Información Tecnológica
Vol. 29(3), 257-270 (2018)
http://dx.doi.org/10.4067/S0718-07642018000300257

Sistema Predictivo Bayesiano para Detección del Cáncer de Mama

Omar D. Castrillón[(1)], Eduardo Castaño[(2)] y Luis F. Castillo[(1,3)]
(1)Universidad Nacional de Colombia, Sede Manizales. Facultad de Ingeniería y Arquitectura, Departamento de Ingeniería Industrial, GTA en Innovación y Desarrollo Tecnológico, Campus la Nubia - Manizales, Colombia. (e-mail: odcastrillong@unal.edu.co, lfcastilloos@unal.edu.co)
(2) Universidad de Caldas. Facultad de Ciencias para la Salud, Programa de Medicina, Colombia. (e-mail: eduardo.castano@ucaldas.edu.co)
(3) Universidad de Caldas, Facultad de Ingenierías, Depto. de Sistemas e Informática, GITIR Grupo Inv. Tecnologías Información y Redes, Manizales, Colombia. (e-mail: luis.castillo@ucaldas.edu.co)

Recibido May. 24, 2017; Aceptado Ago. 3, 2017; Versión final Sep. 15, 2017, Publicado Jun. 2018

Resumen

Se propone un método predictivo para detectar el cáncer de mama, basado en las siguientes variables: edad, peso, talla, índice de masa corporal, escolaridad, estrato socioeconómico, seguridad social, fumador, cuando dejo de fumar, fumador pasivo, consume licor, cantidad de licor, herencia familiar de cáncer, edad de la menarca, menopausia, embarazos, partos, edad del primer parto, lactancia, consumo de anticonceptivos orales, cuanto años consumió anticonceptivos orales, tiempo de suspensión de anticonceptivos orales, terapia de reemplazo hormonal y presencia del gen GSTM1. Tomando como referencias pacientes de la región central de Colombia (Caldas), se definieron dos bases de datos, una de personas sin cáncer y otra de personas con cáncer. La misma base de datos de entrenamiento fue empleada para prueba. La metodología propuesta, define y entrena un sistema de clasificación bayesiano, con una base de datos de pacientes con cáncer y sin cáncer. Posteriormente, se realiza una validación del sistema con el fin de determinar el número de aciertos y errores en el reconocimiento de esta enfermedad. Como resultado, se logra un porcentaje de aciertos del 100%.

Palabras clave: clasificador bayesiano; cáncer de mama; entrenamiento; ***detección*** *automatizada de enfermedades*

Bayesian Predictive System for Detection of Breast Cancer

Abstract

We propose a predictive method to detect breast cancer, based on the following variables: Age, weight, height, body mass index, schooling, socioeconomic stratum, social security, smoker, when quit smoking, passive smoker, consumption of liquor, quantity of liquor, family inheritance of cancer, age of menarche, menopause, pregnancies, age of first birth, breastfeeding, consumption of oral contraceptives, how many years of oral contraceptive use, oral contraceptive suspension time, hormone replacement therapy, and the presence of the GSTM1 gene. Taking as reference patients from the central region of Colombia (Caldas), two databases were defined, one of people without cancer and another of people with cancer. The same training database was used for testing. The proposed methodology defines and trains a Bayesian classification system, with a database of patients with cancer and without cancer. Subsequently, a system validation is performed in order to determine the number of successes and errors in the recognition of this disease. As a result, a 100% success rate is achieved.

Keywords: bayesian classifier; breast cancer; training; automated disease detection

INTRODUCCIÓN

El cáncer de mama es el cáncer invasivo más común que afecta a las mujeres de todo el mundo. El desarrollo de métodos de detección han aumentado la incidencia, pero la mortalidad ha disminuido de manera constante; y aun así, esta patología es la segunda causa de muerte por cáncer en las mujeres (Ban et al, 2014). Se estima, que en el mundo, aproximadamente se diagnostican al año 1.4 millones de casos nuevos de cáncer de mama, de los cuales al menos 458.000 les ocasiona la muerte (Ban et al, 2014). En Colombia, las tasas de incidencia y de mortalidad de cáncer de mama durante el periodo del 2007-2011, fueron ascendiendo según el grupo etario considerado, (Ver Tabla 1); igual comportamiento se refleja en el departamento de Caldas (Instituto Nacional de cancerología, 2017).

La etiología del cáncer de mama es muy heterogénea. Se registra mayor incidencia en la raza blanca, pero con menor tasa de mortalidad y mejor sobrevida a cinco años; mientras que en la raza afrodescendiente es un poco menor la incidencia, pero hay mayor tasa de mortalidad y menor sobrevida a los cinco años (Ban et al, 2014). Los aspectos que contribuyen al desarrollo del cáncer de mama, son muy heterogéneos; entre ellos se encuentran el estrato socioeconómico, el acceso a los servicios de salud; la disparidad en estos aspectos incide en un oportuno diagnóstico, lo cual es de vital importancia para poder iniciar tempranamente el tratamiento. Entre los principales factores influyentes se encuentran:

Factores reproductivos: La edad, a mayor edad hay mayor riesgo a desarrollar esta neoplasia, debido a la tendencia de los receptores de estrógenos positivos, cuya incidencia aumenta con la edad y es más común en mujeres postmenopaúsicas (Kapil et al, 2014; Parkin, 2011; Syamala et al, 2008) – La edad de la menarquia, si ocurre tempranamente, antes de los doce años, la mujer va a tener un tramo de vida con mayor exposición a hormonas endógenas, en este caso a estrógenos y a mayor estimulación hormonal durante más tiempo y este evento se ha asociado con el desarrollo de cáncer de mama. – La edad del primer embarazo a término, se ha asociado una edad menor del primer embarazo a término (20 años), como un factor protector del desarrollo a cáncer de mama; si el primer embarazo es a una edad de 35 años o mayor, el riesgo de desarrollar cáncer de mama aumenta. – La paridad; una mujer multípara presenta menos riesgo, con el tiempo, a desarrollar cáncer de mama, que una mujer nulípara

A largo plazo, las células epiteliales de mama se diferencian después del primer embarazo, sus ciclos celulares son más largos y por tanto son menos sensibles a los efectos de los agentes cancerígenos y tienen más tiempo para someterse a la reparación del ADN. - La lactancia disminuye el riesgo de desarrollar cáncer de mama, porque induce la diferenciación de los conductos y durante la lactancia se disminuyen los niveles de estrógenos.- El aborto inducido o espontáneo parece aumentar el riesgo a cáncer de mama, pero los estudios que se han realizado no son concluyentes (Guo et al, 2015)- La edad de la menopausia, a mayor edad, mayor riesgo a desarrollar cáncer de mama. Se ha establecido que una ooferoctomía bilateral, ocasiona una menopausia artificial y reduce dramáticamente el riesgo a cáncer de mama.- Exposición a hormonas exógenas,(anticonceptivos) aumenta ligeramente el riesgo a cáncer de mama.- El uso de terapia hormonal después de la menopausia, incrementa el riesgo a desarrollar cáncer de mama, dependiente de la dosis y del tiempo.

