CLASIFICACIÓN DE TUMORES EN MAMOGRAFÍAS MEDIANTE USO COMBINADO DE RBP Y FILTROS SOBEL

Enrique Calot, Hernán Merlino, Paola Britos, Ramón García-Martinez

Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA
Centro de Ingeniería del Software e Ingeniería del Conocimiento. ITBA

enrique@calot.info, {hmerlino, pbritos, rgm}@itba.edu.ar

Resumen

La aplicación de sistemas inteligentes a áreas de la medicina es aún un campo abierto a la investigación. Este artículo propone una mejora a la metodología que aplica redes neuronales para clasificar tumores en mamografías de manera automática. Para esto se procederá a introducir un filtro Sobel sobre la imagen preprocesada para así obtener mejores datos de entrada que alimentarán a la red neuronal. Esta red neuronal es la que realizará la clasificación final. Por las características específicas de los tipos de tumores, este filtro y la metodología propuesta para aplicarlo ha probado ser eficiente y nos ha dado mejores resultados que los obtenidos por la metodología anterior.

Palabras Clave. Mamografía, Filtro Sobel, Back Propagation, Tumor, Procesamiento de Imágenes, Sistemas Inteligentes, Contornos, Patrones Visuales.

1. INTRODUCCIÓN

Mediante la utilización de sistemas inteligentes es posible automatizar el procesamiento de grandes volúmenes de información [Zrimec, 2007]. En la medicina esto abre las puertas a un área de desarrollo incipiente y novel [Zorman *et al*., 2003]. Se estima que en un futuro cercano este tipo de técnicas podrá servir -bajo supervisión médica- de pre-diagnósticos [Chan *et al*., 1987].

El cáncer de mama es el tumor más frecuente en la mujer, representando el 31% de todos los tumores de la población femenina. Aproximadamente una de cada ocho mujeres habrá desarrollado un cáncer de mama en el curso de su vida. Éste tipo de cáncer ocupa el primer lugar entre las causas de muerte por cáncer en la mujer adulta, con una tasa ajustada de mortalidad de 27,32 cada cien mil mujeres en Argentina. Los beneficios del *screening* mamario han sido demostrados en numerosos estudios aleatorios desde mediados de la década de 1980 a la fecha. En éstos se ve una reducción del índice de mortalidad por cáncer de mama en por lo menos un 25% [AMA, 2006]. Es por ello que, físicos, ingenieros y médicos están en la búsqueda de nuevas herramientas para combatirlo y permitir al médico obtener una segunda opinión [Gokhale & Aslandonga, 2003; Simoff *et al*., 2002].

Se han utilizado varios métodos para clasificar anomalías en imágenes medicas, como *wavelets*, teoría de fractales, métodos estadísticos, los cuales en su mayoría han utilizado técnicas tomadas de la rama principal del procesamiento de imágenes. Además otros métodos se encuentran presentes en la literatura, como los basados en la teoría de conjuntos difusos, modelos de Markov y redes neuronales. La mayoría de los métodos asistivos mostraron ser herramientas potentes capaces de asistir al personal médico en hospitales permitiendo así obtener mejores resultados al diagnosticar un paciente [Ferrero *et al*, 2006; Antonie *et al*, 2001].

Enfocar este problema mediante redes neuronales está comenzando a ser un modelo a seguir y hay varios proyectos de desarrollo de software relacionados, sin embargo todos se encuentran en estado experimental. Uno de los últimos desarrollos ha obtenido un 60% de éxito [Ferrero *et al*, 2006] .

2. EL MODELO PROPUESTO

Este artículo propone una mejora al modelo propuesto en [Ferrero *et al*, 2006] donde se abordó el problema mediante varias capas de procesamiento.

2.1 Base Utilizada por la Propuesta

Luego de ser escaneada la imagen, se la somete a un proceso de ajuste de brillo y contraste para resaltar las diferencias obtenidas, luego ésta es utilizada en crudo como entrada para una red neuronal que será quien la clasifique obteniendo como resultado el tipo de anomalía.
La base de datos de mamografías utilizada por el citado trabajo fue la de la *Mammographic Image Analysis Society* (MIAS).

