation of the training set is assigned; otherwise, the new label example takes the most frequent class of the k observations. The observations in the dataset are the time series extracted from colposcopic images. There are different ways to measure the similarity but the most common is the Euclidean distance [11].

2.7.2. Naïve Bayes (NB). This method can predict the probability that a new example belongs to a class given the observations (time series) contained in the database. This method is based on Bayes' theorem, which is useful because it provides a way to calculate these probabilities.

$$p(c_j|d) = \frac{p(d|c_j)p(c_j)}{p(d)} \tag{2}$$

where:
$p(c_j|d)$ = probability of class c_j given the observation d
$p(d|c_j)$ = probability of the observation d given class c_j
$p(c_j)$ = probability of occurrence of class c_j
$p(d)$ = probability of occurrence of the observation d

2.7.3. C4.5. The induction of a decision tree is learned from a set of training observations with known class labels. The decision tree is a structure resembling a flow chart just as a tree where each internal node denotes a test of an attribute. In this case each discrete value of the time series was considered as an attribute. The decision tree has branches and each branch represents a departure from the test and each terminal node ("leaf node") has a class label. The node at the top of the tree is the root node [10].

The general strategy to build a decision tree is:

- The algorithm takes as input a data partition D. Initially, this partition consists of the full set of training observations (time series) with their associated class labels. A list of attributes (discrete values of the time series) that describe the observations and a selection method, which specifies a procedure for selecting the attribute that "best" discriminates the observations according to the class, are also considered.
- The tree starts with a single node N that represents the training observations in D.
- If the observations in D are all of the same class, the node N becomes a leaf node and is labeled with that class.
- Otherwise, the algorithm uses the attribute selection method to determine the criterion of discrimination or division of the observations. This criterion indicates which attribute evaluated at node N provides the "best" way to separate the observations of D on individual classes. The selection criterion also indicates which branch should grow from the node N.
- For each of the outputs of the attribute evaluated by the selection criterion grows a branch from node N and the observations in D are divided in each branch according to different attribute values [10].

It should be mentioned that for the method kNN the time series were processed with continuous values but the methods NB and C4.5 used the discretized representations PLA and PSA.

2.8. Statistical methods

There are some methods used in order to assess and compare the learning algorithms. The cross-validation method is one of them and consists of dividing the data into two sets: training and validation. In the basic form of cross-validation, the database is divided into k-partitions (k-fold cross validation, k-fold CV) of equal or almost equal size. The procedure consists in making k iterations of training and validation executed consecutively. Thus, within each iteration, a different partition of data is maintained for validation while the remaining k-1 partitions are used for learning. Additionally a stratification procedure can be applied, which is done to reorganize the data in order to ensure that each partition is a good representation of the proportion of classes in the full data set [18]. The leave-one-out method (LOO) is a special case of cross-validation where the training set is composed with the data except for one observation,

Table 1. Confusion matrix

Actual \Predicted	Class = 1	Class = 0
Class = 1	f_{11}	f_{10}
Class = 0	f_{01}	f_{00}

which is used to validate the model. The procedure is applied as many times as examples the set of data has [18].

Furthermore, the evaluation of the performance of a classification model is based on the number of test cases correctly and incorrectly predicted by the model. These calculations are tabulated in a table called confusion matrix (1). In this work, f_{01} is the number of cases of class 0 (negative) incorrectly predicted as class 1 (positive). According to the entries in the confusion matrix, the total number of correct predictions made by the model is $(f_{11} + f_{00})$ and the total number of incorrect predictions is $(f_{10} + f_{01})$. This information can be summarized in a metric to compare the performance of different methods such as accuracy [26], which is defined as follows:

$$accuracy = \frac{\text{number of correct predictions}}{\text{total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \tag{3}$$

Sensitivity is also a metric to assess the performance of a classification model, it refers to the number of positive cases (precursor lesions of cervical cancer) actually identified as positive by a test. On the other hand, specificity is the number of negative cases (no cervical cancer precursor lesions) identified as negative by the test. The metrics of accuracy, sensitivity and specificity are expressed in percentages.