Tabla 1: Incidencia de Cáncer de mama en Colombia y Caldas; periodo 2007-2011 TEI es la tasa específica de incidencia por 100.000 habitantes y TEM es la tasa específica de mortalidad por 100.000 habitantes. Elaborada a partir de las siguientes fuentes. (López et al 2012; Instituto Nacional de Cancerología, 2016; López-Guarnizo et al 2012)

Colombia					Caldas			
Grupo/años	Incidencia	TEI	Mortalidad	TEM	Incidencia	TEI	Mortalidad	TEM
15-44	1394	13.2	327	3.1	29	13	7	3.1
45-54	1988	78	531	20.8	58	92.2	15	23.9
55-64	2000	123.5	548	33.8	63	145.5	17	39.9
65 +	2245	138.6	820	50.6	71	158.1	26	57.9

Factores genéticos: Se registra mayor incidencia en la raza blanca, pero con menor tasa de mortalidad y mejor sobrevida a cinco años; mientras que en la raza afrodescendiente es un poco menor la incidencia, pero hay mayor tasa de mortalidad y menor sobrevida a los cinco años (Ban et al, 2014). Las mujeres hispanas presentan menor incidencia y menor mortalidad; lo anterior sugiere un compromiso de factores genéticos, ambientales y sociales. Pero ya al considerar los grupos familiares se aprecia un compromiso genético; la historia familiar de cáncer de mama ha sido bien documentada como un factor de riesgo; una mujer cuya madre o hermana presentan la patología, tiene el doble del riesgo, con respecto a la población general, de padecer la neoplasia de mama. Además, la predisposición genética incluye un comienzo temprano de la enfermedad y en muchas ocasiones más agresiva; dependiendo de si los genes sean de alto riesgo como mutaciones en *BRCA1, BRCA2, P53, PTEN* o de bajo riesgo como los que participan en los mecanismos de

activación (CYPs) y detoxificación de xenobióticos (GSTs). Los de altos riesgo son poco frecuente y los de bajo riesgo son muy frecuentes en la población. En la característica histológica de la neoplasia, puede ocurrir que lesiones benignas se vuelvan malignas; se ha encontrado un ligero incremento de riesgo del cáncer de mama, cuando se presentan adenomas, fibroadenomas, o papiloma intraductal (Yoon et al, 2013; Sakoda et al 2008).

Factores relacionados con el estilo de vida: El elevado consumo de alcohol incrementa hasta en un 32% el riesgo de cáncer de mama. El consumo de tabaquismo, sobre todo desde edad temprana, elevadas la dosis y por mucho tiempo, se asocian con el desarrollo del cáncer de mama; y es de anotar que muchas de las mujeres que consumen alcohol también fuman (Knight et al, 2017; Van Emburgh et al, 2008) La actividad física regular, sobre todo en mujeres adultas pre menopaúsicas, disminuye el riesgo a desarrollar el cáncer de mama. Respecto a las dietas, hay numerosos factores en las dietas que contienen compuestos potencialmente protectores, como las isoflavonas (fitoestrogenos) que abundan en la soya; otros con propiedades antioxidantes como las vitaminas (A,C,E y beta-carotenos). El peso corporal; el efecto de la obesidad depende del estatus menopausico. Un índice de masa corporal elevado después de la menopausia está asociado con un mayor riesgo de presentar cáncer de mama. La exposición a la radiación aumenta el riesgo a desarrollar el cáncer de mama; incluso la mamografía, y por ello no es aconsejable realizar dicho procedimiento en mujeres jóvenes.

Desafortunadamente, es difícil evaluar el verdadero efecto de cada uno de los factores de riesgo nombrados, sobre el desarrollo del cáncer de mama. No obstante, es necesario implementar modelos que permitan considerar la carga que aportan las diferentes combinaciones de estos factores con el riesgo a desarrollar el cáncer de mama, y poder hacer predicciones más ajustadas a la realidad y permitir detectar a tiempo las personas con alto riesgo, para incluirlas en un programa de vigilancia epidemiológica. Estas detecciones tempranas quizás se puedan hacer con modelos bayesianos ajustados a las mediciones de las características de interés. Una revisión de las diversas contribuciones en los últimos 20 años, muestra que, si bien se han diseñado algunos sistemas de detección de cáncer de mama, basados en técnicas de inteligencia artificial, estas no han sido exploradas en su totalidad. En la Tabla 2 se relacionan algunos estudios publicados en los últimos 20 años.

Tabla 2: Contribuciones sobre sistemas inteligentes aplicados a la detección de cáncer de mama.