2.2 La Propuesta

En este artículo se proponen dos mejoras sustanciales. La primera es la utilización de una base de datos *The Digital Database for Screening Mammography* (DDSM) [Heath, 1998; 2001] de la University of South Florida (USF), que posee imágenes de mayor resolución y tiene una cantidad mucho más grande de estudios agrandando así el tamaño de la muestra. Además se cuenta con mayor información sobre cada imagen, como el contorno de los tumores y varias otras clasificaciones que no estaban en la base de MIAS.
La segunda mejora es la utilización de un filtro de imágenes que permite encontrar gradientes y detectar contornos. Además proponemos analizar la anomalía en sí y no la mama entera utilizando la información de contorno que esta base de datos provee.

2.3 Definición de los Contornos

A diferencia de trabajos anteriores, este artículo plantea la idea conocer la ubicación del tumor y no analizar la mama por separado. Para poder ubicar el o los potenciales tumores será necesario conocer sus respectivos contornos.
Este artículo propone abstraerse de la forma en que se obtiene el contorno alrededor de un tumor; se asume que este es dato y la forma en que vendrá dado puede ser tanto por selección manual como por detección automática [Lee, 2006].
Una vez obtenido el contorno, se aplican técnicas especiales que dependen de la distancia a la zona contorneada y de la imagen superpuesta obteniendo información que va a servir para alimentar las redes neuronales. Existen dos formas de almacenar el contorno de un tumor para aplicar estos procesos, la primera es de manera vectorial y la segunda es como un mapa de bits o *bitmap*.
Si se aproxima el contorno por una elipse o circunferencia, el almacenamiento será mucho menor, solo deben ser guardados el centro de la circunferencia o los focos de la elipse.
La utilización del mapa de bits, en cambio, es mucho más precisa pero consume más recursos. En nuestros resultados experimentales, trabajando con imágenes de alta definición (16 bits), en

mamografías de más de 2200 por 4000 píxeles, obtenidas de la base DDSM [Heath, 1998], pueden ocupar aproximadamente 20Mb cada una. En el caso de la elipse, una forma de almacenarla vectorialmente es mediante sus focos. Si se parte de allí, una forma de obtener una buena imagen para ser superpuesta sobre un tumor, es la resultante de la ecuación 1.

$$I(x) = F\left(\frac{d(x, f_1) + d(x, f2)}{d(f_1, f_2)}\right) \qquad \textbf{(1)}$$

Esto se debe a que por definición de elipse, $d(x,f_1)+d(x,f_2)+d(f_1,f_2)=cte$ para todo x perteneciente a la elipse; y como la distancia focal $d(f_1,f_2)$ es constante en si misma, se obtiene que $d(x,f_1)+d(x,f_2)$ debe ser constante. Aplicando la desigualdad triangular es fácil de probar que el valor mínimo de esa expresión será $d(f_1,f_2)$ y esto ocurrirá cuando x pertenezca al segmento recto que une ambos focos. Es por esta razón que $(d(x,f_1)+d(x,f_2))/d(f_1,f_2)$ será un número representativo de varias elipses concéntricas que irán desde el segmento recto que une ambos focos (obteniendo el valor 1) hasta elipses de tamaño infinito. Estos valores pertenecientes al rango $[1;\infty)$ pueden ser fácilmente transformados en valores del rango $[0;1]$ mediante distribuciones acumulativas de probabilidad que tengan como media un valor cercano al del contorno del tumor y una varianza relacionada con la resolución de imagen. La F(x) representada en la ecuación 1 se refiere a esta transformación. Las distribuciones recomendadas deben ser aquellas con una baja curtosis para garantizar que la mayor pendiente se haga cerca de la media (esto viene dado gracias a que en la función de densidad los números más grandes estarían cercanos a la media y ésta, al ser la derivada de la función acumulativa, indicaría una pendiente muy pronunciada).