$$sensitivity = \frac{f_{11} * 100}{f_{11} + f_{10}} \tag{4}$$

$$specificity = \frac{f_{00} * 100}{f_{00} + f_{01}} \tag{5}$$

3. Results and Discussion

The results of kNN using the leave-one-out method (LOO) were obtained with a tool implemented in Matlab. The assessment of kNN, NB, and C4.5 using k-fold CV was obtained using Weka software module explorer 3.7.1. Table 2 summarizes the best results obtained from different automatic classification methods based on the acetowhite temporal patterns ($Awtp$) under different representations. The metrics of sensitivity, specificity and accuracy were evaluated. The results obtained in [3, 4] and those reported in [16] were included for comparative purposes.

In the kNN method, the table 2 shows that higher accuracy values correspond precisely to those obtained in this work using LOO with both: the normalized data representation and the data adjusted to polynomial model (70%). This shows an improvement over that reported in [3]. Particularly, there is an increase in the metric of specificity. Comparing the results obtained in this work using NB and C4.5 methods, the highest value of accuracy was achieved with C4.5 and the discretized representation PLA (71%). However, making a general comparison among all of the methods on table 2, it can be observed that the highest accuracy is obtained using the method NB (PLA) [4].

Table 2. Comparative summary of the classification methods and its representations using acetowhite temporal patterns. The name of each method was coded as follows: automatic classification method, statistical analysis method and data representation. The letter k in the header of the table represents the number of neighbors for kNN method.

	Methods	Window	k	Sensitivity(%)	Specificity(%)	Accuracy(%)
a	kNN LOO adjusted [3]	NA	20	71	59	67
b	kNN LOO normalized	1	3	68	71	70
c	kNN LOO adjusted	1	15	61	77	70
d	kNN LOO parameters	1	25	51	66	59
e	kNN k-fold CV normalized	19	10	68	69	69
f	kNN k-fold CV adjusted	1	25	60	75	68
g	kNN k-fold CV parameters	19	5	58	72	66
h	NB PLA [4]	NA	NA	67	85	76
i	NB PSA [4]	NA	NA	61	80	70
j	NB k-fold CV PLA	7	NA	60	76	69
k	NB k-fold CV PSA	15	NA	67	67	67
l	C4.5 k-fold CV PLA	11	NA	65	77	71
m	C4.5 k-fold CV PSA	1	NA	53	73	64
n	Colposcopist [16]	NA	NA	91	57	74

To investigate the behavior of the classification methods in this work, a class-level analysis (these results are not shown in this paper) was carried out and showed the presence of similarity between acetowhite temporal patterns of precursor lesions of cervical cancer and normal cases. This could explain the lower values of accuracy obtained in this study compared with those reported in [3, 4, 6].

Clinically, this finding can be interpreted from two points of view. First, it is necessary to consider that the acetowhite appearance is not exclusive in early stages of cervical cancer, which means that the dynamics of this reaction may be similar between cases considered as negative to be precancerous lesions and actually abnormal cases [22]. Moreover, it is possible that the difficulty faced by specialists to establish a proper diagnosis has been also a problem of classification methods when they are based only on the temporal dynamics of acetowhitening phenomenon.

4. Conclusions and Future work

In this paper, we assessed the performance of three classification methods using a temporal approach. The analysis showed the similarity of acetowhite temporal patterns of precancerous lesions and normal cases. This can be explained by the difficulty of discrimination between some normal and abnormal epitheliums even for colposcopist and apparently this was also a problem for the automatic classification methods. Furthermore, because the acetowhite appearance is not always associated with a precancerous lesion, the acetowhite temporal patterns by themselves apparently do not constitute a sufficient element to discriminate between normal and abnormal tissues by an automatic method. Therefore, it is advisable to consider the characteristics of the cervical epithelium additionally to the acetowhitening reaction.

Subsequent studies might include features such as patterns of blood vessels (punctuation and mosaic), shape of borders, location and size of the lesion. It is also important to integrate cytologic, colposcopic, and histopathologic data to establish a proper diagnosis.