Autores	*Contribución*
(Wang et al, 1999)	Mejora la detección del cáncer de mama a partir de la mamografía, en un porcentaje superior al 80%, al emplear una red bayesiana simple, en vez de combinaciones hibridas de redes independientes, en la cual se integra la imagen y las características que no son imagen.
(Abbass, 2002)	Diseña una red artificial para la predicción del cáncer de mama. Este algoritmo trabaja sobre la base datos de Wisconsin. Los resultados son comparados frente a una programación evolucionaria y programación hacia atrás. Este sistema logra un porcentaje de aciertos en la clasificación del 98.1%.
(Chou et al, 2004)	Propone un modelo de clasificación de datos para identificar el cáncer de mama, integrando un modelo de regresión adaptativa a una red neuronal, obteniéndose resultados del 99.7% de efectividad en la detección de esta enfermedad.
(Sahan et al, 2007)	Propone un sistema inmune basado en una máquina de aprendizaje para la detección del cáncer de mama. Obteniéndose clasificaciones efectivas del 99.14%.
(Cruz-Ramirez et al, 2007)	Evalúa diferentes clasificadores bayesianos en el diagnóstico del cáncer de mama. Los resultados son comprados frente a datos recolectados por simples observadores y múltiples observadores. Obteniéndose resultados del 93.04% y 83.31% respectivamente.
(Abbod et al, 2007)	En este trabajo hacen una revisión de las principales técnicas inteligentes para el manejo del cáncer urológico: Próstata, vejiga, riñón, testículo. Esto sistemas presentan una efectividad entre el 75% y 100%, según la técnica y el tipo de cáncer.
Chakraborty, S. 2009a)	Propone un modelo bayesiano de dos clases basado en un núcleo, semi paramétrico bayesiano. El problema es simplificado reduciendo la dimensionalidad de sus variables. El modelo es probado en 3 diferentes tipos de cáncer (Leucemia, próstata y colon) con muy buenos resultados.
Chakraborty, S. 2009b).	Establece un modelo bayesiano integrado a una técnica de selección de datos, para identificar diferentes clases de cáncer: Mama, Colon, Leucemia y Glioma. El modelo es probado con 100.000 iteraciones, obteniéndose muy bajos errores en la clasificación (cerca al 1%)
(Cruz-Ramírez et al, 2009)	Evalúa dos árboles de decisión y cuatro redes bayesianas (con el apoyo del programa Weka) en el diagnóstico del cáncer de mama. Se emplearon bases de datos colectadas por simples observadores y múltiples observadores. Encontrándose diferencias en ambos resultados.
(Catto et al, 2010)	Logra buenos resultados en el diseño de un sistema inteligente basado en lógica difusa y redes neuronales, para la identificación del cáncer de vejiga
(Cedeño et al 2011)	Presenta un nuevo método en el entrenamiento de las redes neuronales, en el cual se da prioridad a la actualización de los pesos, variando las activaciones. Encontrándose resultados de clasificación del 99.26%

Tabla 2 (continuación)

Autores	*Contribución*
(Keleş et al, 2011)	Desarrolla un sistema experto para la detección del cáncer de mama, a partir de la mamografía. Con un 96% efectividad en predicciones positivas y 81% predicciones negativas.
(Kalderstam et al, 2013)	Construye una red neuronal para la detección del cáncer de mama la cual es entrenada por medio de un algoritmo genético con la mitad de los datos, con resultados superiores al 90%.
(López et al, 2013)	En este trabajo se desarrolla un sistema para la identificación automática del tamaño de los tumores de cáncer de mama. Esta metodología se basa en los algoritmos: J48, LADtree, NaiveBayes. Obteniéndose resultados superiores al 96%.
(Dheeba et al, 2014)	Emplea un algoritmo basado partículas inteligentes optimizadas por la red neuronal de wavelet, en la detección del cáncer de mama a partir de mamografías. Con resultados superiores al 94%.
(Kim et al, 2014)	Se realiza un algoritmo de aprendizaje evolucionario en la detección del cáncer de mama y mieloma múltiple, el cual es optimizado por medio de un secuencia bayesiana simple, con muy buenos resultados
(Calderon et al, 2014)	Se propone una metodología para la detección de cáncer de mama, con frecuencias no ionizantes.
(Karabatak et al, 2015)	Emplea un clasificador bayesiano ponderado en la detección del cáncer de mama. con resultados superiores al 98%
(Liu et al, 2015)	Propone un sistema de detección de cáncer de mama sobre las mamografías. Este sistema emplea una técnica basada en máquinas de soporte vectorial (SVM). Obteniéndose solo un 5.3% de fallos en la detección.
(Papageorgiou et al 2015)	Desarrolla un soporte de decisión medico basado en lógica difusa, con un 95% en el diagnóstico del cáncer de mama.
(Magna et al, 2016)	Diseña un sistema basado en un conjunto adaptativo de redes inmunes para detectar el cáncer de mama partiendo de una mamografía. Como resultado se obtiene un nivel de precisión máxima de 0,90, la sensibilidad de 0,93 y especificidad de 0,87.
(Sheikhpou et al, 2016)	Se propone un modelo basado en partículas inteligentes para clasificar y distinguir si un tumor de cáncer de mama es maligno o no. Con este sistema se encuentran resultados correctos de clasificación del 98.53%.
(Herrera et al, 2016)	Se propone un sistema automatizado, basado en máquinas de aprendizaje y el procesamiento de imágenes, para detectar el grado de malignidad de un tumor de cáncer de mama, obteniéndose muy buenos resultados.
(Kozegar et al, 2017)	Propone un sistema de detección de cáncer de mama basado en ultrasonidos. Mediante un clasificador se delimitan los límites de los tumores.
(Van Zelst et al, 2017)	Analiza el software para la detección de cáncer de mama por medio de ultrasonidos.

Como se observa en la Tabla 2, predomina el uso de técnicas inteligentes como redes neuronales, algoritmos evolutivos, máquinas de soporte vectorial, sistemas difusos, entre otros. Sin embargo el uso de clasificadores bayesianos, no ha sido muy usado en la detección del cáncer de mama y otros tipos de cáncer encontrándose pocos trabajos entre otros (Wang et al, 2014). Así mismo, es importante resaltar que los trabajos encontrados en la literatura, toman como base el estudio de exámenes especializados como análisis de mamografías, y otro tipo de datos. No obstante, en este documento se propone un sistema de diagnóstico de cáncer de mama tomando como referencia las variables: edad, peso, talla, índice de masa corporal, escolaridad, estrato socioeconómico, seguridad social, fumador, cuando dejo de fumar, fumador pasivo, consume licor, cantidad de licor, herencia familiar de cáncer, edad de la menarca, menopausia, embarazos, partos, edad del primer parto, lactancia, consumo de anticonceptivos orales, cuanto años consumió anticonceptivos orales, tiempo de suspensión de anticonceptivos orales, terapia de reemplazo hormonal y presencia del gen GSTM1.