Figura 1. Contorno *bitmap*.

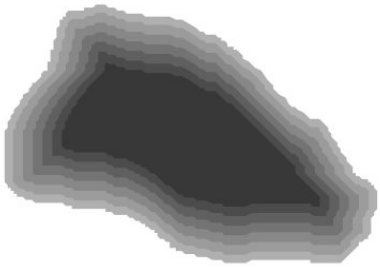

Figura 2. Contorno *bitmap* con sus regiones internas diferenciadas

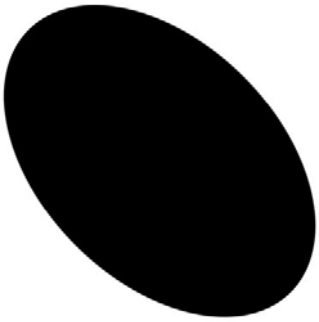

Figura 3. Contorno vectorial elíptico.

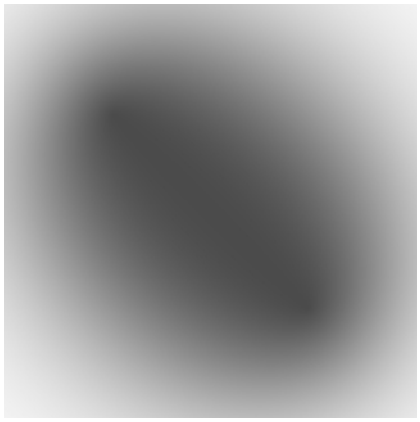

Figura 4. Contorno vectorial elíptico con regiones internas diferenciadas

Cabe destacar que en la figura 4 puede observarse que los focos son más oscuros que el resto de la elipse, esto es sólo una ilusión óptica y, por lo explicado líneas arriba, matemáticamente puede demostrarse que los focos son los valores máximos y son igual de oscuros que el resto del segmento recto que los une.
Este modelo tiene un problema y es que las elipses concéntricas generadas por esta expresión no están distanciadas uniformemente. Otros modelos pueden resolver este problema pero son más complicados y costosos a la hora de calcular.
La utilización de contornos por mapas de bits permite una mejor aproximación a los bordes del tumor, pero para obtener gradientes alrededor de ellos, las regiones circundantes deben ser procesadas y calculadas mediante algoritmos recursivos que midan la distancia al contorno. Esto puede ser realizado aplicando el algoritmo de Bellman-Ford [Bellman, 1958], si se considera al mapa de bits como un grafo donde cada píxel está unido con los cuatro inmediatamente adyacentes.

2.4 Capas Concéntricas

Una vez obtenidas las capas concéntricas se las toma como regiones que van a alimentar a las redes neuronales. Éstas pueden ser interiores o exteriores a la figura, la cantidad y tamaño depende de la resolución de la imagen y puede ser variada hasta observar mejores resultados. A estas regiones se les agrega una región más que comprende el centro del tumor. Cada una posee un tamaño que depende del contorno específico y por esto su tamaño es un valor posible como entrada a la red neuronal. El valor debe ser normalizado en el rango *[0;1]*. Para ello utilizamos la función acumulativa de una distribución exponencial negativa con una media igual al promedio de los tumores de la base DDSM, en nuestro caso 25000 píxeles obteniendo un $\lambda=1/25000$.
También se utilizó el promedio de la luminosidad de las regiones y la varianza existente.

