ISSN 1870-4069

Feature Subset Selection and Typical Testors Applied to Breast Cancer Cells

Alexis Gallegos, Dolores Torres, Francisco Álvarez, Aurora Torres

Universidad Autónoma de Aguascalientes, Aguascalientes,
Mexico

alexisEdm@gmail.com, mdtorres@correo.uaa.mx,
fjalvar@correo.uaa.mx, atorres@correo.uaa.mx

Abstract. One of the most significant concerns around the world has been human health. So far, there have been little discussion about how the Computer Science can help to improve diagnostics of many diseases. In this paper, the aim is to prove that the Computer Science can offer techniques to help the improvement of the diagnosis of medical pathologies. For that, Feature Subset Selection and Typical Testors will be applied to a breast cancer database. Nowadays, cancer is a medical condition difficult to ignore, which has special interest by the specialists for finding effective methods to prevent and cure it. The database contains a feature set of breast cancer cells, which were subjected to an analysis of testors in order to find the minimum feature set that best describes benign and malignant cells. As a result, each typical testor found contains risk factors recognized by medical experts.

Keywords. Typical testor, feature subset selection, breast cancer, cancer cells, combinational logic.

1 Introduction

Breast Cancer is the most common female neoplasia[1] and the first death cause from tumor disease among women worldwide [2], accounting for 16% of all female cancers [3]. Nowadays, cancer has been become increasingly difficult to ignore. Every day, new studies on the causes and treatments are published; however, everyone agrees that the critical point of this studies is the timely diagnosis or cancer screening [4]. According to the National Cancer Institute, cancer screening means checking for cancer in people who have no symptoms that help doctors find and treat several types of cancer.

[1] Neoplasia is an abnormal mass of tissue whose growth exceeds the normal tissues, and it persists after cease the stimulus that triggered the change [1] D. Zicre, "Cátedra de Anatomía y Fisiología Patológicas," in *Neoplasia*, ed: Facultad de Ciencias Médicas. UNR, 2012.

Early detection is important because when abnormal tissue or cancer is found timely, it may be easier to treat. By the time symptoms appear, cancer has begun to spread and is harder to treat [5]. However, there are effective methods for screening only for some cancers.

Despite the usefulness of early detection, it can have some risks, as well as the methods used. For example, screening test can present a false-positive results; it means that the test indicates that the cancer is present when this is not true. Also, the test can have false-negative results, this indicates that cancer is not present even though it is.

Furthermore, over diagnosis is possible, this happens when screening test correctly shows that a person has cancer, but the cancer is slow growing and would not have harmed that person in his or her lifetime [5]. This justifies the need to improve the diagnosis of cancer.

The clinical diagnosis is a cognitive process that starts from the sensitive concrete thinking. It is linked to objective reality; it develops in abstract thinking and has the criterion of truth in practice [6]. It involves training, experience, pattern recognition and calculation of conditional probability, among other components, that human treatment has and therefore, is not free of errors that can cause sickness, damages, expenses and even death, especially in sensitive diseases such as cancer [7].

The errors represent an estimated 150 out of 1000 patients with misdiagnosis [8]. As such, the medical field is one of the areas that could be most benefit from close interaction with Computing Science and Mathematics to improve processes such as medical diagnosis [7]. Reason why it is decided to apply comprehensive mathematical methods to support diagnosis and prognosis of diseases such as cancer, in this case breast cancer.

This paper has been divided into three parts and organized the following way. The first part deals with important concepts in Typical Testors, Featured Subset Selection in computer science, Breast Cancer and its impact around the world. Next, the second section will examine the framework of this analysis, such as the previous research related to Testors Theory and the review of the methodology applied to breast cancer cells. Finally, the third section describes the results of the methodology and its review.

2 Important Concepts

2.1 Typical Testors

Testors Theory was formulated as an independent scientific direction of Mathematical Cybernetics in the 60's in the former Union of Soviet Socialist Republics (USSR), whose origin is linked to the use of mathematical logic methods for locating faults in electrical circuits that perform Boolean functions [9].

