Método Heurístico para el Diagnóstico de Cáncer de Mama basado en Minería de Datos

Sandro Saul Camacho Centellas
Postgrado en Informática
Universidad Mayor de San Andrés - UMSA
La Paz, Bolivia
sandrex@live.com.mx

***Resumen*— El cáncer de mama es una de las principales causas de muerte en las mujeres; por esta razón, existe una amplia evidencia de que la detección temprana juega un papel importante en la reducción de la mortalidad de este cáncer, muchos métodos han sido investigados para mejorar una detección temprana; actualmente se utiliza la mamografía, mamografía 3d y la Tomografía axial; entre otros. Sin embargo, la mamografía es el medio más accesible y menos dañino para proporcionar información útil sobre la presencia de cáncer de mama. Debido a la dificultad en la interpretación de las mamografías, son necesarias múltiples lecturas de un único examen; de este modo, en este trabajo se propone aumentar la fiabilidad de un diagnóstico con la utilización de procesamiento de imágenes digitales y análisis de las mismas; mediante un Método Heurístico basado en Minería de Datos para extraer información esencial de las imágenes mamográficas y transformarlas en patrones. Posteriormente, se clasificará en sub-grupos de patrones para la conformación de familias mediante la homogeneidad y maximización de índices de coincidencia. El enfoque propuesto reduce la intervención humana y mejora la precisión de los resultados computacionales y es relativamente fácil de adquirir los datos.**

Palabras clave*— *Heuristica; Mineria de Datos, patrones

I. INTRODUCIÓN

El cáncer de mama es el más frecuente en las mujeres; según el Instituto Nacional del Cáncer Americano, una de cada ocho mujeres que viva hasta los 70 años, desarrollará cáncer de mama a lo largo de su vida y, entre 25 a 30 mujeres mueren por esta enfermedad. La incidencia del cáncer mamario ha aumentado de un 5% en 1940 a un 12% en 1995. También en España es el tumor de mayor incidencia en la mujer (entre 40 a 70 mujeres por cada 100.000), según el Registro Español de Cáncer. En las mujeres, el cáncer en general, es la causa de muerte en el 19,9% de los casos, de los cuales, el más frecuente (16,6%) es el de mama, Se ha determinado que el cáncer de mamas es la segunda causa de muerte de mujeres con cáncer, se estima que 22 de cada 100.000 mujeres poseen dicha enfermedad (Abalo, 2003).

El Ministerio de Salud de Bolivia en el año 2010 informó que en Bolivia, 26,57 por cada 100.000 mujeres desarrollaron Cáncer de Mama, entre tanto la mortalidad por ese mal alcanzó a 8,71 de cada 100.000. Por esa razón, el Ministerio de Salud elaboró y aplicó líneas estratégicas dentro del Plan Nacional de Prevención, Control y Seguimiento del Cáncer de Mama que tiene como fin reducir la alta incidencia y la mortalidad por esta dolencia.

El cáncer de mama es una de las pocas enfermedades cancerosas que se pueden diagnosticar precozmente; antes de que se note algún síntoma; la mamografía es una prueba de imagen por rayos X que detecta la presencia del tumor en la mama, antes de que sea perceptible al tacto, los especialistas recomiendan que todas las mujeres se realicen esta prueba cada año, a partir de los 50 a 55 años.

La mamografía permite detectar lesiones en la mama hasta dos años antes de que sean palpables y cuando aún no han invadido en profundidad ni se han diseminado a los ganglios ni a otros órganos. Cuando el tumor se detecta en estas etapas precoces es posible aplicar tratamientos menos agresivos; los cuales dejan menos secuelas físicas y psicológicas en la mujer (Antonie M., 2001) .

II. DESCRIPCIÓN DEL PROBLEMA

El diagnóstico clínico es un proceso cognitivo y complejo que precisa de: capacitación, experiencia, reconocimiento de patrones y cálculo de probabilidad condicional.

Muchos médicos han adquirido en el tiempo una intuición que facilita la tarea; la misma es producto de la experiencia de haber visto cientos de casos similares a lo largo de su carrera.

Los médicos menos experimentados pueden reducir sus errores al momento de diagnosticar a un paciente, realizando un proceso consciente y sistemático; tomando en cuenta la confiabilidad de sus apoyos diagnósticos y las fuentes más frecuentes de error.