2.5 Aplicación del Filtro Sobel

Debido a que los datos utilizados normalmente no aportan información relacionada con el principio de localidad de los focos de luminiscencia (es decir si hay saltos bruscos en la región, como contornos o ramificaciones) es necesario aplicar una estrategia que pueda aportar esta información a la red que hará la clasificación. La varianza es una buena medida de la diferencia de luminosidad, pero no depende de la posición en la que se encuentran los píxeles. A modo de ejemplo, si se tienen tres imágenes: a) una con gris al 50%, b) la otra con 50% blanco y 50% de negro distribuidos uniformemente en dos bloques, c) una imagen 50% blanco y 50% negro pero con valores distribuidos de manera alternada como se muestra en la figura 5 y por último, d) una imagen con un gradiente de grises, podríamos decir que los cuatro casos tienen un color medio de 0.5, sin embargo el caso a) no tiene varianza, y los casos b), c) y d) tienen la misma varianza. No obstante estamos ignorando el hecho de que la distribución en un caso, el b) es en bloques, en el otro, el c) es alternada y en el tercero, el d), es una escala de grises. Esta información debe ser provista a la red de alguna manera.

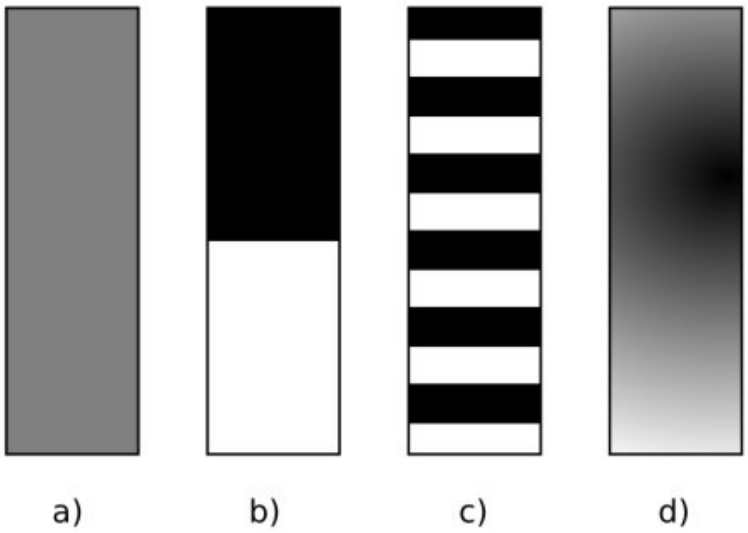

Figura 5. Ejemplos de distribuciones de luminosidad en una región.

Los tumores malignos tienden a producir ramificaciones, que son lugares por los que las células malignas pertenecientes al tumor escapan del contorno definido que los contiene e intentan avanzar sobre el tejido sano. En los tumores benignos, en cambio, este fenómeno no ocurre, permitiendo observar contornos bien definidos. Para detectar estas ramificaciones, es necesario utilizar una medida de la localidad de la luminosidad de las regiones aledañas al tumor. En nuestro ejemplo, un tumor maligno se asemejaría más al caso c) ó d) mientras que uno benigno se asemejaría a un b).
Por esta razón se decidió aplicar un filtro Sobel a la imagen antes de ser ingresada a la red neuronal, ya que este filtro es una medida del gradiente de diferencia de luminosidad y permite distinguir muy fácilmente ambos ejemplos que con la varianza no eran posibles de ser detectados. [Berhrend, 2006]
El filtro Sobel parte de la convolución de dos matrices con la imagen. Una matriz vertical y otra vertical que producen dos imágenes con las diferencias del gradiente en coordenadas cartesianas.

$$G_x = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix} * A; G_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} * A \qquad (2)$$

Una vez obtenidas ambas imágenes se aplica la transformación a polar mediante la expresión 3.

$$G = \sqrt{G_x{}^2 + G_y{}^2}; \Theta = arctan\left(\frac{G_y}{G_x}\right) \qquad (3)$$

Aplicándose esta transformación para cada píxel se producen dos imágenes, una que representa el módulo del gradiente y la otra su argumento [Sobel, 1968]. El módulo y el argumento nos indicarían respectivamente cuán pronunciado es el contorno a evaluar y la dirección de la mayor pendiente.
En nuestras experiencias se aplicaron las tres imágenes por igual: la original (figura 6), y las dos imágenes obtenidas por el filtro Sobel en coordenadas polares (figuras 7 y 8).