Later, testors were used to perform supervised classification and selection of variables in problems of geology [9, 10]. The use given in this article to the testors and typical testors is related to feature subset selection, whose precursor is Dmitriev, Zhuravlev and colleagues [10].

In this way, a testor is a subset of features that distinguishes objects from different classes [10]. According to Santiesteban and Pons [11], Shulcloper [9], and Torres [10], a typical testor is a testor that it is no longer possible to remove any feature without losing its status of testor. Otherwise, a typical testor is already formed by the minimum set of features needed to ensure the identification of the class to which a specific object belongs.

Typical testors determinne issues such as evaluation of informational weight of traits and selection of variables. They can reduce the dimension of the space of representation of objects [11] and the can be used as a set of support for classification algorithms [12]. So, consequently, the aim of this study is to prove that testor analysis can help to classify cells based on a real dataset; this will be explained in chapter 3.

2.2 Featured Subset Selection

Regularly, Featured Subset Selection (FSS) [13] is used to reduce dimensionality [14], which is used to efficiently reduce the number of variables, attributes or characteristics with which should describe the objects and to find their influence in a problem. This is an alternative method that starts by using the set of typical testors, taking out irrelevant or redundant features [11, 14].

FSS really has importance because reducing the number of features may help to decrease the cost of acquiring data and also make the classification models easier to understand [14, 15]. Also, the number of features could affect the accuracy of classification. Some authors have also studied the bias feature subset selection for classification learning [14].

The FSS problem has been studied by the statistics and machine learning communities for many years with high attention because of the enthusiastic research in data mining [16]. There are many potential benefits of variable and feature selection [17]:

- Facilitating data visualization and data understanding,
- Reducing the measurement and storage requirements,
- Reducing training and utilization times,
- Defying the curse of dimensionality to improve prediction performance.

Selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant [17]. For example, the brute-force selection method evaluates exhaustively all possible combinations of the input features, and then finds the best subset [16]. Later, chapters 3 and 4 will describe a brute-force method applied to the already mentioned dataset related to breast cancer cell with the aim of evaluating if a cell is malign or benign and calculate the informational weight of every feature.

2.3 Informational Weight

The use of the informational weight for feature subset selection is an excellent tool that shows tangible results [14]. Informational weight of a feature is a score, in other words, is the measure of significance to predict whether an object belongs to a group or to another (Classification) [10, 18].

2.4 Breast Cancer

Firstly, cancer is a collection of related diseases. In all types of cancer, some body's cells begin to divide without stopping and spread into surrounding tissues. Cancer can start almost anywhere in the human body, which is made up of trillion of cells [19]. It is the result of mutations, or abnormal changes in the genes that regulate cell growth. Normally, human cells grow and divide to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place [19, 20].

When cancer develops, this orderly process breaks down. Mutations can „turn on" certain genes and „turn off" others in a cell. The modified cell acquires the ability to divide without any control or order, which produces more identical cells and generates a tumor [19, 20].

Consequently, breast cancer is a malignant tumor that has been developed from the breast cells [21]. The breast is made up of glands called lobules that can make milk and thin tubes called ducts that carry milk from the lobules to the nipple, generally breast cancer originates in cells of those lobules [20, 21].

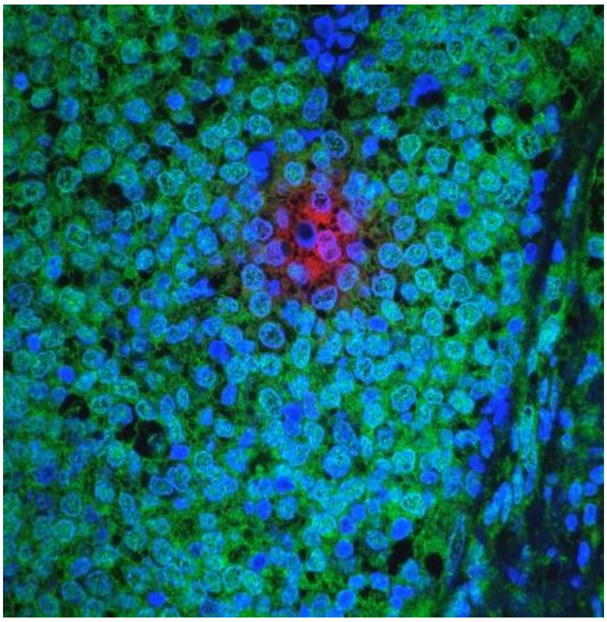

Fig. 1. Invasive breast cancer tumor [19].