Un médico puede sentirse fatigado o abordar de manera superficial a un paciente, puede no estar familiarizado con su enfermedad o estar influido por la opinión diagnóstica del colega que lo evaluó antes; y como consecuencia, no considerar un diagnóstico con un abordaje apropiado. Otro error común es la asociación de toda la sintomatología al diagnóstico inicial; utilizándolos solo para confirmar lo que se pensó en un inicio y descartando aquellos síntomas que no estén asociados.

En el cáncer de mama existen diagnósticos errados por factores humanos, debido a que la detección de la totalidad de los carcinomas

visibles a través de análisis retrospectivos de las imágenes, muchas veces resulta complicado; las lesiones varían desde alteraciones en partes blandas de distintas formas y márgenes, hasta calcificaciones de diferente morfología, tamaño y distribución, que pueden ser representativas de malignidad; por lo que son necesarias múltiples lecturas de un mismo examen.

A través de la aplicación de diferentes investigaciones sobre minería de datos para el procesamiento de imágenes médicas han obtenidos muy buenos resultados mediante las redes neuronales para tareas de clasificación y agrupamiento (Antonie M., 2001). Pero, el presupuesto de los hospitales en Bolivia, no alcanza a cubrir el alto costo de dichas herramientas por lo cual se dificulta el acceso a esta tecnología.

Importancia para afrontar el problema

La interacción de diferentes disciplinas permite resolver problemas complejos, algunas de ellas son bases de datos, estadística, aprendizaje automático y procesamiento de imágenes (Han, 2001).

La Heurística es un procedimiento para resolver un problema de optimización bien definido mediante una aproximación intuitiva, en la que la estructura del problema se utiliza de forma inteligente para obtener una buena solución.

La Minería de Datos es el conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos; y describir de forma automatizada modelos previamente desconocidos (Piatetski-Shapiro, 1996).

Objetivo general

Elaborar un método heurístico que permita la selección de patrones obtenidos a través de la minería de datos para facilitar el diagnóstico oportuno del cáncer de mama.

Hipótesis

La aplicación de un método heurístico para la selección de patrones obtenidos a través de la minería de datos facilitará el diagnóstico oportuno del cáncer de mama.

III. MARCO TEÓRICO

Estado del Arte.

Se ha encontrado varios grupos de investigación en el diagnóstico del cáncer de mama, la mayor parte se encuentra en una fase experimental, con respecto a la explotación de datos aplicadas al procesamiento de imágenes, las investigaciones se viene realizado a través de grupos de investigación de grandes organizaciones gubernamentales como la Agencia Nacional Aéreo Espacial del gobierno de los Estados Unidos de Norteamérica (NASA) (Simoff S., 2002).

Actualmente la resolución de tareas de explotación de datos se está realizado con herramientas propietarias ofrecidas por las diferentes empresas líderes en tecnología de bases de datos, como ORACLE o MICROSOFT (Liebstein, 2002).

Heuristica

La evolución cronológica que ha llevado el trasplante renal en España. Apoyados en la heurística (Albacete, 2006).

Las personas ciegas, su cuerpo, el espacio y la representación mental, el planteo metodológico está centrado en la heurística, es decir, en el conocimiento, maduración y proyección de las fuerzas del individuo, que busca su genuina expresión, como miembro de una sociedad (Carla Beatriz, 2010)

Un enfoque heurístico para la programación de la producción en la industria de la fundición mediante lógica difusa, En este artículo se describe la concepción, el desarrollo y aplicación de una metodología para la programación de la producción en la industria de la fundición (Raul Landmann, 2011).

Diseño de una hiperheurística para la programación de la producción en ambientes job shop, El objetivo del trabajo es disminuir el tiempo de proceso (Makespan) e incrementar el tiempo de trabajo de las maquinas, diminuyendo el tiempo de ocio en ambientes de job shop, a través del diseño de una hiperheurística basada en colonia de hormigas y algoritmos genéticos (Omar Danilo Castrillón, 2010).

