COMPARACIÓN DE CLASIFICADORES BASÁNDOSE EN DATOS EXTRAÍDOS DE MAMOGRAMAS DIGITALES

COMPARISON OF CLASSIFIERS BASED ON DATA EXTRACTED FROM DIGITAL MAMMOGRAMS

Miriam Martínez Arroyo
Tecnológico Nacional de México/Instituto Tecnológico de Acapulco
miriamma_ds@hotmail.com

José Antonio Montero Valverde
Tecnológico Nacional de México/Instituto Tecnológico de Acapulco
jamontero1@infinitummail.com

Eduardo de la Cruz Gámez
Tecnológico Nacional de México/Instituto Tecnológico de Acapulco
gamezeduardo@hotmail.com

José Francisco Gazga Portillo
Tecnológico Nacional de México/Instituto Tecnológico de Acapulco
jfgazga@hotmail.com

Resumen

Este trabajo se enfoca al desarrollo de la etapa de clasificación de un sistema para identificar la lesión de distorsión arquitectural en mamogramas, se realiza un análisis comparativo de clasificadores, con la finalidad de determinar el método que mejor se ajuste a los datos extraídos. Las técnicas aplicadas para la comparación de clasificadores son matriz de confusión y matriz costo-beneficio. Se toma en cuenta la sensibilidad al costo del error de clasificación de cada técnica, ya que en muchas situaciones los errores producidos por un modelo predictivo no tienen las mismas consecuencias. Para la realización de las pruebas, se hace uso del *UCI Machine Learning Repository*, donde dos BD contienen datos de historial médico (BD1, BD2) y una contiene datos extraídos de mamogramas digitales (BD3), para esta última se determina que NB obtiene los mejores resultados.

Palabras clave*:* árboles de decisión, clasificador bayesiano simple (NB), distorsión arquitectural, Máquina de Vector de Soporte (VSM), Perceptrón *Multicapa* (MLP).

Abstract

This work focuses on the development of the classification stage of a system to identify the architectural distortion lesion in mammograms, a comparative analysis of classifiers is performed, in order to determine the method that best fits the data extracted. The techniques applied are confusion matrix and cost-benefit matrix. The sensitivity analysis versus the cost of the classification error of each technique is taken into account, since in many situations the errors produced by a predictive model do not have the same consequences. NB, TAN, J48, SVM and MLP are the classifiers used. To carry out the tests, a set of classification domains was selected from the UCI Machine Learning Repository collection, which contains medical history data (BD1, BD2) and data extracted from digital mammograms (BD3), for the latter was determined that NB obtains the best results.

Keywords: *architectural distortion, decision trees, Multi-Layer Perceptron (MLP), Naive Bayes classifier (NB), Support* Vector *Machines (VSM).*

1. Introducción

De acuerdo con datos proporcionados por la Secretaría de Salud en México, el cáncer de mama es la principal causa de mortalidad entre todos los cánceres para las mujeres. Asimismo, indica que una de cada ocho mujeres está propensa a desarrollar esta enfermedad durante su vida [OMS, 2016]. Actualmente el uso de la mamografía de rayos-X es el método más utilizado en las instituciones de salud para detectar esta enfermedad en su etapa temprana. Si el tratamiento es oportuno, el número de muertes por esta causa puede ser reducido al menos en un 30%. Sin embargo, debido a varios factores, tales como la experiencia del radiólogo y la calidad en la imagen observada, hacen que la sensibilidad en el diagnóstico no sea perfecta. En este sentido y con el fin de apoyar el trabajo de los radiólogos se han diseñado sistemas de Diagnóstico Asistidos por Computadora (CAD), los cuales tienen como objetivo, mejorar la calidad de la imagen y mostrar regiones de interés en un mamograma, los que generalmente constan de cuatro etapas básicas [Samulski, 2006]:

- ***Preprocesamiento***. Debido a las limitaciones de las imágenes de mamogramas, ya que éstas se caracterizan por tener un bajo contraste y un alto contenido de información no deseada (p.e. etiquetas de nombre y otros datos), se utilizan técnicas para mejorar la calidad de la imagen, de modo que se realcen características importantes y se delimiten las zonas (se puede eliminar el fondo de la imagen y el músculo pectoral) para la detección y el diagnóstico de lesiones.
- ***Segmentación***. La segmentación es el proceso que subdivide una imagen en sus partes constituyentes u objetos, en esta etapa se extraen los objetos de interés separando la lesión del tejido normal, es decir trata de aislar las regiones sospechosas (regiones de interés –ROI-) del resto de la imagen.
- ***Extracción y selección de características***. Se buscan los descriptores (generalmente morfológicos y de textura) más discriminantes para la distinción entre diferentes tipos de tejido que mejor representen el objeto. Se consideran tres tareas en el proceso de obtención de características: la *extracción*, que hace referencia a la transformación de la señal del dominio primario a un dominio más adecuado para tratar el problema; la *selección*, con la que se escogen aquellas características que aportan información valiosa para la identificación del objeto, desechando la información no discriminante o redundante; y la *reducción* del número de características, que contribuirá a reducir el coste computacional.
- ***Clasificación***. En esta etapa, se realiza la clasificación de imágenes sospechosas a través de un proceso de aprendizaje automatizado que simula el trabajo del experto humano, en el que se toma el vector que representa a la imagen para obtener la lesión más probable. Pueden clasificarse regiones de una imagen en tipo de lesión o clasificar lesiones en benigno y maligno.

Este trabajo se enfoca a esta última etapa, es decir seleccionar el mejor clasificador para identificar la lesión de distorsión arquitectural en Mamogramas.

Esta tarea resulta de gran importancia ya que el uso del clasificador adecuado garantiza en gran medida el resultado emitido en el diagnóstico. Aunque en realidad, aún no existe una regla general acerca de qué métodos de clasificación son los más apropiados para cada tipo de problema, existen diversos estudios realizados que demuestran que esto depende en gran medida de los datos que son utilizados y de las características que presente cada técnica, por lo que esto implica analizar el comportamiento y complejidad de dichas técnicas, tomando en cuenta la capacidad de representación, la legibilidad, la robustez ante ejemplos de entrenamiento con un número variado de instancias, el porcentaje de instancias clasificadas correctamente, el costo de clasificación u otras propiedades deseables.

2. Métodos

Para llegar al objetivo del proyecto se siguió una metodología que consiste de las cuatro fases mostradas en la figura 1.

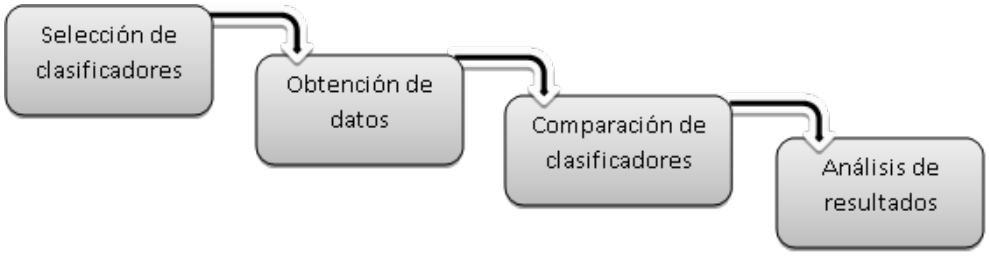


Figura 1 Metodología para evaluación de clasificadores.

Selección de clasificadores

En esta etapa se eligen los algoritmos de clasificación que se utilizan para la realización del objetivo de este trabajo (conocer el clasificador que ofrece mejor desempeño al momento de clasificar datos para el diagnóstico del cáncer de mama). Éstos fueron seleccionados con base a trabajos relacionados con la detección de cáncer de mama [Marin-Castro y Franco-Vázquez, 2017], [Flores, 2016], [Martínez, 2016], [Oommen y otros, 2008], [Argañaraz y Entraigas, 2010], [Barrientos y otros 2008], [Bellaachia y otros, 2010]. Se consideraron: el Naive

Bayes (NB) y su extensión TAN, los Árboles de decisión (J48), Máquina de Vector de Soporte (SVM) y Redes Neuronales Artificiales (RN):