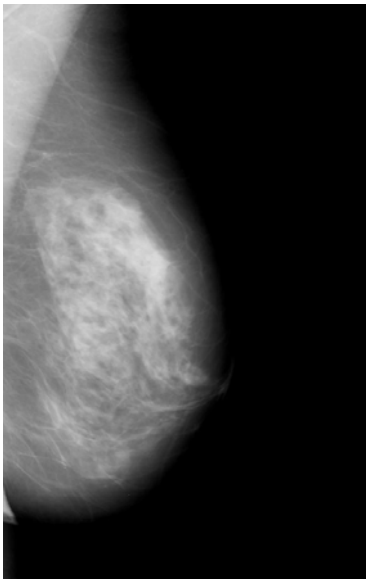

Figura 6. Imagen preprocesada

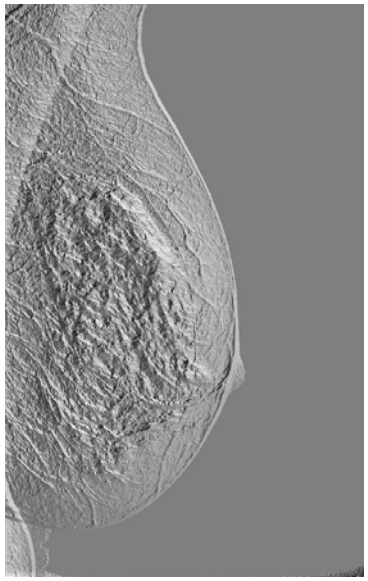

Figura 7. Argumento del gradiente de Sobel

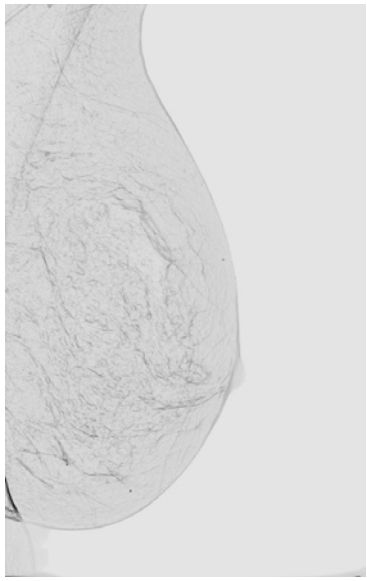

Figura 8. Módulo del gradiente de Sobel

Una vez obtenidas las imágenes, se deben generar las entradas para alimentar la red neuronal. Estas entradas serán las resultantes de superponer las regiones seleccionadas para cada imagen tomando la luminosidad media y varianza de cada región en cada imagen y agregándole el tamaño de la muestra como se muestra en el cuadro 1.

		tamaño	**media**	**varianza**
imagen		1	2	3
sobel	**modulo**		4	5
	argumento		6	7

Cuadro 1. Distribución de las posibles entradas para la red neuronal.

Cabe destacar que la información proporcionada por el filtro Sobel no es del todo redundante ya que al tomar la media y la varianza sobre una región se pierde mucha información. Con estos pasos se intenta utilizar el conocimiento proporcionado por los expertos en diagnóstico por imágenes a la red neuronal, es decir, la utilización del filtro dará información más procesada a al red proporcionando así información muy relevante. Esto se debe especialmente a que las ramificaciones en tumores malignos pueden ser detectadas muy fácilmente mediante este filtro.
Un gran problema con el que se enfrenta este tipo de soluciones es la sobrecarga de información de entrada, es posible que al haber mucha información para clasificar, la red no llegue a entrenarse del todo o necesite una muestra de mayor tamaño, por eso, para obtener los resultados experimentales se corrieron varias pruebas con distintas configuraciones detectando así qué entradas son las más importantes y cuales no tanto o tienen información potencialmente redundante.