Mamografía

En 1913 el cirujano alemán Albert Salomón, publicó la importancia de realizar radiografías a las piezas de mastectomía, para demostrar la extensión del tumor a los ganglios axilares, así como, la diferencia radiológica entre los márgenes del carcinoma infiltrante versus el circunscrito, describiendo por primera vez, los cambios radiográficos correspondientes a la presencia de microcalcificaciones en estos tumores (A.Salomon, 1913).

La siguiente publicación sobre radiología mamaria data de 1927 cuando otro cirujano, Otto Kleinschmidt, menciona por primera vez las indicaciones de la mamografía en un capítulo de un libro de texto (Keinsshmidt O, 1927).

En 1930 Stafford L. Warren, un radiólogo del Rochester Memorial Hospital, Rochester, Nueva York, publicó la utilización de la técnica estereoscópica para las mamografías en vivo, usó película de grano fino, pantallas de refuerzo de grano fino, parrilla móvil para disminuir la radiación dispersa, 50-60 kV y 70 mA, con una distancia del tubo a la placa de 25 pulgadas y un tiempo de exposición de 2,5 segundos. (Warren SL, 1930).

Al año siguiente (1931), Walter Vogel en Leipzig (Vogel, 1932) y Paul Seabold en América (Seabold PS, 1931), publicaron de forma independiente sus investigaciones sobre mamografía, incluyendo la diferenciación entre enfermedad benigna y carcinoma.

En 1937 Hicken publicó la utilización del contraste en mamografía, realizando magnificas radiografías que mostraban los ductos lactofóricos normales y anómalos, los quistes, papilomas y carcinomas (Hicken NH Mammography, 1937).

Minería de Datos

La idea de la minería de datos viene desde los años 60, cuando los estadísticos de esa época manejaban términos como data fishing, data mining o data archeology, más tarde en los años 80, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros empezaron a fortalecer los términos de data miningy KDD.

A finales delos años 80 solo existían un par de empresas quienes se dedicaban a esta tecnología; para el 2.002 este número se multiplicó considerablemente, ya que existían más de 100 empresas en el mundo con un portafolio de más de 300 soluciones que utilizaban la tecnología.

En la actualidad se ha incursionado con mayor fuerza en el desarrollo de aplicaciones que utilizan la minería de datos. Existen un conjunto de técnicas y herramientas capaces de ayudar a la toma de decisiones de los expertos. A pesar de ser relativamente joven, la minería de datos presenta aplicaciones en casi todos los sectores de la sociedad. En la salud, a nivel internacional se destaca la "Aplicación de técnicas de minería de datos para el diagnóstico prematuro del cáncer de mamas". Este sistema se encarga de realizar un diagnóstico del cáncer de mama a partir de una base de datos de imágenes de mamografías (Vallejo Delgado N, 2012).

En Cuba se han desarrollado investigaciones como por ejemplo "Aplicaciones de la minería de datos para el análisis de la Información Clínica". Este estudio se basa en el apoyo a la toma de decisiones a partir de coronariografías realizadas a pacientes que padecen cardiopatías isquémicas (Rosete Suárez A, 2009).

La UCI tampoco ha estado ajena al desarrollo de aplicaciones que emplean la minería de datos, y en ese sentido se destaca el "Diagnóstico de enfermedades de transmisión sexual mediante técnicas de inteligencia artificial (Bañobre Corpas Y).

IV. IDENTIFICACIÓN DEL PROBLEMA

Definición del Problema

Se hace necesario contar con un Método Heurístico basado en minería de datos basados con los principios del Software Libre y distribución libre que permita coadyuvar con el diagnóstico de cáncer de mama

Objetivos

Obtener una base de datos relacional de imágenes mamográficas pre-procesadas con minería de datos y patrones.

Diseccionar los patrones y asociarlos en familias de patrones.

Recoger toda la información con datos básicos en una base de datos relacional.

Evaluar el nivel de uniformidad en las familias, a través del cálculo de los índices de homogeneidad.

Determinar rectángulos de contención de familias y abstraer los niveles de coincidencia, mediante el algoritmo de colonia de hormigas.

Generar las recomendaciones basadas en los resultados de experimentación.

