Detección de cáncer de mama usando técnicas avanzadas de minería de datos con redes neuronales

Jorge Armando Ortiz Murillo[1-4], José María Celaya Padilla[1-4], María Del Rosario Martínez Blanco[1-4], Luis Octavio Solís Sánchez[1-4], Rodrigo Castañeda Miranda[1-4], Idalia Garza Veloz[1-3], Margarita Martínez Fierro[1-3], Yamilé López Hernández[1-3], José Manuel Ortiz Rodríguez[1-4]

Universidad Autónoma de Zacatecas, Av. Ramón López Velarde 801, Col. Centro 98000 Zacatecas, México:

[1]Centro de Investigación e Innovación Tecnológica Industrial (CIITI)

[2]Laboratorio de Innovación y Desarrollo Tecnológico en Inteligencia Artificial, (LIDTIA)

[3]Laboratorio de Bioinformática

[4]Unidad Académica de Ingeniería Eléctrica (UAIE)

Abstract

El cáncer de mama es uno de los problemas de salud más grandes a nivel mundial, es el cáncer más diagnosticado en mujeres y la prevención parece imposible ya que su causa es desconocida, debido a esto, la detección temprana juega un papel fundamental en el pronóstico del paciente. En países en desarrollo como México donde el acceso a servicios especializados de salud es escaso, la revisión clínica regular es poco frecuente y no se cuenta con suficientes radiólogos. La forma más común de detección del cáncer de mama es la auto examinación pero esta solo se detecta en etapas más avanzadas, cuando ya es palpable. Debido a lo anterior, el objetivo del presente trabajo es crear un sistema de Diagnóstico Asistido por computadora (CADx) empleando técnicas de análisis de información como minería de datos y técnicas avanzadas de Inteligencia Artificial, buscando ofrecer un diagnóstico médico previo o segunda opinión, como si fuera un segundo radiólogo con el objetivo de ayudar a reducir los índices de mortalidad por cáncer de mama. En este trabajo se presentan avances obtenidos en el diseño de algoritmos computacionales empleando técnicas de visión computacional para la extracción de características derivadas de mamografías. Utilizando técnicas de análisis de información big data con minería de datos, es posible identificar a pacientes con un alto riesgo de cáncer de mama. Con la información obtenida del análisis de las mamografías, el objetivo en la etapa siguiente será establecer una metodología para la generación de bio-marcadores imagenológicos que permitan establecer

un índice de riesgo de cáncer de mama para pacientes mexicanos, en esta primera etapa se presentan resultados de la clasificación de pacientes con alto y bajo riesgo de padecer cáncer de mama utilizando redes neuronales.

Keywords: *Cáncer de mama, visión computacional, minería de datos y redes neuronales.*

1.- INTRODUCCION

1.1 Industria 4.0

El termino Industria 4.0 fue mencionado por primera vez en a "Hannover Fair" con la presentación de la iniciativa "Industry 4.0". La primera revolución industrial "Mecanización" como resultado de la invención de la máquina de vapor, la segunda "Producción en masa" con la ayuda de la electricidad, la tercera "Digitalización" con el uso la Electrónica y las Tecnologías de Información, esto marca la venida de la cuarta revolución industrial con el uso de los sistemas físicos cibernéticos por sus siglas en ingles CPS y el internet de las cosas y servicios[1]. El objetivo de la Industria 4.0 es la aparición de fábricas digitales con las siguientes características:

- Creación de redes inteligentes[1].
- Movilidad[1].
- Flexibilidad[1].
- Integración de clientes[1].
- Nuevos modelos de negocio innovadores[1].

1.2 Salud 4.0

La industria 4.0 llegó para quedarse, la digitalización de datos vitales, las recomendaciones para los pacientes, el registro de hábitos cotidianos, es el Big Data aplicado al sector de la salud, el objetivo de muchos productos y soluciones están alineados con incrementar la experiencia de los usuarios, darles información relevante para su salud y ofrecer soluciones, a fin de cuentas se trata de utilizar la tecnología para mejorar la calidad de vida.

1.3 Cáncer de Mama

El cáncer de mama es el problema de salud más significante del mundo, en México cuenta con aproximadamente 148 mil nuevos casos de cáncer diagnosticados y 78 mil mujeres mueren de cáncer de mama cada año[2], el cáncer de mama es el cáncer más frecuente en mujeres, por el cual el 1% de todas las mujeres viven con el mismo y de morir por el mismo es de 1 cada 26.8[3].

La prevención primaria parece imposible ya que su causa de este cáncer es desconocida, donde la detección temprana es la clave, Los estudios por mamografías son lo más aceptado a nivel mundial, han demostrado tener una efectividad de reducir su índice de mortalidad entre un 30 y 70%[3], pero cuanta con limitaciones de observadores humanos y es difícil para los radiólogos proveer resultados acertados y uniformes es difícil debido a la cantidad de mamografías generadas[2].

Las mujeres cuyos tumores fueron encontrados tempranamente por mamografías, su tasa de supervivencia de cinco años son del 82% y en las que no de 60%[2].

Las limitaciones de los observadores humanos tienen entre 10 y 30% de errores en lesiones de mama, con el procesamiento digital de imagen, reconocimiento de patrones e inteligencia artificial, los radiólogos en promedio incrementan en 10% la sensibilidad[2], debido a los problemas que presentan los revisores humanos, la comunidad científica ha tratado de ayudarles para esto se crearon los sistemas CAD.

Los radiólogos utilizan los sistemas CAD para asistirlo en las fallas de detección de signos de cáncer visibles en las mamografías (CADe) y para asistir en la clasificación de cáncer benigno o maligno y diagnóstico de lesiones (CADx)[4].