Breast cancer has had a great impact worldwide, according to the Pan American Health Organization (PAHO) in the American continent breast cancer is the most common between the women with 29% of the cancer cases. PAHO estimates more than 596,000 new cases and more than 142,100 deaths in the region by 2030, mainly in the area of Latin America and the Caribbean [22]. Next figure shows the incidence of breast malignant tumors in women over age 20 years divided by age group, year 2014:

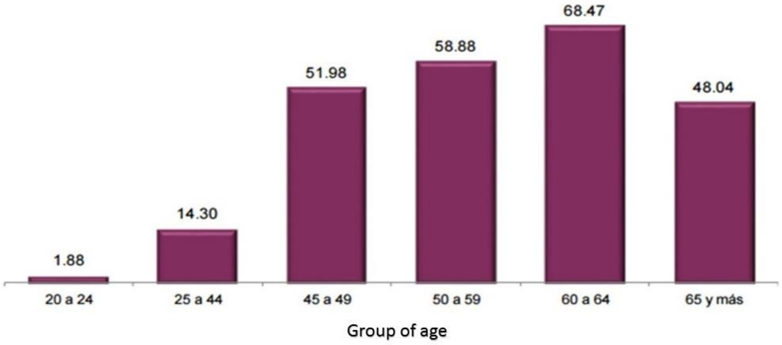

Fig. 2 Incidence of breast malignant tumors in women over age 20 years divided by age group. For 100 thousand women for each age group, INEGI [22].

In general, any type of cancer represents an important impact to the physical state of the person, his or her emotional sphere, a high cost of treatment and can even undermine the economy of the countries; so the prevention and a timely diagnosis are critical to address this problem [23]. Hence, this paper is focused in the application of Feature Subset Selection and Typical Testors to improve the diagnosis of cancer in body’s cells. Next chapters will explain the study done.

3 Framework

The general methodology used for this paper is shown in figure 3 below:

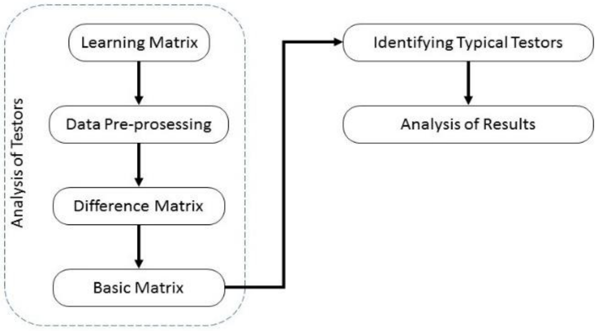

Fig. 3. General Methodology.

As seen if Figure 3 the methodology used needs a Learning Matrix (LM), which is the source of the information that contains descriptions of objects [9, 11]. For this paper, the LM comes from the University of California and their Machine Learning Repository. The repository is Wisconsin Diagnostic Breast Cancer [24].

The database contains the diagnosis and 10 features computed from a digitized image of a fine needle aspirate of a breast mass and describes characteristics of a cell nucleus present in the image [24]. The image below shows an example of this images.

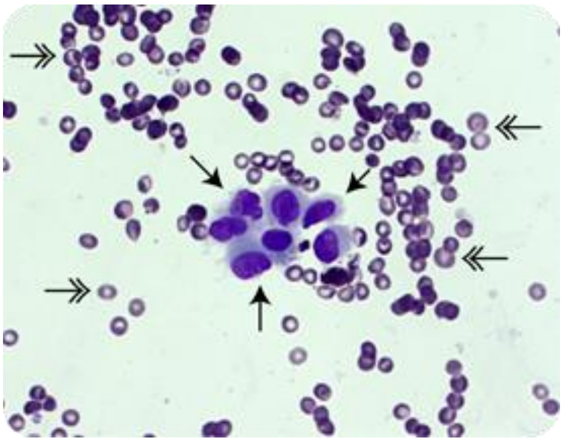

Fig. 4. Fine needle aspirate of a breast mass [25].