Marco Legal

a. Constitución Política del Estado

Sección II "Derecho de la salud y a la Seguridad Social" se tomó en cuenta los siguientes artículos: Art. N° 35. Inc. I y II, Art. N° 36 Inc. I y II, Art. N° 37, referente a que "El estado tiene la obligación indeclinable de garantizar y sostener el derecho a la salud, que se constituye en una función suprema y primera responsabilidad financiera. Se priorizará la promoción de la salud y la prevención de las enfermedades"

En el capítulo segundo principios, valores y fines del estado, inciso 5. Garantizar el acceso de las personas a la educación, a la salud y al trabajo.

En el capítulo segundo derechos fundamentales Artículo 18.

I. Todas las personas tienen derecho a la salud.

II. El Estado garantiza la inclusión y el acceso a la salud de todas las personas, sin exclusión ni discriminación alguna.

III. El sistema único de salud será universal, gratuito, equitativo, intracultural, intercultural, participativo, con calidad, calidez y control social. El sistema se basa en los principios de solidaridad, eficiencia y corresponsabilidad y se desarrolla mediante políticas públicas en todos los niveles de gobierno

En el capítulo cuarto derechos de las naciones y pueblos indígena originario campesinos, Artículo 30 inciso 13. Al sistema de salud universal y gratuito que respete su cosmovisión y prácticas tradicionales.

V. ESBOZO DE LA SOLUCIÓN

Resolución del problema

El proceso de extracción de conocimiento de las imágenes mamográficas, se lo realizara en torno a las siguientes fases como se muestra en la Figura 1.

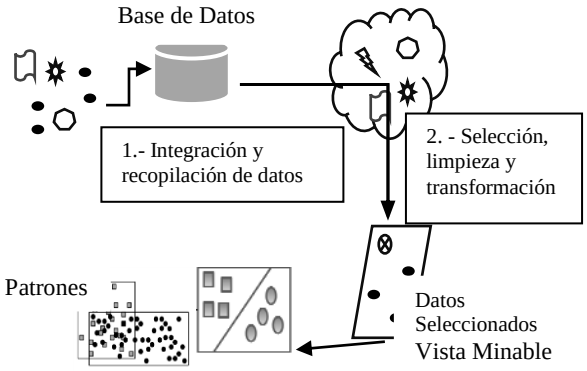

Fig. 1 Fases del proceso de extracción de Patrones de imágenes mamográficas con la minería de datos

En la fase de integración y recopilación de datos se determinan las fuentes de información que pueden ser útiles y donde conseguirlas. A continuación, se transforman todos los datos a un

formato común, frecuentemente mediante un almacén de datos que consiga unificar de manera operativa toda la información recogida, detectando y resolviendo las inconsistencias. Este almacén de datos facilita enormemente la navegación y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados. Dado que los datos provienen de diferentes fuentes, pueden contener valores erróneos o faltantes. Estas situaciones se tratan en la fase de selección, limpieza y transformación, en la que se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. Además, se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes con el objetivo de hacer más fácil la tarea propia de minería y para que los resultados de la misma sean más útiles. La selección incluye tanto una fusión horizontal (filas o registros) como vertical (atributos). Las dos primeras fases se suelen englobar bajo el nombre de preparación de datos. En la fase de minería de datos, se decide cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elige el método que se va a utilizar. En la fase de evaluación e interpretación se evalúan los patrones y se analizan por los expertos y, si es necesario, se vuelve a las fases anteriores para una nueva Introducción a la minería de datos iteración. Esto incluye resolver posibles conflictos con el conocimiento que se disponía anteriormente Finalmente, en la fase de difusión se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles usuarios.

Una visión general del metodo propuesto

El primer paso es el procesamiento de imágenes mamográficas y almacenarlos en una base de datos relacional, para obtener los datos necesarios. El segundo paso es aplicar la minería de datos para la obtención de patrones y diseccionar los mismos para conformar las familias de patrones. Los datos básicos y la información se recogen y se almacenan en otra base de datos. En el tercer paso, se realiza una evaluación del nivel de uniformidad en los patrones, mediante el cálculo de los índices de homogeneidad, para la conformación de familias de patrones. El cuarto paso es el uso la heurística para maximizar el nivel de coincidencia en la familia con caracteres específicos. El paso final es la generación de recomendaciones basadas en los resultados experimentales, ver figura 2.