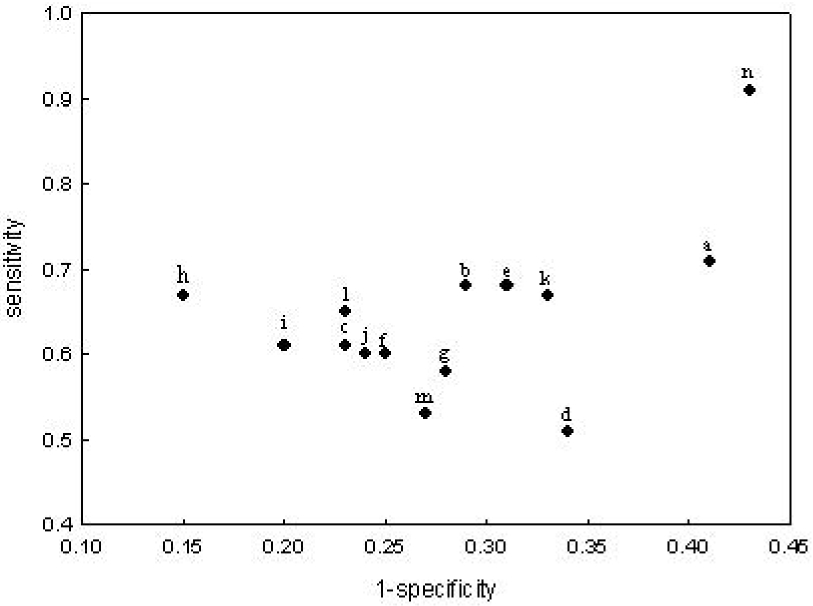

Figure 3. ROC curve

This work also provided a set of colposcopic image sequences. This data will be part of a repository to evaluate different methods of preprocessing and classification. This will allow the continuation of research on automatic classification of cervical cancer precursor lesions.

Acknowledgments

The authors thank the Mexican National Council for Science and Technology for the economic support of this project under the research grant: Fondo Sectorial de Investigación en Salud y Seguridad Social SSA/IMSS/ISSSTE-CONACYT (Salud-2008-01-86566) and the doctoral scholarship (171310). The authors also thank the valuable collaboration of Dr. Antonio Alberto Martínez Virues, Dr. Luis Landeros López, and Dr. Hormisdas Alvarado Carrasco.

References

[1] Acosta-Mesa H G, Cruz-Ramírez N, Llaguno-Roque J L, Hernández-Jiménez R, Cocotle-Ronzón B E 2007 Clasificación de patrones temporales para caracterizar lesiones cervico uterinas en imágenes colposcópicas *Avances en la Ciencia de la Computación* (México: Sociedad Mexicana de Ciencia de la Computación)

[2] Acosta-Mesa H G, Cruz-Ramírez N, Hernández-Jiménez R, García-López D A 2007 Modeling aceto-white temporal patterns to segment colposcopic images *Lecture Notes in Computer Science* (Springer)

[3] Acosta-Mesa H G, Cruz-Ramírez N, Hernández-Jiménez R, and Cocotle-Ronzón B E 2009 Aceto-white temporal patterns classification using k-NN to identify precancerous cervical lesion in colposcopic images *Comput Biol Med* **39** 9 778-784

[4] Acosta-Mesa H G, Cruz-Ramírez N, Gutiérrez-Fragoso K, Barrientos-Martínez R E, Hernández-Jiménez R 2010 Assessing the possibility of identifying precancerous cervical lesions using aceto-white temporal patterns in *Decision Support Systems Advances* Devlin G (Ed) InTech

[5] Allgower E L and Kurt G 1988 Estimates for piecewise linear approximations of implicity defined manifolds *SIAM Journal of Numerical Analysis* **24** 452-469