The real-valued features for each cell nucleus are [24, 26]:

1. Diagnosis (M=malignant, B=benign)
2. Radius,
3. Texture,
4. Perimeter,
5. Area,
6. Smoothness,
7. Compactness,
8. Concavity,
9. Concave points,
10. Symetry,
11. Fractal dimension

Diagnosis is the final result of evaluating the characteristics of the cell with a computer vision diagnostic system [26-28]. Each cell in the database has one of two possible diagnoses, it can be a malignant cell registered with an uppercase letter M or a benign cell registered with an uppercase letter B.

The radius of a cell was measured by averaging the length of radial lines segments defined by the centroid of the cell and the individual points in the boundary of the cell. The radial lines were defined by Street, Wolberg and Mangasarin in [26, 27] as can be seen in figure 5.

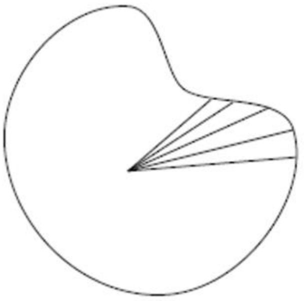

Fig. 5. Radial lines measured in a cell [26].

As mentioned earlier, each cell feature was extracted by a computer vision system, so, the texture was measured by finding variance of gray scale intensities in computer pixels [26, 27]. See Figure 6.

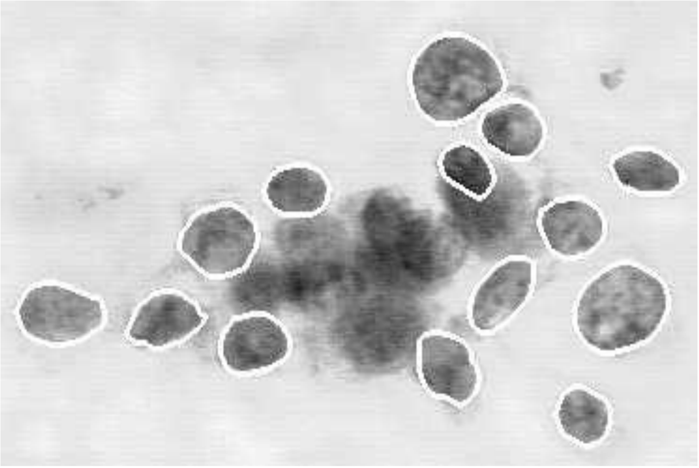

Fig. 6. Example of image taken by a computer vision system and boundary cell [26].

The perimeter is defined as the total distance between individual points named snake points in [26]. Those individual points comprise the white lines in the boundary of the cells (see Figure 6).

The area is measured by counting the number of pixels on the interior of the white line adding one-half of the pixels in the perimeter [26].

Meanwhile, the smoothness of a nuclear contour is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it [26], see Figure 5. Basically, the smoothness is a local variation in radius lengths [24].

Perimeter and area are combined to calculate a measure of compactness, which is a measure of shape [26, 27]. The compactness if given by the formula:

Compactness=perimeterˆ2/area.

This number is minimized by a circular disk and increases with the irregularity of the boundary and also increases for elongated cell nuclei, which can indicate an increased probability of malignancy [26].

Concavity analyzes the shape irregularities in a cell nucleus. Street, Wolberg and Mangasarian [26] measure the number and severity of concavities or indentations in a cell nucleus. They draw chords between non-adjacent white points and measure the extent to which the actual boundary of the nucleus lies on the inside of each chord. See Figure 7.