El modelo propuesto emplea un algoritmo evolutivo, con el fin de realizar una selección efectiva de las variables descritas. Una vez seleccionadas las variables, se define un modelo bayesiano basado en dos funciones, una para las muestras de personas con cáncer y otra para las muestras de personas sin cáncer. Definidas las funciones, los nuevos registros de las personas son evaluados en las dos funciones anteriores, suponiéndose que el registro pertenece a la clase cuya función presente el máximo valor. Los resultados son comparados contra test especializados, con el fin de establecer la efectividad del sistema. Finalmente se encuentra que a partir de las diversas variables empleadas en la construcción de este sistema, se logran porcentajes de aciertos en la detección del cáncer iguales al 100%.

MATERIALES Y MÉTODOS

La metodología desarrollada, para la detección del cáncer de mama, comprende los siguientes pasos metodológicos: Paso 1. Bases de datos; Paso 2. Formalización matemática; y Paso 3. Algoritmo de Selección y clasificación.

Paso 1. Bases de datos: Tomando como referencias pacientes de la región central de Colombia –Caldas (Con su respectivo consentimiento), se definieron dos bases de datos, una de personas sin cáncer y otra de personas con cáncer. La misma base de datos de entrenamiento fue empleada para probar la metodología. Las columnas de la Tabla 3, representan las características de las bases datos definidas:.

Tabla 3: Características de la base de datos

Paciente	*Características*													*¿Cáncer?*
	C1	C2	C3	…	…	…	..	…	…	…	C22	C23	C24	
1														
.														
N														

En esta tabla, C1 = Edad, C2 = peso, C3 = talla, C4 = índice de masa corporal, C5 = escolaridad, C6 = estrato, C7 = seguridad social, C8 = fumador, C9 = cuando dejo de fumar ?, C10 = fumador pasivo ?, C11 = consume licor ?, C12 = cantidad de licor ?, C13 = herencia familiar de cáncer, C14 = merca, C15 = menopausia, C16 = embarazos, C17 = partos, C18 = edad del primer parto, C19 = lactancia, C20 = consumo de anticonceptivos orales, C21 = cuanto años consumió anticonceptivos orales, C22 = tiempo de suspensión de anticonceptivos orales, C23 = terapia de reemplazo hormonal, C24 = presencia del gen GSTM1.

Una vez creadas las bases de datos todos los valores no numéricos de los registros fueron transformados en números con el fin de poder definir los clasificadores bayesianos. Esta transformación se realizó según el siguiente proceso: a) Los valores de respuesta sí o no fueron transformados en 1 ó 0 respectivamente. b) Las diferentes respuestas para los valores de escolaridad (C5) fueron transformados en valores entre 1 y 7, (1 mínimo grado de escolaridad, 7 máximo grado de escolaridad). c) Los valores de seguridad social (C7) fueron transformados en valores entre 1 y 4, según las respuestas (1 el mínimo, 4 el máximo).

Paso 2. Formalización matemática: La definición matemática del clasificador bayesiano es representada en los siguientes conjuntos de ecuaciones: a) Una híper matriz de tres dimensiones, es definida. Las columnas representan el conjunto de características (C1:C24) ilustradas en la Tabla 3. Así mismo, las filas representan el número de pacientes en cada base de datos que serán objeto de análisis. La tercera dimensión de esta matriz, representará, el número de clases, en este caso 2 (¿tiene Cáncer?: Si o No). b)
Se calcula el vector de medias y la matriz de covarianza de dónde *i* representa la dimensionalidad de cada una de las matrices expresadas en las ecuaciones 1-4 y *K* es una constante definida a partir de la matriz *Xi. c)* Para cada clase se define una función de probabilidad según el sistema de ecuaciones propuesto por Duda, Hart y Stork (2001: pp 41):

$$\mathrm{P_i(x)} = \mathrm{X^t W_i X} + w^t{}_\mathrm{i}\mathrm{X} + \mathrm{w_{io}} \tag{1}$$

Donde,

$$W_i = -\frac{1}{2}\sum_i^{-1} \tag{2}$$

$$w_i = \sum_i^{-1} u_i \tag{3}$$

$$w_{io} = -\frac{1}{2} u_i^t \sum_i^{-1} u_i \; - \frac{1}{2}\ln(\,|\textstyle\sum_i|\,) \quad + \ln(k) \tag{4}$$

Paso 3. Algoritmo de Selección y clasificación. El algoritmo propuesto está conformado por los siguientes pasos: a) Selección del tamaño de los padres (Número de características). Este proceso selecciona un numero K (K=1...24). Este número definirá las características que serán analizadas. K, define el tamaño de los padres iniciales. b) Población Inicial. El tamaño P establecido en el paso anterior, define el tamaño de la población inicial. Esta se hace mediante un vector de tamaño K, el cual contiene un conjunto de números aleatorios, (K=1...24). Estos números definirán las características que serán objeto de análisis. c) Operadores Genéticos y Fitness. Con los operadores de mutación (3%) y combinación (97%) se definen los nuevos hijos. Con cada uno de los individuos definidos (Padres e Hijos), Se definen las funciones bayesianas establecidas por la ecuación 1. Dos funciones son definidas, (personas con cáncer y sin cáncer). Una vez definidas las funciones según la ecuación 1, cada registro de la base de datos, establecida para la validación, son probadas en estas funciones, suponiéndose que el registro de datos pertenece aquella función cuyo resultado sea

mayor. Los resultados establecidos son comparados con los test especializados. El porcentaje de aciertos positivos, será el resultado de la función Fitness d). Condiciones de parada. El procedimiento anterior, se repite hasta que se encuentre un número determinado de iteraciones, sin que se haya logrado mejorar el porcentaje de aciertos positivos. e) Efectividad del sistema. Se definirá como el porcentaje de aciertos positivos, en la clasificación de la enfermedad. Considerando la revisión literaria encontrada, se deberá permitir la evolución del algoritmo hasta que el porcentaje de aciertos sea por lo menos al 95%.

RESULTADOS Y DISCUSIONES

Como resultado de aplicar la metodología en las bases de datos definidas, se obtienen los resultados que se presentan en lo que sigue:

Paso 1.Bases de Datos. En esta parte del trabajo, se definieron dos archivos, uno para las personas sin cáncer y otro para las personas con cáncer.

Tabla 4: Datos Generales.