Paso 1. Extracción de datos mediante el procesamiento de imágenes mamográficas

Paso 2. Aplicación de la Minería de Datos, para la obtención de patrones

Paso 3. Conformación de familias de patrones, mediante el índice de homogeneidad y uniformidad.

Paso 4. Maximización del nivel de coincidencias, mediante el método heurístico

Paso 5. Obtención de Resultados y generar recomendaciones

Fig. 2 Fases del Método Heurístico

Paso 1: Disección de patrones y conformación de Familia de Productos patrones.

En el primer paso, seleccionamos y diseccionamos los diferentes patrones obtenidos mediante la minería de datos. Esto, para evaluar el carácter en común hasta el nivel más bajo, dividiéndolos en subgrupos y conformando familias de patrones. Los datos se almacenan en una base de datos de patrones.

Paso 2: Recolección de datos

Recoger los datos necesarios obtenidos en el paso 1, mediante los siguientes datos:

- Tamaño y Geometría: Esta información se utiliza para comparar si los datos son únicos, son comunes o son variantes en una familia de patrones.

2.4 Paso 3: Evaluación de Similitud

Para medir la similitud dentro de una familia de patrones, se han propuesto varios índices de homogeneidad en la literatura (Jiao, 2000) (Thevenot, 2006). Un índice común es una métrica para evaluar el grado de coincidencia dentro de una familia de patrones. Se basa en parámetros del número en común de componentes para el análisis de una familia existente. Ellos están destinados a proporcionar información valiosa sobre el grado de coincidencia alcanzado dentro de una familia y la forma de mejorar y aumentar la uniformidad. En este trabajo, utilizamos la Métrica Integral de Similitud (MIS) para evaluar el carácter común de la familia de patrones.

Nombre		Desarrollado por	Medida de la similitud	Cero Similitud	Similitud Completa
GSI	Grado de similitud del índice	Collier (1981)	Toda la Familia	1	
IDCT	Indice de similitud de la constante	Wacher and Trleva (1986)	Toda la Familia	0	1
ISLP	Índice de similitud de la línea de producto	Kota, Sethuraman and Miller(2000)	Toda la Familia	0	100
PIDI	Porcentaje de índice de similitud	Siddike, Rose and Wang (1998)	individual	0	100
IS	Índice de similitud	Martin and ishii(1997)	Toda la Familia	0	1
PCS	Parte componente de similitud	Jiao and Tseng (2000)	Toda la Familia	1	
MCS	Métrica comprensiva para la similitud	Thevenot and Simpsom(2007)	Toda la Familia	0	1

Tabla 1 Lista de similitud de índices

Paso 4: Optimización mediante el Método Heurístico

En este trabajo, se utilizará el Algoritmo Genético (AG) para maximizar la (MIS). Un AG es un algoritmo de optimización estocástica de adaptación, que implican la búsqueda y optimización (Goldberg, D. E., 1989).

En este trabajo, cada atributo de un patrón se codifica como un entero, que más tarde se convierte en una representación binaria del AG. El AG maximiza la MIS, con sujeción a las siguientes características:

- Microcalcificaciones
- Masas circunscritas
- Masas espiculadas
- Distorsiones y Asimetrías

Para poder determinar:

- CDIS: carcinoma ductal in situ
- CDI: carcinoma ductal invasivo
- CLI: carcinoma lobular invasivo
- Cáncer de mama inflamatorio
- CLIS: carcinoma lobular in situ
- Carcinoma tubular de la mama
- Carcinoma medular de la mama
- Carcinoma mucinoso de la mama
- Carcinoma papilar de la mama
- Carcinoma cribiforme de la mama
- Cáncer de mama en hombres
- Enfermedad de Paget en el pezón
- Tumor filoides de la mama

Mediante la adición de estas características, se especifica un número máximo de análisis. De ahí que el AG ofrece recomendaciones que más influyen en el carácter común, ayudando al enfoque de análisis en patrones críticos para analizar. Actualmente no hay pautas para elegir el valor adecuado para este análisis. Sin embargo, se puede tomar un determinado porcentaje del número total de parámetros para este análisis.