2.6 Redes Neuronales y Entrada

La red neuronal utilizada es la de *back propagation*. Esta tiene una configuración de una capa de neuronas de entrada, una o varias capas ocultas en el medio y una capa de neuronas de salida. Cada unión entre una neurona y otra está asociada a un peso. Se define a w_{ij} como el peso que alimenta a

la neurona *j* desde la neurona *i* y se dice que el valor de la neurona *j*, x_j, será la sumatoria de $w_{ij}\ x_i$ para todo *i*.
Para entrenar la red mediante el algoritmo de *back propagation*, se realiza este procedimiento calculando el valor de todas las neuronas. Sus pesos son inicializados de manera aleatoria. Una vez que se llega a la capa de salida se compara los resultados obtenidos con los deseados obteniendo una medida del error de la red para el dato evaluado. Se realiza una retropropagación de este error ajustando así todos los pesos. Este procedimiento se realiza una vez por cada dato de un conjunto de entrenamiento y al finalizar se itera una cierta cantidad de veces para asegurarse de que los resultados tiendan al del conjunto de datos y los errores sean cada vez más chicos.
Al finalizar este procedimiento la red queda entrenada y puede ser corrida con entradas obteniendo así una salida relacionada, en nuestro caso, la clasificación del tumor.
En los resultados experimentales siempre hemos utilizado dos tercios de la muestra como datos de entrenamiento y el tercio restante (el cual es completamente independiente de los datos de entrenamiento) como datos de prueba. Conociendo los resultados médicos de ambos casos es posible calcular los errores reales ocurridos con el conjunto de prueba y así obtener una medida de la calidad del experimento.

3. RESULTADOS EXPERIMENTALES

Se experimentó con las imágenes de la base DDSM cuyo preprocesamiento fue ajustado de la misma forma que en artículos anteriores. Fueron detectados los valores mínimo y máximo de luminosidad para luego reajustar la luminosidad de cada pixel a un rango de 65536 valores posibles en una escala de grises (16 bits).
Una vez realizado este paso se procedió a calcular el filtro sobel en sus dos componentes para todas las imágenes cuya mama presentaba anomalías contorneadas.
Se utilizaron varias configuraciones de entrada y redes neuronales, en varios casos se obtuvo el máximo de 73%. Nuestra medida de exactitud siempre fue en la clasificación benigno-maligno, en otras categorías la red fue mucho más precisa. No consideramos el error de cada estudio, sino que el numero expuesto significa que de todas las muestras evaluadas, el 73% de ellas obtuvo el resultado correcto independientemente de su intervalo de error.
También se observó que la varianza no fue necesaria en ninguna de las imágenes, sino que tendía a sobrecargar de información a la red. Los mejores resultados fueron con 10 capas de cada lado del contorno.
La configuración que produjo mejores resultados fue la de una red de tres capas, es decir con una sola capa intermedia, la cual poseía 12 neuronas.
El tiempo de procesamiento de 271 estudios, que incluye la descompresión y el cálculo del filtro Sobel para cada imagen del estudio, fue de 42 minutos en promedio utilizando una máquina relativamente potente para la fecha. Los tiempos de entrenamiento estuvieron en un intervalo de 3 a 5 minutos para las redes de 12 neuronas. El tiempo de corrida de las redes fue instantáneo.

4. CONCLUSIONES

Con un 73% de resultados correctos, se obtuvo un 13% más que en trabajos anteriores y se estima que pueden ser aún mejores si se realiza un *clustering* antes de utilizar las redes *back propagation*.
Las corridas experimentales muestran que es posible mejorar la calidad de la detección mediante el esquema de regiones concéntricas propuesto, siempre y cuando se cuente con el contorno del tumor como dato. La obtención del contorno por medios automatizados puede ser un tema interesante a

abordar en futuras publicaciones, así como también lo sería la utilización de *clustering* antes de obtener la entrada de los datos.
Siendo más específicos podemos decir que la utilización de un filtro Sobel hace innecesaria la inclusión de la varianza de luminosidad de la región y que el argumento del filtro Sobel tampoco influye en los resultados. Por lo tanto es recomendable incluir solamente el módulo obtenido por el filtro. Tampoco es necesaria una configuración de red muy compleja y con pocas neuronas intermedias es posible obtener buenos resultados.