- ***Naive Bayes**:* está basado en el teorema de Bayes, una fórmula que calcula una probabilidad por recuento de la frecuencia de los valores y combinaciones de valores de los datos. Este algoritmo permite la construcción rápida del modelo y altamente escalable. Se escala linealmente con el número de predictores y filas, es muy sencillo, además de que puede ser utilizado para datos binarios y problemas de clasificación multiclase, ha generado buenos resultados en muchos dominios, razones por las que fue elegido para este proyecto [Lozano,2011]. También es utilizada la extensión TAN del Naive Bayes, en la que este clasificador es aumentado a una red.
- ***Árboles de decisión:*** al igual que Naive Bayes, se basa en las probabilidades condicionales, pero a diferencia de éste, los árboles de decisión generan reglas automáticamente, que son las sentencias condicionales que revelan la lógica utilizada para construir el árbol y que pueden ser fácilmente comprendidos y utilizados por las personas dentro de una base de datos, para identificar un conjunto de registros [Quinlan, 1993].
- ***Máquina de Vector de Soporte:*** son un poderoso algoritmo basado en la regresión lineal y no lineal. SVM puede modelar problemas complejos del mundo real, tales como la clasificación de texto e imagen, reconocimiento de escritura a mano, y la bioinformática. Se desempeña bien en conjuntos de datos que tienen muchos atributos, incluso si hay muy pocos casos para entrenar el modelo. La naturaleza de los datos determina en gran parte qué algoritmo de clasificación es la mejor solución a un problema dado. El algoritmo puede diferir con respecto a la precisión y el tiempo de finalización. En la práctica tiene sentido la comparación de varios clasificadores y luego seleccionar el de mejor desempeño [Betancourt, 2005].
- ***Redes Neuronales Artificiales**:* están constituidas por los elementos que se comportan de forma similar a la neurona biológica en sus funciones más

comunes. Estos elementos están organizados de una forma parecida a la que presenta el cerebro humano. En los trabajos e investigaciones revisadas *Multilayer Perceptron* (MLP) presenta buenos resultados [Botia y otros, 2009].

Obtención de los Datos

Para la realización del proceso de comparación se toman tres bases de datos (BDs) del repositorio de máquinas de aprendizaje de la UCI (Universidad de California - Irvine), debido a que se está trabajando a la par con la etapa de extracción de características del proyecto (*"Reconocimiento de la Distorsión Arquitectural en Mamogramas*) y aun no se contaba con los datos disponibles, se optó por esta base de datos, ya que contiene información extraída de mamografías digitales. Estas BDs recogen la información digitalizada de las imágenes tomadas en pacientes con tumores benignos y malignos. El archivo fue creado en 1987 por David Aja y compañeros estudiantes de postgrado de la UCI. Desde entonces, ha sido ampliamente utilizado por los estudiantes, educadores e investigadores de todo el mundo como una fuente primaria de los conjuntos de datos de la máquina de aprendizaje. La versión actual del sitio web fue diseñada en 2007 por Arthur Asuncion y David Newman, este proyecto se realiza en colaboración con Rexa.info en la Universidad de Massachusetts Amherst [Merz y Murphy, 1996].

Comparación de clasificadores

Cada clasificador es ejecutado con las diferentes BDs y evaluado su grado de exactitud. El criterio de evaluación de los clasificadores conocido como proceso de validación, permite efectuar una medición sobre la capacidad de predicción del modelo generado a partir de un clasificador. No existe un consenso en la forma en la cual se deba reportar el desempeño de los algoritmos de detección. Algunos autores reportan el desempeño, simplemente, en función del número de verdaderos positivos (VP) y de falsos positivos (FP). Una métrica usada para reportar el desempeño es la matriz de confusión (mostrada en la tabla 1), que

consiste en verdaderos negativos (VN), verdaderos positivos (VP), falsos positivos (FP) y falsos negativos (FN), por medio de la cual se puede observar la distribución de los errores cometidos por un clasificador a lo largo de las distintas categorías del problema. En dicha matriz se cruza la clase predicha por el clasificador con la clase real [Rodríguez, 2012].

Tabla 1 Matriz de confusión.