[6] Balas C J, Themelis G C, Prokopakis E P, Orfanudaki I, Koumantakis E and Helidonis E S 1999 In vivo detection and staging of epithelial dysplasias and malignancies based on the quantitative assesment of acetic acid tissue interaction kinetics *Photochemistry and Photobiology B: Biology* **53** 153-157

[7] Balas C J 2001 A novel optical imaging method for the early detection, quantitative grading, and mapping of cancerous and precancerous lesions of cervix *IEEE Transactions on Biomedical Engineering* **48** 1 96-104

[8] Brown L G 1992 A survey of image registration techniques *ACM Computing Survey* **24** 4 325-376

[9] Conzuelo-Q A E 2002 *Nuevas Alternativas en el Tratamiento de Papilomavirus* (México: Editorial Prado)

[10] Han J and Kamber M 2006 *Data Mining: Concepts and Techniques* (USA: Elsevier)

[11] Hastie T, Tibshirani R, and Friedman J 2001 *The elements of statistical learning (Data mining, Inference and Prediction)* (USA: Springer)

[12] Keogh E and Pazzani M 2000 A simple dimensionality reduction technique for fast similarity search in large time series databases *Proc. of the 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, Current Issues and New Applications* (UK: Springer-Verlag London)

[13] Maddox P, Szarewski A, Dyson J and Cuzick J 2004 Cytokeratin expression and acetowhite change in cervical epithelium *Journal of Clinical Pathology* **47** 1 15-17

[14] Mitchell T M 1997 *Machine Learning* (USA: McGrawHill)

[15] World-Health-Organization(WHO) 2008 The GLOBOCAN project http://globocan.iarc.fr/

[16] World-Health-Organization(WHO) 2007 Control integral del cáncer cervicouterino. Guía de prácticas esenciales

[17] De Palo G, Dexeus S and Chanen W 2007 *Patología y tratamiento del tracto genital inferior* (España: Elsevier)

[18] Payam R, Lei T, and Huan L 2009 Cross validation in *Encyclopedia of Database Systems* Tamer zsu M, Ling L (Eds) (EUA: Springer)

[19] Pérez-Palacios G 1997 Norma Oficial Mexicana NOM-014-SSA2-1994, para la prevención, detección, diagnóstico, tratamiento, control y vigilancia epidemiológica del cáncer cervicouterino Diario Oficial de la Federación. Gobierno de México

[20] Pogue B W, Kaufman H B, Zelenchuk A, Harper W, Burke E E and Harper D M 2001 Analysis of acetic acid-induced whitening of high-grade squamous intraepithelial lesions *Journal of Biomedical Optics* **6** 4 397-403

[21] Rechy-Ramírez F, Acosta-Mesa H G, Mezura-Montes E, and Cruz-Ramírez N 2011 Time series discretization using evolutionary programming Sidorov B *Proc. of the 10th Mexican International Conf. on Artificial Intelligence MICAI 2011* (Berlin: Springer-Verlag)

[22] Sellor J W and Sankaranarayanan 2003 *Colposcopy and treatment of cervical intraepithelial neoplasia. A beginner's manual* (Francia: International Agency for Research on Cancer, IARC)

[23] Schmid-Saugeon P, Pitts J D, Kaufman H B, Zelenchuk A and Harper D M 2004 Time-resolved imaging of cervical acetowhitening *DRAFT* 1-42

[24] Stefanaki I M, Tosca A D, Themelis G C, Vazgiouraki E M, Dokianakis D N, Panayiotidis J G, et al. 2001 In vivo detection of human papilloma virus induced lesions of anogenital area after application of acetic acid: a novel and accurate approach to a trivial method *Journal of Photochemistry and Photobiology* **65** 115-121

[25] Syrjänen K J 2005 The colposcopy, cytology and histopathology of genital HPV infections *CME Journal of Gynecologic Oncology* **10** 46-41

[26] Tan P N, Steinbach M, and Kumar V 2006 *Introduction to Data Mining* (Boston: Pearson Addison Wesley)

[27] Zitova B and Flusser J 2003 Image registration methods: a survey *Image and vision computing* **21** 11 977-1000