5. REFERENCIAS

AMA. 2006. *Consenso Nacional Inter-Sociedades sobre Cáncer de Mama: Pautas para el Diagnóstico y Manejo de las Lesiones Mamarias Subclínicas*. Asociación Médica Argentina. 2006.

Antonie, M.; Zaïene, O.; Coman, A. 2001. *Application of data mining techniques for medical image classification*. Proceedings of the Second International Workshop on Multimedia Data Mining. San Francisco.

Bellman, R. 1958. *On a Routing Problem*. En Quarterly of Applied Mathematics, 16(1), pp.87-90.

Berhrend, P. 2006. Identificación de marcas en la industria siderúrgica. Reportes técnicos en ingeniería del software, 8(2):43-46. ISSN: 16775002

Chan, H. P.; Doi, K.;Galhotra, S.; Vyborny, C. J.; MacMahon, H.; Jokich, P. M. 1987. *Image Feature Analysis and Computer-Aided Diagnosis in Digital Radiography: Part 1 Automated Detection of Microcalcifications in Mammography*, Medical Physics, vol. 14, pp. 538-548,

Ferrero, G.; Britos, P.; García-Martínez, R. 2006. *Detection of Breast Lesions in Medical Digital Imaging Using Neural Networks*. IFIP International Federation for Information Processing, Volume 218, Professional Practice in Artificial Intelligence, eds. J. Debenham, (Boston: Springer), pp. 1-10.

Gokhale, M.; Aslandonga Y. 2003. *A Visualization Oriented Data Mining Tool for Biomedical Images*. Departament of Computer Science and Engineering, University of Texas at Arlington.

Heath, M.; Bowyer, K.; Kopans, D; Kegelmeyer, W. P.; Moore, R.; Chang, K.; MunishKumaran, S. 1998. *Current status of the Digital Database for Screening Mammography*, Digital Mammography, 457-460, Kluwer Academic Publishers; Proceedings of the Fourth International Workshop on Digital Mammography.

Heath, M.; Bowyer, K.; Kopans, D; Moore, R.; Kegelmeyer, W. P. 2001. *The Digital Database for Screening Mammography*, Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., 212-218, Medical Physics Publishing, ISBN 1-930524-00-5.

Lee, N. J. 2006. Computer-*Aided Diagnostic Systems for Digital Mammograms*. Tesis de Mágister, Louisiana State University.

Selman, S. 2000. *Data Mining of Digital Mammograms* Will Aid in War against Cancer. www.gatech.edu. Página vigente al 27 de marzo de 2008.

Simoff, S.; Djeraba, C.; Zaïane, O. 2002. *Multimedia Data Mining between Promise and Problems*. 3rd Edition of the International Workshop on Multimedia Data Mining. ACM SIGKDD Explorations 4(3): 118-121. December 2002.

Sobel, I., Feldman,G.. 1968. *A 3x3 Isotropic Gradient Operator for Image Processing*, presentado en la Stanford Artificial Project.

Zorman, M.; Kokol, P.; Lenic, M.; Povalej, P.; Stiglic, B.; Flisar, D. 2003. *Intelligent platform for automatic medical knowledge acquisition: detection and understanding of neural dysfunctions*. Lab. for Syst. Design, Maribor Univ., Slovenia; Computer-Based Medical Systems, Proceedings. 16th IEEE Symposium, pp. 136-141. ISSN 1063-71258

Zrimec, T.; Busayarat, S. 2007. *A System for Computer Aided Detection of Diseases Patterns in High Resolution CT Images of the Lungs*. Computer-Based Medical Systems. CBMS apos;07. Twentieth IEEE International Symposium on Volume, Issue, 20-22 June, pp. 41-46