		Predicción	
		Clase1	**Clase2**
Clase Real	Clase 1	Verdadero positivo (VP) (diagnóstico positivo enfermedad presente)	Falso negativo (FN) (diagnóstico negativo enfermedad presente)
	Clase 2	Falso positivo (FP) (diagnóstico positivo enfermedad ausente)	Verdadero negativo (VN) (diagnóstico negativo enfermad ausente)

Sin embargo, existen determinadas situaciones en la que los errores no se valoran por igual, a veces se prefiere extraer más instancias asumiendo que alguno de los propuestos no lo sean, por lo que el costo de un falso positivo es menor que el de un falso negativo. Bajo estas circunstancias es posible incorporar los costos para cada tipo de error. De esta forma se puede integrar en una matriz los beneficios y costos donde *Bs* representan beneficios y *Cs* representan costos (tabla 2). La integración de una matriz de confusión y una matriz de costos-beneficios nos brinda las siguientes valoraciones del modelo obtenido (ecuaciones 1 y 2) [Corso, 2009].

$$Beneficio = VP.Bvp + VN.Bvn \quad (1)$$

$$Costo = FP.Cfp + FN.Cfn \quad (2)$$

Tabla 2 Matriz Costo-Beneficio.

		Predicción	
		Clase1	**Clase2**
Clase Real	Clase 1	Beneficio VP	Costo FN
	Clase 2	Costo FP	Beneficio VN

Análisis de resultados

Se observan los resultados obtenidos por los clasificadores, tomando en consideración qué tan costoso resulta uno con respecto a los demás. Los valores que puede arrojar un clasificador para el diagnóstico de enfermedades que resultan más costosos, son los *falsos negativos* (es decir que la enfermedad este presente y sea clasificada como ausente), pues un diagnóstico falso negativo puede traer consecuencias fatales, los *falsos positivos* también tienen un costo pero no tan elevado ya que no conllevan a las mismas consecuencias que un falso negativo, es por eso que para los sistemas *CAD´s (Diagnóstico Asistido por Computadora)* es preferible implementar un clasificador que cueste menos a uno que clasifique mejor. Para fines de conocer el clasificador más costoso se asignaron valores a los errores de clasificación, estos valores están en un rango de 0 a 1. Un falso negativo tiene un costo de 0.8 y un falso positivo tiene un costo de 0.2. Con estos valores podemos construir la matriz de costos y a partir de ella conocer los costos de clasificación.

3. Resultados

Para las pruebas se utilizan tres BDs (referenciadas como: *Wisconsin Prognostic Breast*), extraídas del repositorio de máquinas de aprendizaje de la UCI y se construye cada clasificador con estos datos (las características de cada BDs son descritas en la tabla 3).

Tabla 3 Características de las BDs de prueba.

BD de Prueba	Instancias	Atributos	Clases	Distribución
BD1	286	10	2	201/85
BD2	699	10	2	458/241
BD2	194	34	2	101/93

Para la implementación se hace uso del conjunto de herramientas Weka y el lenguaje Java. El objetivo de realizar la prueba con cada clasificador es conocer la precisión con la que clasifica cada uno de ellos, el número de instancias clasificadas de manera correcta es un elemento clave para que un clasificador sea

elegido. Algo que tiene gran importancia, sobre todo en el diagnóstico médico, es conocer que tan costoso resulta la clasificación, por lo que este aspecto será tomado muy en cuenta durante el proceso.

Para determinar el conjunto de entrenamiento y prueba, se hace uso de la validación cruzada de K iteraciones o *K-fold cross-validation,* donde los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. Las *K iteraciones* realizadas en las 3 pruebas diferentes fueron 10, 30 y 60 *fold*. En la a tabla 4, se muestran los resultados obtenidos en la primera prueba, donde está remarcado el mejor porcentaje promedio de clasificación, que pertenece al clasificador MLP con un 73.83%. La tabla 5 muestra los costos de clasificación para la prueba 1.

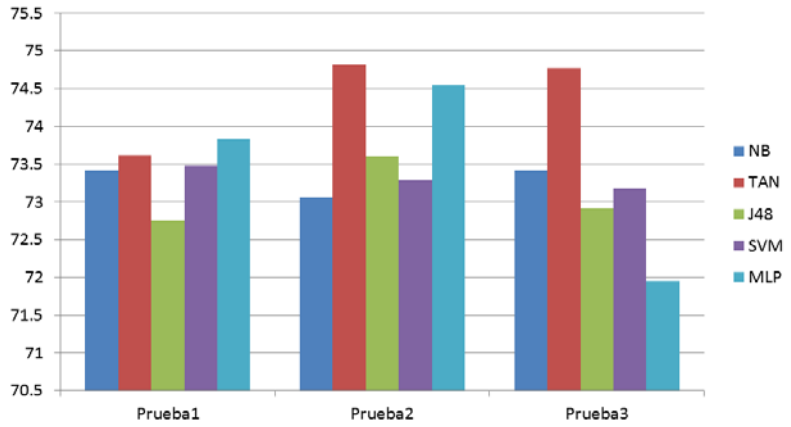
Tabla 4 Concentrado de resultados de la prueba 1 (%).