Nombre	*Definición*	*Número de registros*
Sin Cáncer	Persona sin Cáncer –Entrenamiento y Validación.	4
Con Cáncer	Personas con Cáncer –Entrenamiento y Validación.	44

Posteriormente, siguiendo con lo estipulado en la metodología todos los valores no numéricos fueron trasformados en números, con el fin de poder construir los clasificadores bayesianos. Las características generales de estos archivos y sus estadísticas básicas son definidas en las Tablas 5, 6, 7, 8, 9,10 respectivamente para las personas con y sin cáncer.

Tabla 5: Estadísticas básicas. Personas sin Cáncer (Características C_1-C_8)

Parámetro	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Promedio	57,50	57,75	1,60	22,48	3,25	3,25	1,25	0,25
Desviación	15,00	11,09	0,10	3,11	1,71	0,96	0,50	0,50
Mínimo	37,00	46,00	1,48	18,90	1,00	2,00	1,00	0,00
Máximo	73,00	72,00	1,69	25,82	5,00	4,00	2,00	1,00
Mediana	60,00	56,50	1,62	22,60	3,50	3,50	1,00	0,00
Varianza	225,00	122,92	0,01	9,68	2,92	0,92	0,25	0,25

Tabla 6: Estadísticas básicas. Personas sin Cáncer (Características C_9-C_{16})

Parámetro	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
Promedio	2,50	0,50	0,50	1,00	0,25	14,00	32,50	2,00
Desviación	5,00	0,58	0,58	1,15	0,50	0,82	21,89	2,45
Mínimo	0,00	0,00	0,00	0,00	0,00	13,00	0,00	0,00
Máximo	10,00	1,00	1,00	2,00	1,00	15,00	46,00	5,00
Mediana	0,00	0,50	0,50	1,00	0,00	14,00	42,00	1,50
Varianza	25,00	0,33	0,33	1,33	0,25	0,67	479,00	6,00

Tabla 7: Estadísticas básicas. Personas sin Cáncer (Características C_{17}-C_{24})

Parámetro	C_{17}	C_{18}	C_{19}	C_{20}	C_{21}	C_{22}	C_{23}	C_{24}
Promedio	1,25	13,25	7,25	0,25	1,00	4,50	15,00	0,00
Desviación	1,50	15,31	9,91	0,50	2,00	9,00	30,00	0,00
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Máximo	3,00	27,00	21,00	1,00	4,00	18,00	60,00	0,00
Mediana	1,00	13,00	4,00	0,00	0,00	0,00	0,00	0,00
Varianza	2,25	234,25	98,25	0,25	4,00	81,00	900,00	0,00

Tabla 8: Estadísticas básicas. Personas con Cáncer (Características C_1-C_8)

Parámetro	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Promedio	56,23	59,41	1,55	24,75	3,87	3,05	1,44	0,36
Desviación	13,00	8,21	0,05	3,14	2,09	1,32	0,88	0,49
Mínimo	33,00	45,00	1,46	18,49	1,00	1,00	1,00	0,00
Máximo	82,00	76,00	1,68	31,63	7,00	6,00	4,00	1,00
Mediana	56,00	59,00	1,55	24,89	4,00	3,00	1,00	0,00
Varianza	169,13	67,41	0,00	9,84	4,38	1,73	0,78	0,24

Tabla 9: Estadísticas básicas. Personas con Cáncer (Características C_9-C_{16})

Parámetro	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
Promedio	4,59	0,23	0,33	1,38	0,21	12,54	38,67	2,97
Desviación	10,59	0,43	0,48	2,25	0,41	2,57	18,79	2,56
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Máximo	50,00	1,00	1,00	6,00	1,00	17,00	56,00	11,00
Mediana	0,00	0,00	0,00	0,00	0,00	13,00	48,00	3,00
Varianza	112,25	0,18	0,23	5,09	0,17	6,62	353,12	6,55

Tabla 10: Estadísticas básicas. Personas con Cáncer (Características C_{17}-C_{24})

Parámetro	C_{17}	C_{18}	C_{19}	C_{20}	C_{21}	C_{22}	C_{23}	C_{24}
Promedio	2,46	20,05	10,21	0,18	1,26	1,62	2,85	0,69
Desviación	2,32	10,22	12,56	0,39	3,35	5,07	10,51	0,47
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Máximo	11,00	38,00	48,00	1,00	16,00	20,00	60,00	1,00
Mediana	2,00	21,00	5,00	0,00	0,00	0,00	0,00	1,00
Varianza	5,36	104,52	157,80	0,15	11,20	25,66	110,55	0,22

Paso 2. Formalización matemática: Como resultado de esta proceso se definieron cada una de las funciones matemáticas para las dos clases que deben ser identificadas (personas sin cáncer y personas con cáncer). Estas funciones son ilustradas en las ecuaciones 5 y 6 respectivamente:

$$\mathrm{P_{normal}(x)} = \mathrm{X}^{t}_{(1xk)}\mathrm{W_{i\_sincancer(kxk)}}X_{(kx1)} + w^{t}{}_{i\_sincancer(1xk)}\ \ X_{(kx1)} + \mathrm{w}_{io\_sincancer} \quad (5)$$

$$\mathrm{P_{Cancer}(x)} = \mathrm{X}^{t}_{(1xk)}\mathrm{W_{i\_concancer(kxk)}}X_{(kx1)} + w^{t}{}_{i\_concancer(1xk)}\ \ X_{(kx1)} + \mathrm{w}_{io\_cancer} \quad (6)$$

Donde, las matrices $W_{i\_sincancer}$, $w_{i\_sincancer}$ y la constante $W_{io\_sincancer}$ y las matrices $W_{i\_concancer}$, $w_{i\_concancer}$ y la constante $W_{io\_concancer,}$ son definidas tomando como referencia las ecuaciones 2, 3, 4 y el conjunto de datos de entrenamiento establecidos para los pacientes sin cáncer y con cáncer. Adicionalmente, la dimensionalidad de cada una de las matrices empleadas en las ecuaciones 6 y 7, es definida en el paréntesis anexo al lado de cada matriz *X, W y w* como subíndice; esto es: (1xk), (kxk) y (kx1).