Sobre la base de estas características, se eligen los patrones. Dentro de este conjunto de componentes, se consideran cuatro atributos (1) Microcalcificaciones, (2) Masas circunscritas, (3) Masas espiculadas, (4) Distorsiones y Asimetrías. Para una familia dada, si un atributo es común entre todos los patrones que utiliza esta familia, entonces este atributo no se considera durante la optimización.

Paso 5: Obtención de los resultados y recomendaciones

Una vez que la optimización es completa, la AG propone una nueva secuencia de análisis de patrones, que puede ser comparado con los patrones originalmente extraídos con la Minería de Datos. El AG no comprueba actualmente la viabilidad de la solución en cuenta; más bien, ofrece una lista clasificada de los parámetros que más influye en el grado de coincidencia en la familia de patrones. Esto puede ser visto como una reducción del análisis de patrones, donde se comprueba la viabilidad de la solución a posteriori en la lista de recomendaciones propuestas, en lugar de comprobar la viabilidad de una solución de volver analizar los patrones.

REFERENCIAS

A.Salomon. (1913). *Beiträge zur pathologie und klinik der mammakarzinome.* Arch Klin Chir.

Abalo, E. (17 de 8 de 2003). *Información para el público en general.* Obtenido de www.samas.org.ar

Albacete, M. P. (2006). Evolución cronológica del trasplante renal en España.

Antonie M., Z. O. (2001). *Application of Data Mining Techniques for Medical Image Classification* (segunda ed.). (S. Explorations, Ed.) International Workshop on Multimedia Data Minig.

Bañobre Corpas Y, B. G. *Diagnóstico de Enfermedades de Transmisión Sexual mediante técnicas de Inteligencia Artificial.* . Universidad de las Ciencias Informáticas, La Habana.

Carla Beatriz, G. (2010). LAS PERSONAS CIEGAS, SU CUERPO, EL ESPACIO Y LA REPRESENTACIÓN MENTAL.

Fundación Científica de la Asociación Española contra el Cáncer. (2004). *El cáncer en España.*

Goldberg, D. E. (1989). Genetic Algorithm in Search, Optimization and Machine Learning, Addison-Wesley Publishing Company Inc. Reading, PA.

Han, J. Y. (2001). *Data Mining. Concepts and Techniques. 548.* Editorial Morgan Kaufmann.

Hicken NH Mammography. (1937). *The roentgenographic diagnosis of breast tumours by means of contrast media.* Surgery, Gynaecology and Obstetrics.

Jiao, J. a. (2000). *Understanding Product Family for Mass Customization by Developing Commonality Indices*, (11 ed., Vol. III). Journal of Engineering Design.

Keinsshmidt O, B. Z. (1927). *Die Klinik der bösartigen geschwulste.*

Liebstein, L. (2002). Data Mining. Teoría e Práctica. *(16)*. Rio, Brasil.

Omar Danilo Castrillón, W. A. (2010). DISEÑO DE UNA HIPERHEURÍSTICA PARA LA PROGRAMACIÓN DE LA PRODUCCIÓN EN AMBIENTES JOB SHOP. *Ingeniare. Revista chilena de ingeniería*, 203-214.

Piatetski-Shapiro, G. U. (1996). *From data mining to Knowledge.* California, EEUU: AAAI Press/MIT Press.

Raul Landmann, R. H. (2011). Un enfoque heurístico para la programación de la producción en la industria de la fundición mediante lógica difusa.

Rosete Suárez A, R. D. (2009). Predicción de pacientes diabéticos. Preprocesado para Minería de Datos. *Informatica Médica*.

Seabold PS. (1931). *Roentgenographic diagnosis of diseases of the breast.* Surg Gynecol Obstet.

Simoff S., D. C. (2002). *Multimedia Data Mining between Promise and Problems* (Tercera ed.). SIGKDD Explorations.

Thevenot, H. J. (2006). *A Comprehensive Metric for Evaluating Commonality in a Product Family*, (17 ed., Vol. II). Journal of Engineering Design.

Vallejo Delgado N, R. J. (2012). *Aplicación de técnicas de minería de datos para el diagnóstico prematuro de cáncer.*

Vogel, W. (1932). *Die roentgendarstellung der mammatumoren.* Arch Kin Chir.

Warren SL. (1930). *Roentgenologic study of the breast.*