	NB	**TAN**	**J48**	**SVM**	**MLP**
BD1	72.37	**73.07**	**73.07**	**73.07**	69.58
BD2	95.84	**96.27**	94.69	95.84	94.69
BD3	52.06	51.54	50.51	51.54	**57.21**
Promedio	73.42	73.62	72.75	73.48	**73.83**

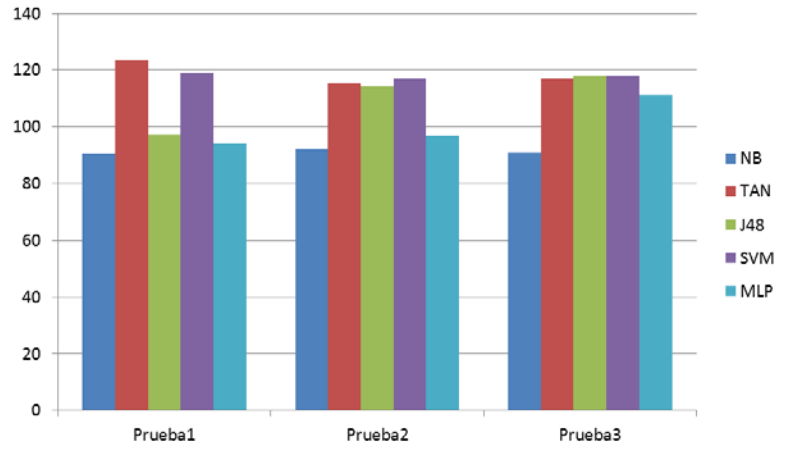
Tabla 5 Costos de clasificación prueba 1.

	NB	**TAN**	**J48**	**SVM**	**MLP**
BD1	34.4	32.2	23.8	**22.6**	31.2
BD2	19	**16**	20	21.4	18.8
BD3	37.2	75.2	53.4	75.2	**42.4**
T. Costo	**90.6**	123.4	97.2	119.2	94.2

En la figura 2a se observa un concentrado de los resultados de las pruebas, donde podemos notar que en la Prueba 1, el clasificador MLP generó la mejor clasificación con un 83% de icc (instancias clasificadas correctamente), en la Prueba 2, TAN obtuvo un 82% de icc y para la Prueba 3 nuevamente TAN logra un mayor número de icc con un 77%. En la figura 2b, se muestra el costo de clasificación de cada clasificador en cada prueba. Los resultados muestran que el clasificador NB tiene una menor suma de costos.



a) Clasificación.



b) Costos.

Figura 2 Resultados.

4. Discusión

Este trabajo se enfocó en la etapa de clasificación correspondiente al proyecto "Reconocimiento de la Distorsión Arquitectural en Mamogramas", se realizó un análisis comparativo de clasificadores para conocer el desempeño y el costo de clasificación de cada uno, determinando el clasificador que mejor se ajuste a los datos que se relacionan al padecimiento de cáncer de mama. Las pruebas consistieron en correr en cada clasificador tres diferentes BDs, de las cuales dos BDs contienen datos de historial médico de pacientes y la tercera contiene datos extraídos de las ROIs de mamografías.

Se utilizó el método de validación cruzada en cada prueba para determinar el número de folds (particiones) para entrenamiento de cada clasificador, variando

10, 30 y 60 para cada prueba (respectivamente) corriendo las 3 bases de datos con los diferentes clasificadores. Para la elección del mejor clasificador (en este caso para la detección del cáncer de mama), se consideraron algunos puntos: el clasificador que genere menor error de clasificación, el clasificador que genere menor costo de clasificación y los datos utilizados. Para el primer punto, la matriz de confusión es el método por medio del cual se realizó la comparación de los clasificadores, tomando en cuenta el número de instancias clasificadas correcta e incorrectamente y los valores que son arrojados directamente por este método (verdaderos positivos y negativos, y falsos positivos y negativos), el TAN obtuvo mejores resultados en dos de las pruebas. Para el segundo punto, el clasificador NB generó el menor costo en cada una de las pruebas. Y con respecto al tercer punto, a las pruebas realizadas directamente con la BD3, se observa que el clasificador NB es el más adecuado para el proyecto, a pesar de que se posicionó como el segundo mejor clasificador en las pruebas, aunque no se determina diferencia significativa entre los otros clasificadores.