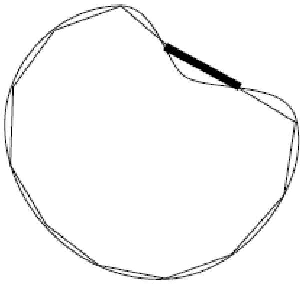

Fig. 7. Chords used to compute Concavity [26].

Concave points uses similar measure than Concavity but this feature only measures the number, rather than the magnitude, of contour concavities [26].

Symmetry is found the longest chord through the center. Then, according to [26], the length difference between lines perpendicular to the longest chord to the cell boundary in both directions was measured. This is illustrated in Figure 8.

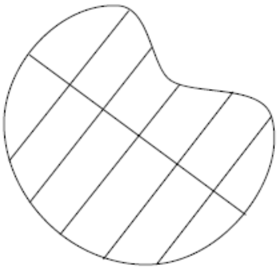

Fig. 8. Segments used in Symetry computation [26].

Finally, the Fractal Dimension is a shape feature [27], so, a higher value corresponds to a less contour and thus to a higher probability of malignancy [26]. The fractal dimension is aproximated using the coastline approximation by Madelbrot [26, 29]. The perimeter of the nucleus is measured using increasingly larger ‚rulers‘. This is, as the ruler size increases, decreasing the precision of the measurement, the observed perimeter decreases. Now, Plotting these to values on a log scale and measuring the downward slope gives the negative of an approximation of the fractal dimension [26]. This measurement is illustrated in Figure 9.

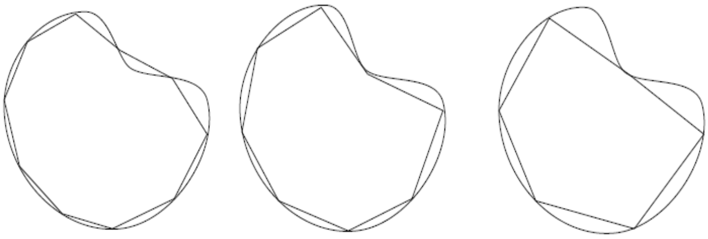

Fig. 9. Sequence of measurements for computing Fractal Dimension [26].

The database contains a total of 569 instances with 357 benign instances and 212 malignant instances. This dataset was preprocessed (Data Preprocessing, see Fig. 3), it means that is necessary a depth analysis of the database looking for duplicate instances in each class and contradictions (delete equal records but different diagnosis). Consequently, the database most contain unique instances.

At the end of Data Preprocessing, the dataset contains 130 benign instances and 146 malignant instances. Those 276 instances were analysed with the next steps of the methodology.

With the LM and C_1, ..., C_n comparison criteria specified [11] by a pathologist of each feature described earlier, a Difference Matrix (DM) is computed by comparing each instance from a class with each instance in the other classes following the comparison criteria of the corresponding feature. When a couple of feature are equal the matrix receives a 0, and 1 when the features are different [11]. A DM contains information that distinguishes objects from differents classes, which contains descriptions of objects [11, 30, 31].

Following this, the Basic Matrix (BM) is determined. The BM contains the basic differences from the DM without duplicates [11, 30]. According to Pons and Shulcloper [9, 32]: be i_q a row of the Difference Matrix, the row i_q is essential if there is not a row i_p that is subline of i_q, then the Basic Matrix contains the essential rows of the DM.

Next step is identifying typical testors. As mentioned earlier, typical testors are formed by the minimum set of features needed to ensure the identification of the class to which a specific object belongs.

The subset $\tau = \{X_{i_1}, \ldots, X_{i_s}\}$ of features from a LM is a Testor if each column from its BM is deleted, except those corresponding elements of τ, there is not a row composed by full of zeros. The subset τ is a Typical Testor whether any feature is deleted, the subset is no longer a testor [9, 11, 32].

Finally, the informational weight is calculated and it is possible begin an analysis of results. This information is explained below.

4 Results and Conclusions

At the end of the process, a validation is required which is done by a specialist in the area. For breast cancer and its cells, the specialists are the Oncologist and Pathologist. The final information must be attached to reality, for this reason a Computer Science specialist can not validate this information.

Finally, the tyical testors and their informational weight is shown below:

Table 1. Typical Testors.