Paso 3. Algoritmo de Selección y clasificación. Las diferentes ecuaciones de entrenamiento derivadas de las ecuaciones 5 y 6 son definidas para cada uno de los posibles subgrupos de tamaño *k* (Donde, *k* varía entre 1 y 24). Estas características son establecidas por medio del algoritmo evolutivo explicado en el paso 3 de la metodología. En la definición de la función bayesiana de las personas sin cáncer no fue incluida la característica C24, en la definición de la función bayesiana de las personas con cáncer si fue incluida esta característica. Esto último con el fin de determinar de manera particular, la incidencia de esta característica en el reconocimiento de la enfermedad.

Las mismas características indicadas por cada subgrupo de tamaño k, usadas en la fase de entrenamiento, son empleadas para seleccionar cada una de las columnas de la matriz que representa el conjunto de datos de validación, para las personas con cáncer y sin cáncer. Estos datos seleccionados, permitirán resolver las ecuaciones 6 y 7. Lo cual dará como resultado un número real. Se supondrá, que el paciente pertenece a la clase cuyo resultado es mayor. Esta clasificación es comparada, para cada paciente, con los resultados de exámenes especializados. El porcentaje de aciertos será el valor de la función fitness del algoritmo evolutivo. La evolución de este algoritmo se hizo durante 100.000 iteraciones. La Tabla 11, ilustra los principales

resultados obtenidos, que permitieron obtener una clasificación del 100%. En las tablas que siguen, C1=Edad, C2=peso, C3=talla, C4=índice de masa corporal, C5=escolaridad, C6=estrato, C7=seguridad social, C8=fumador, C9=cuando dejo de fumar?, C10=fumador pasivo , C11=consume licor ?, C12=cantidad de licor?, C13= herencia familiar de cáncer, C14=merca, C15=menopausia, C16=embarazos, C17=partos, C18=edad del primer parto, C19=lactancia, C20=consumo de anticonceptivos orales, C21=cuanto años consumió anticonceptivos orales, C22=tiempo de suspensión de anticonceptivos orales, C23=terapia de reemplazo hormonal, C24= presencia del gen GSTM1.

En total, se obtuvieron 216 grupos de 5 características, los cuales permitieron una clasificación de los datos al 100%. A continuación, se ilustran las características más empleadas en referencia a los 216 grupos de 5 características que permitieron una clasificación de los datos al 100%: GSTM 1 (100%), Merca (45.83%), IMC (37.03%), Edad (32.8%), Lactancia (22.22%), Escolaridad (20.37%), Edad del primer parto(20.37%), Talla(19.9%), Sesofumar(17.59%), Menopausa(16.37%), Fumador(16.2%), Cantidad licor(16.2%), Estrato(15.74%), Gravidez(14.35%), ACOSUP(13.88%), Peso(12.5%), ACOano(9.7%), TRHmes(9.7%), ACO(7.87%), SS(5.5%), HFCAM(5.5%). No obstante, a pesar de que la características (C24 = GSTM1) está en todos los grupos, la misma por sí sola, no es suficiente para identificar el cáncer de mama. Se requiere de la combinación de otras características.

Tabla 11: Grupos de 5 Características que permitieron una clasificación del 100% (a)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
1								1						1	1								1
1			1												1			1					1
		1	1	1												1							1
													1	1	1	1							1
1											1				1					1			1
			1						1				1				1						1
					1								1					1	1				1
			1	1					1			1											1
		1			1			1						1									1
1					1		1						1										1
			1	1					1									1					1
1								1					1				1						1
1											1		1	1									1
			1	1				1														1	1
		1	1							1			1										1
		1			1							1					1						1
1								1					1			1							1
		1	1						1				1										1
											1		1			1		1					1
		1						1		1								1					1
1											1				1						1		1
			1		1										1		1						1
1								1					1	1									1
1			1										1			1							1
1							1						1	1									1
								1					1	1	1								1
			1	1					1				1										1
1		1	1								1												1
1			1		1				1														1
1					1								1		1								1
1								1					1								1		1
1										1			1	1									1
			1		1							1			1								1

Tabla 11: Grupos de 5 Características que permitieron una clasificación del 100% (b)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
1								1					1						1				1
1													1	1					1				1
		1	1								1						1						1
1														1	1		1						1
			1	1						1			1										1
		1			1													1		1			1
1			1								1	1											1
1	1		1														1						1
		1	1		1						1												1
1			1								1		1										1
			1	1					1					1									1
			1	1					1						1								1
			1											1		1		1					1
									1					1			1				1		1
			1					1			1											1	1
	1				1											1					1		1
1									1						1						1		1
		1											1					1		1			1
1												1	1			1							1
1										1					1				1				1
1				1			1											1					1
										1			1		1			1					1
1								1					1		1								1
1													1	1							1		1
		1			1								1			1							1
			1	1							1		1										1
			1					1	1												1		1
									1					1			1		1				1
												1	1				1					1	1
												1	1				1				1		1
										1			1			1		1					1
		1		1					1								1						1
1							1						1				1						1
1								1					1									1	1
	1			1						1										1			1
			1	1						1								1					1
1		1	1						1														1
1			1													1		1					1
1							1							1			1						1
			1							1			1					1					1
						1							1					1		1			1
	1			1							1									1			1
	1							1					1					1					1
1													1	1		1							1
1		1											1				1						1
		1			1												1			1			1
1										1					1					1			1
	1			1					1											1			1
				1			1							1				1					1
					1	1							1				1						1
1			1			1			1														1
1			1										1		1								1
			1					1			1									1			1
		1	1								1		1										1
			1	1					1							1							1

Tabla 11: Grupos de 5 Características que permitieron una clasificación del 100% (c)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
1				1				1										1					1
		1			1												1	1					1
1											1		1									1	1
1	1						1											1					1
				1				1						1				1					1
1	1												1					1					1
	1						1		1												1		1
		1											1				1		1				1
1													1	1	1								1
1													1		1			1					1
			1						1				1					1					1
									1				1			1		1					1
		1	1	1					1														1
			1	1							1								1				1
					1							1			1							1	1
	1			1					1												1		1
		1			1								1				1						1
1				1									1					1					1
		1			1												1				1		1
	1			1							1										1		1
								1					1	1			1						1
1													1		1		1						1
			1					1	1													1	1
1										1					1						1		1
		1		1						1							1						1
1			1	1					1														1
1										1			1										1
		1			1													1			1		1
1													1	1								1	1
			1					1			1										1		1
1			1							1			1										1
		1	1		1																		1
						1							1				1			1			1
	1				1											1				1			1
1											1				1				1				1
			1				1	1					1										1
		1											1			1							1
		1	1								1							1					1
											1		1		1			1					1
1								1					1							1			1
1							1						1			1							1
			1					1		1										1			1
			1	1				1													1		1
1													1	1						1			1
							1						1	1			1						1
		1		1							1						1						1
1					1								1			1							1
			1								1		1										1
		1					1						1										1
						1							1					1	1				1
			1	1							1						1						1
1	1							1					1										1
		1	1						1								1						1
			1	1			1			1													1