5. Pares Revisores

Revisor 1

Nombre:	Cristina Barrera de Jesús
Institución:	Tecnológico Nacional de México/Instituto Tecnológico de San Marcos
Cédula Profesional:	0917015
Área de conocimiento:	Ciencias Computacionales (Procesamiento de Lenguaje Natural
Correo electrónico:	c_barje@hotmail.com
Teléfono:	777 728 9452

Revisor 2

Nombre:	Félix Florentino Álvarez Paliza
Institución:	Universidad Central “Marta Abreu” de Las Villas
Cédula Profesional:	5010/1606680
Área de conocimiento:	Telemática
Correo electrónico:	fapaliza@uclv.edu.cu
Teléfono:	53 5353 3505

6. Bibliografía y Referencias

[1] Argañaraz J. y Entraigas I. (2011). Análisis comparativo entre máquinas de vectores soporte y clasificador de máxima probabilidad para la discriminación de cubiertas de suelo. Revista de teledetección: Revista de la Asociación Española de Teledetección, ISSN 1133-0953, Nº. 36, 2011, págs. 26-39.

[2] Barrientos M. R. E., Cruz R. N. y otros (2008). Evaluación potencial de Redes Bayesianas en la clasificación de datos médicos. Revista médica de la universidad Veracruzana. Vol. 8 Num. 1, junio 2008.

[3] Bellaachia, Abdelghani and Guven, Erhan (2005). Predicting breast cancer survivability using data mining techniques. The George Washington University.

[4] Botia J., Sarmiento H. e Isaza C. (2009). Redes neuronales artificiales de base radial como clasificador difuso: Una aplicación en diagnóstico médico. Universidad de Antioquia.

[5] Betancourt, G. A. (2005). "Las Máquinas de Soporte Vectorial (SVMs)". Scientia et Technica Año XI, No 27. UTP. ISSN 0122-1701.

[6] Flores G. H. (2016). "Redes Neuronales Aplicadas a la Detección de Cáncer de Mama". Tesis IPN.

[7] Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando WEKA. Universidad Tecnológica Nacional, Facultad Regional Córdoba.

[8] Lozano P. F. (2011). Integración del algoritmo CTC en la plataforma WEKA, Universidad del País Vasco. Irun, España.

[9] Marin-Castro H.M. y Franco-Vázquez P.E. (2017). "Estudio de Herramientas de Minería de Datos para la Tarea de Clasificación". Universidad Politécnica de Victoria, Av. Nuevas Tecnologías 5902, Parque Científico y Tecnológico de Tamaulipas, C.P. 87138, Cd Victoria, Tamaulipas, México. Tecno Intelecto 2017, 14(1):1-9.

[10] Martínez F. C. (2016). Detección Automática de Anomalías Presentes en mamografías Digitales. Tesis IPN.

[11] Merz, C.J., and Murphy, P.M. (1996). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

[12] Oommen, t., Misra and Twarakavi, N. K. C., Prakash, Sahoo, b. & Bandopadhyay, S. (2008). An objective analysis of Support vector machine-based classification for remote sensing.

[13] John Ross Quinlan (1993). C4.5 Programs for machine Learning. Morgan Kaufmann Publishers, San Mateo.

[14] Rodríguez L. V. Análisis de imágenes de mamografía para la detección de cáncer de mama (2012). Temas de Ciencia y Tecnología. 2012; 15(47): 39-45.

[15] Organización Mundial de la Salud (OMS, 2016). "Estadísticas a Propósito del Día Mundial de la Lucha Contra el Cáncer de Mama". http://www.inegi.org.mx/saladeprensa/aproposito/2016/mama2016_0.pdf.

[16] M.R.M. Samulski (2006). "Classification of breast lesions in digital mammograms," Master's thesis, University Medical Center Nijmegen, 2006.

[17] UCI Machine Learning Repository (2017). Center for Machine Learning and Intelligent Systems: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.