Feature	Typical Testors (In column)	
	1	2
Radius	0	1
Texture	1	1
Perimeter	1	1
Area	1	0
Smoothness	1	1
Compactness	1	1
Concave points	1	1
Symetry	1	1
Fractal dimension	1	1

As can be seen, Table 1 shows the typical testors. The process found two typical testors and their resulting information that states that most of the features are critical to determine if a instance of the cell is malignant or benignant.

The Typical Testor 1 states a 0 in the radius feature and 1 in the other features. On the other hand, the Typical Testor 2 sets a 0 in the area and 1 in the other features. This means that the radius and the area ar interchangeable.

Table 2. Informational weight according to the typical testors.

Feature	Informational weight
Radius	50%
Texture	100%
Perimeter	100%
Area	50%
Smoothness	100%
Compactness	100%
Concave points	100%
Symetry	100%
Fractal dimension	100%

In other words, it is possible to classify a cell instance knowing at least one of the two features, the radius or the area. For example, an instance can be classified knowing its texture, perimeter, smoothness, compactness, concave points, symetry and fractal points but if radius is unknown, the area must be known. However, if the area is unknown, the radius must be known. Finally, in the best case, both values are known but is not possible to classify a cell if both features are unknown.

The informational weight is obtained by calculating a percentage factor that indicates the frequency of each variable in the set of typical testors [33]. Table 2 shows the informational weight of each feature.

The value of the informational weight represents the degree of importance of each feature analyzed in a classification process. A value of 100% indicates that the feature is critical and it can not be ignored in any case.

Also, it can be possible that one or more features have 0% of informational weight meaning that is not necessary, therefore the number of features is reduced and make the problem easier. Remember, this is one of the objective of the analysis.

5 Future Work

The analysis described in this paper is a exhaustive method to find the typical testors. Next step, is apply a Metaheuristics as alternative.

The main goal is to find typical testors through metaheutistc algorithms taking advantage of the already identified basic matrix. In this way, the alternative process will be an hybrid method.

Moreover, it is intended to apply both processes (hybrid and exhaustive) in more cases they represent the behavior of different pathologies.

Acknowledgements. We thank Dr. Luis Muñoz Fernandez for all his assistance and advices and Angela Paulina Pérez Díaz for her advices. We really appreciate your help.

References

1. Zicre, D.: Cátedra de Anatomía y Fisiología Patológicas. In: Neoplasia, ed: Facultad de Ciencias Médicas, UNR (2012)
2. Guerra Merino, I.: Factores pronostico del cáncer de mama en 108 mujeres menores de 36 años. Universidad Complutense de Madrid (2000)
3. CEAMEG: Cancer de Mama. Cancer de Mama, Vol. 1, pp. 1 (2014)
4. Canceronline. Detección Precoz de Cáncer. Available: http://www.canceronline.cl/index.php?option=com_content&view=article&id=48&Itemid=57
5. NIH.: Cancer Screening. Available: http://www.cancer.gov/about-cancer/screening (2015)
6. Pérez, N. M.: El diagnóstico médico: algunas consideraciones filosóficas. (2009)