Tabla 11: Grupos de 5 Características que permitieron una clasificación del 100% (d)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
			1	1							1									1			1
		1						1			1							1					1
		1					1			1								1					1
								1					1	1		1							1
		1			1		1							1									1
			1	1						1									1				1
1			1										1	1									1
			1								1		1					1					1
1			1						1								1						1
			1		1	1										1							1
		1			1													1	1				1
1													1	1				1					1
					1								1		1	1							1
	1		1							1												1	1
	1		1				1														1		1
			1					1		1									1				1
			1	1	1								1										1
1							1						1		1								1
1													1			1	1						1
			1	1			1		1														1
	1							1			1										1		1
1													1			1					1		1
		1			1												1					1	1
	1				1								1				1						1
		1											1					1			1		1
			1				1				1										1		1
1								1						1			1						1
													1	1			1						1
			1				1				1									1			1
1													1	1			1						1
	1						1			1											1		1
							1						1	1		1							1
1						1	1						1										1
			1				1				1									1			1
		1											1					1				1	1
	1		1								1							1					1
			1	1						1							1						1
									1				1		1			1					1
					1								1					1		1			1
	1			1			1															1	1
												1	1					1				1	1
			1	1					1								1						1
													1	1	1		1						1
			1	1					1											1			1
					1	1				1					1								1
	1		1	1																	1		1
	1						1						1					1					1
							1						1	1	1								1
		1											1				1				1		1
		1			1	1											1						1
			1				1			1											1		1
	1			1									1									1	1
					1	1				1						1							1

Tabla 11: Grupos de 5 Características que permitieron una clasificación del 100% (e)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
			1				1		1												1		1
1													1			1						1	1
	1							1	1										1				1
		1									1	1		1									1
			1					1			1								1				1
	1			1						1											1		1
			1					1	1										1				1
			1	1			1												1				1
						1							1					1				1	1
			1	1				1												1			1
	1	1													1			1					1
			1					1		1											1		1
												1	1					1			1		1
			1					1		1												1	1
	1		1					1												1			1
												1	1					1		1			1
			1	1							1							1					1
			1				1			1												1	1
1			1						1				1										1

CONCLUSIONES

Este clasificador bayesiano constituye una herramienta de gran utilidad para ayudar al diagnóstico temprano del cáncer de mama. El nivel de acierto del sistema es del 100% logrando esta efectividad con un mínimo de 5 características. Aunque algunas características, están presenten en la mayoría de grupos de 5 características que identifican correctamente la enfermedad, las mismas por si solas no son suficientes para identificar la enfermedad.

AGRADECIMIENTOS

Se agradece la colaboración a la Universidad Nacional de Colombia y en especial al Departamento de Ingeniería Industrial de la sede Manizales. Igualmente, se agradece a la Universidad de Caldas, especialmente al Programa de Medicina y al Departamento de Sistemas e Informática.

REFERENCIAS

Abbass, H. A., *An evolutionary artificial neural networks approach for breast cancer diagnosis*, Artificial Intelligence in Medicine, 25(3), 265–281 (2002)

Abbod, M. F.; J.W.F Catto; D. Linkens; y F.C. Hamdy, *Application of artificial intelligence to the management of urological cancer*, The Journal of Urology, 178(4), 1150–1156 (2007)

Ban K.A.; C.V. Godellas, *Epidemiology of breast cancer*. Surg Oncol Clin N Am,23(3), 409-22 (2014)

Calderón C. M.; H. M. Pérez; A. M. Benavides. y L. J. Morales, *Artículos de desarrollo de un arreglo circular de antenas utilizando herramientas de electromagnetismo computacional*, Inf. Tecnológica, 25(1), 41-54 (2014)

Catto, J. W. F.; M.F. Abbod; P.J. Wild; D. Linkens; C. Pilarsky; I. Rehman. Y F.C. Hamdy, *The application of artificial intelligence to microarray data: identification of a novel gene signature to identify bladder cancer progression*, European Urology, *57*(3), 398–406 (2010)

Cedeño, M.; J. Quintanilla-Domínguez. y D. Andina, *WBCD breast cancer database classification applying artificial metaplasticity neural network*, Expert Systems with Applications, *38*(8) 9573–9579 (2011)

Chakraborty, S, *Simultaneous cancer classification and gene selection with Bayesian nearest neighbor method: An integrated approach*, *Computational Statistics & Data Analysis*, *53*(4), 1462–1474 (2009b)

Chakraborty, S. *Bayesian binary kernel probit model for microarray based cancer classification and gene selection*, *Computational Statistics & Data Analysis*, 53(12), 4198–4209 (2009a)

Chou, S.M.; T.S. Lee; Y.E. Shao. Y I.F. Chen, *Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines*. Expert Systems with Applications, *27*(1), 133- 142 (2004)

Cruz-Ramírez, N.; H.G. Acosta-Mesa; H. Carrillo-Calvet. y R.E. Barrientos-Martínez, *Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks*. Applied Soft Computing, 9(4), 1331–1342 (2009)

Cruz-Ramírez, N.; H.G. Acosta-Mesa; H. Carrillo-Calvet; L.A. Nava-Fernández. y R.E. Barrientos-Martínez, Diagnosis *of breast cancer using Bayesian networks: a case study*, Computers in Biology and Medicine, *37*(11), 1553–64 (2007)

Dheeba, J.; N.A. Singh. y S.T. Selvi, *Computer-aided detection of breast cancer on mammograms : A swarm intelligence optimized wavelet neural network approach*, *Journal of Biomedical Informatics*, *49*, 45 - 52 (2014)

Duda, R.; P.E. Hart y D. Stork, (2001). Patter Classification, (2unda ed.), New York: John Wiley. (2001)

Guo, J.; Y. Huang; L. Yang; Z. Xie; S. Song; J. Yin; L. Kuang; W. Qin, *Association between abortion and breast cancer: an updated systematic review and meta-analysis based on prospective studies*, Cancer Causes Control, 26(6) 811-819 (2015).

Herrera, F.; B. Krawczyk. y M. Galar, *Evolutionary under sampling boosting for imbalanced classification of breast cancer malignancy*, *38*, 714–726 (2016)

Instituto Nacional de Cancerologia ESE, *Cancer en cifras,* recuperado de http://www.cancer.gov.co/cancer_en_cifras, Mayo de 2016. Consultado agosto de 2016.

Kalderstam, J.; P. Edén, P.O. Bendahl; C. Strand; M. Fernö. y M. Ohlsson, *Training artificial neural networks directly on the concordance index for censored data using genetic algorithms*, Artificial Intelligence in Medicine, *58*(2), 125–32 (2013)

Kapil, U.; AS. Bhadoria; N. Sareen; P. Singh; SN. Dwivedi, *Reproductive factors and risk of breast cancer, A Review*, Indian J Cancer. 51(4), 571-576 (2014)

Karabatak, M., (2015). *A new classifier for breast cancer detection based on Naïve Bayesian*. Measurement, *72*, 32–36 (2015)

Keleş, A.; A. Keleş, A. y U. Yavuz, *Expert system based on neuro-fuzzy rules for diagnosis breast cancer*, Expert Systems with Applications, *38*(5), 5719–5726 (2011)

Kim S.J; J.W. Ha. y B.T. Zhang, *Bayesian evolutionary hypergraph learning for predicting cancer clinical outcomes*, Journal of Biomedical Informatics, *49,* 101–111 (2014)

Knight, JA; J. Fan; KE, Malone; EM. John; CF. Lynch; R. Langballe; L. Bernstein; RE. Shore; JD. Brooks; AS. Reiner; M. Woods; X. Liang; JL. Bernstein; *WECARE Study Collaborative Group. Alcohol consumption and cigarette smoking in combination: A predictor of contralateral breast cancer risk in the WECARE study,* Int J Cancer, 1, 141(5) 916-924 (2017)

Kozegar E; M. Soryani; H, Behnam; M. Salamati. y T. Tan., *Breast cancer detection in automated 3D breast ultrasound using iso-contours and cascaded RUSBoosts*, Ultrasonics, 79, 68–80 (2017)

Liu, X. y Z. Zeng, *A new automatic mass detection method for breast cancer with false positive reduction*. Neurocomputing, *152*, 388–402 (2015)

Lopez J.; R. Gozalez; C.L. Parra.; A. Martinez; A. Moreno; J. Peinado; V. Suarez; M. Cabeza; B. Quintana; M. Fernandez; M.J. Ortiz, *Application of artificial intelligence for breast cancer classification before radiatiotherapy*, Reports of Practical Oncology & Radiotherapy,*18*, S391–S392 (2013)

López, G.A.; N.E. Arias, W.A. Arboleda, *Cancer incidence and mortality in Manizales 2003-2007*, Colomb. Med, 43 (4) 281-89 (2012)

López-Guarnizo G; N. Arias-Ortiz, W. Arboleda-Ruiz; D. MorantesArenas, *Registro Poblacional de Manizales - Caldas: Incidencia y Mortalidad 2003-2007*, Informe final de investigación, López-Guarnizo G, ed. Manizales: Universidad de Caldas - Instituto Nacional de Cancerología; 11-24 (2012)

Magna, G.; P. Casti.; S.V. Jayaraman; M. Salmeri; A. Mencattini; E. Martinelli. y C. Natale, *Identification of mammography anomalies for breast cancer detection by an ensemble of classification models based on artificial immune system*, Knowledge-Based Systems, 101, 60–70 (2016)

Papageorgiou, E.I.; J. Subramanian; A. Karmegam, y N. Papandrianos, *A risk management model for familial breast cancer: A new application using Fuzzy Cognitive Map method*, Computer Methods and Programs in Biomedicine, *122*(2) 123–35 (2015)

Parkin, D. *Cancer attributable to exposure to hormones in the UK in 2010*, Br J Cancer, 105(52) 542-548 (2011)

Sahan, S.; K. Polat; H. Kodaz. y S. Güneş, *A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis*, Computers in Biology and Medicine, *37*(3), 415–23 (2007)

Sakoda, L; C. Blackston; K, Xue; J. Doherty; R. Ray; M. Lin; et al. *Glutathione S-transferase M1 and P1 polymorphisms and risk of breast cancer and fibrocystic breast conditions in Chinese women,* Breast Cancer Res Treat, 109(1) 143-155 (2008)

Sheikhpour, R.; M. Agha. y R. Sheikhpour, *Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer*. Applied Soft Computing Journal, *40*, 113–131 (2016)

Syamala, V.; L. Sreeja; V. Syamala; P. Raveendran; R. Balakrishnan; R. Kuttan, et al, *Influence of germline polymorphisms of GSTT1, GSTM1, and GSTP1 in familial versus sporadic breast cancer susceptibility and survival*, Fam Cancer, 7(3) 213-220 (2008)

Van Emburgh, B. J. Hu; E. Levine; L. Mosley; L. Case; H. Lin; et al, *Polymorphisms in drug metabolism genes, smoking, and p53 mutations in breast cancer,* Mol Carcinog, 47(2) 88-99 (2008)

Van Zelts, J.C.; T. Tan; B. Platel.; M. de Jong.; A. Steenbakkers.; M. Mourits.; A. Grivegnee.; C. Borelli.; N. Karssemeijer. y R.M. Mann, *Improved cancer detection in automated breast ultrasound by radiologists using Computer Aided Detection*, European Journal of Radiology, 89, 54–59 (2017)

Wang, K.J.; B. Makond. y K.M. Wang, *Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: a case study of Taiwan*, Computers in Biology and Medicine, *47*, 147–60 (2014)

Wang, X.; B. Zheng; W. Good; J. King y Y. Chang, *Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network*. International Journal of Medical Informatics, *54*(2), 115–126 (1999)

Yoon, JY; D. Chitale, *Adenomyoepithelioma of the breast: a brief diagnostic review,* Arch Pathol Lab Med, 137(5) 725-729 (2013)