7. Lugo-Reyes, S. O., Maldonado-Colín, G., Murata, C.: Inteligencia artificial para asistir el diagnóstico clínico en medicina. Artificial Intelligence to Assist Clinical Diagnosis in Medicine, Vol. 61, No. 2, pp. 110–120 (2014)
8. Reed, K.: HealthGrades Patient Safety in American Hospitals Study. Available: https://www.hospitals.healthgrades.com/
9. Ruíz, J., Alba, E., Lazo, M. : Introducción a la Teoría de Testores. Departamento de Ingeniería Electrica, CINVESTAV-IPN, pp. 197 (1995)
10. Torres, M. D., Torres, A., Torres, M. L., Bermudez, L., Ponce, E. E.: Factores Predisponentes en Relajación Residual Neuromuscular. Research in Computing Science, Vol. 93, pp. 163–174 (2015)
11. Santiesteban, Y., Pons, A.: LEX: A New Algorithm for the Calculus of all Typical Testors. Vol. 1, pp. 85–95
12. Lias-Rodríguez, A., Pons-Porrata, A.: Un nuevo Algoritmo de Escala Exterior para el Cálculo de los Testores Típicos. pp. 10, http://www.rcs.cic.ipn.mx/2015_93/Factores%20predisponentes%20en%20relajacion%20residual%20neuromuscular.pdf (2015)
13. Wang, G., Song, Q., Sun, H., Zhang, X.: A Feature Subset Selection Algorithm Automatic Recommendation Method. China: Cornell University Library, pp. 1–34 (2013)
14. Torres, D., Ponce de León, E., Torres, A., Ochoa, A., Díaz, E.: Hybridization of Evolutionary Mechanisms for Featured Subset Selection in Unsupervised Learning. MICAI 2009, Advances in Artificial Intelligence, pp. 610–621 (2009)
15. Pelikan, M., Sastry, K., Cantú-Paz, E.: Scalable Optimization vía Probabilistic Modeling: From Algorithms to Applications. Springer (2006)
16. Deng, K.: OMEGA: On-line Memory-Based General Purpose System Classifier. Doctor, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1998)
17. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, Vol. 3 (2003)
18. Cotilla, M. O.: Un Recorrido por la Sismología de Cuba. Cuba: Editorial Complutense, S. A. (2006)
19. N. C. Institute: What is Cancer? http://www.cancer.gov/about-cancer/understanding/what-is-cancer (2015)
20. Breastcancer.org.: ¿Qué es el Cáncer de mama? http://www.cancer.gov/about-cancer/understanding/what-is-cancer(2014).
21. NIH: Breast Cancer - Patient Version. National Cancer Institute.
22. INEGI: Estadísticas a Propósito del Día Mundial de la Lucha contra el Cáncer de Mama. In: Estadisticas Nacionales, ed. México, Instituto Nacional de Estadística y Geografía (2015)
23. INEGI: Estadísticas a Propósito del Día Mundial Contra el Cáncer. México, Instituto Nacional de Estadística y Geografía (2016)
24. Wolberg, W. H., Street, N., Mangasarian, O. L.: Wisconsin Diagnostic Breast Cancer (WDBC). California, Ed., USA (1995)
25. V. B. Imaging: Fine Needle Aspiration. http://www.breastimaging.vcu.edu/services/guided/fineneedle.html (2016)

26. Street, W. N., Wolberg, W. H., Mangasarian, O. L.: Nuclear Feature Extraction for Breast Tumor Diagnosis. In: International Symposium on Electronic Imaging: Science and Technology, Vol. 1905, pp. 861–870
27. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L.: Computerized breast cancer diagnosis and prognosis from fine- needle aspirates. Archives of surgery (Chicago, Ill.: 1960), Vol. 130, No. 5, pp. 511 (1995)
28. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L.: Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, Vol. 26, No. 7, pp. 792–796 (1995)
29. Mandelbrot, B. B.: The fractal geometry of nature. New York, W.H. Freeman (1982)
30. Ochoa-Somuano, J.: Técnicas de Selección de Atributos para la Categorización Automática de Escenas Visuales. Maestría, Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Morelos (2005)
31. Martínez-Sánchez, N., García-Lorenzo, M. M., García-Valdivia, Z. Z.: Modelo para Diseñar Sistemas de Enseñanza-Aprendizaje Inteligentes Utilizando el Razonamiento Basado en Casos. Revista Avances en Sistemas e Informática, Colombia,Vol. 6 (2009)
32. Pons-Porrata, A.: Desarrollo de Algoritmos para la Estructuración Dinámica de Información y su Aplicación a la Detección de Sucesos. Doctorado, Departamento de Lenguajes y Sistemas Informáticos, Universidad Jaume 1, Castellón (2004)
33. Rodríguez de Léon, P.: Heurística lógico combinatoria para la selección de subconjuntos de características en diabetes mellitus. Tesis (maestría en informática y tecnologías computacionales), Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Aguascalientes, Ags., Méx. (2016)