Los sistemas CAD pueden detectar pequeños errores ocasionados por los humanos y con esto reducir los falsos negativos, el análisis de imágenes médicas por computadora fue introducido en los 60's por Lusted que sugirió sé que puede analizarse automáticamente para distinguir entre una imagen normal y anormal[4].

Estos son algunos de los problemas que se tienen con el uso de sistemas CAD:

- Micro calcificaciones son muy pequeñas, entre .1 y 1 mm, promedio .3mm y algunas son menores a .1mm, las cuales no pueden ser detectadas en film-screen mamografía de ruido de alta frecuencia[2].
- Micro calcificaciones con diferentes tamaños, formas y distribuciones[2].
- Micro calcificaciones pueden ser de contraste bajo entonces la diferencia de intensidad entre áreas sospechosas y los tejidos circundantes puede ser muy delgada[2].
- Micro calcificación pueden estar muy cerca de los tejidos circundantes y los algoritmos de segmentación simple, no funcionan bien[2].
- En tejidos densos, especialmente en los senos de mujeres jóvenes, las zonas sospechosas son casi invisibles, los tejidos densos en mujeres jóvenes pueden ser fácilmente malinterpretadas como micro calcificaciones y con alta tasa de falsos negativos[2].

1.4 Minería de Datos

En la actualidad la minería de datos se utiliza para resolver varios problemas como por ejemplo la clasificación de correos en spam y no spam, las sugerencias de amigos en facebook, las búsquedas en google, etc. La minería de datos se encarga de hacer que las maquinas aprendan sin ser programadas, mediante el uso de redes neuronales imitando cómo funciona el cerebro humano[5].

Para esto existen diferentes tipos de aprendizaje [5]:

- Supervisado al cual le damos un conjunto de datos “Respuestas correctas” y en base a esos se obtiene otra respuesta correcta, estos a su vez se clasifican:
    - Problemas de regresión, se tiene de salida N datos.
    - Problemas de clasificación, se tiene una salida binaria: bueno y malo, si y no, etc.
- No supervisado, este tipo de aprendizaje se basa en que no se sabe porque característica se van a agrupar los datos.

1.5 Redes Neuronales

Las redes neuronales son algoritmos que intentan imitar el cerebro humano, en los 80´s y principios de los 90’s eran muy usados pero a fines de los 90´s disminuyó su popularidad, actualmente resurgió el uso de las técnicas de su estado de arte para muchas aplicaciones ya que se mejoró la velocidad de procesamiento de las maquinas[5].

En este trabajo se propone una metodología para la detección y clasificación de lesiones de cáncer de mama mediante el uso de mamografías con técnicas de visión computacional y minería de datos.

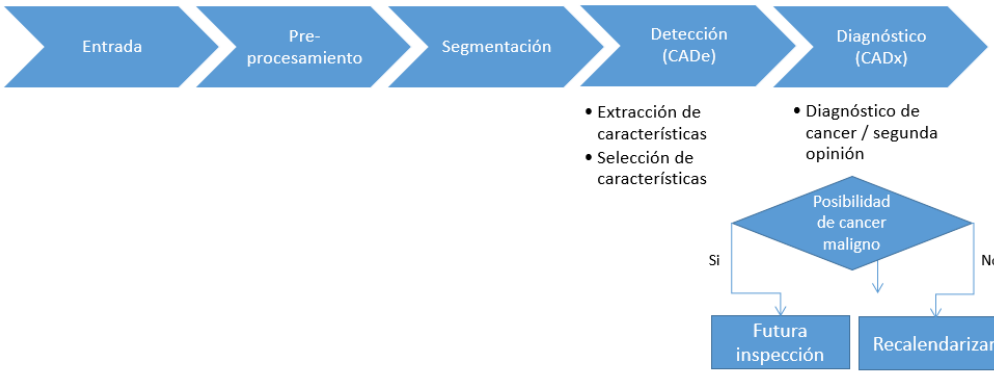
2.- MATERIALES Y METODOS

2.1.- Métodos

La metodología propuesta tiene como objetivo ofrecer una segunda opinión al radiólogo, para esto, se propuso la siguiente metodología en la que se busca imitar los pasos que realiza el radiólogo. Brevemente, primero las imágenes de mamografía son pre-procesadas para realizar y mejorar su visualización. Posteriormente, el radiólogo segmenta las áreas sospechosas donde se encuentra alguna anomalía. Estas lesiones posteriormente son catalogadas como Malignas o Benignas dependiendo del resultado de su biopsia. La metodología propuesta busca evitar el proceso de extracción de biopsias, de tal manera que usando características presentes en la imagen puedan ser usadas como forma de clasificación del tipo de lesión.

En la imagen 1 se muestra el procedimiento que se realizó para la detección de cáncer en las anormalidades sin realizar biopsias. Brevemente, al igual que en la metodología utilizada por el radiólogo: en una primera instancia se pre-procesa la imagen, posteriormente se segmentan las áreas sospechosas con alguna anormalidad. Utilizando estas áreas sospechosas la metodología propuesta, caracteriza y extrae varias características que permitan en una siguiente etapa generar un modelo de clasificación de lesiones Malignas y Benignas, esta clasificación puede ser usada por el radiólogo como una segunda opinión y evitar el proceso de extracción de biopsias in necesarias.

Imagen 1.- Modelo



2.2.- Materiales

En esta investigación se utilizó la base de datos BCDR-D01 del repositorio digital de cáncer de pecho (bcdr.inegi.up.pt)[6], el dataset está conformado por 79 lesiones de cáncer comprobadas por medio de una biopsia realizada a 64 mujeres, de las cuales se sacaron 143 segmentaciones que incluyen información clínica y descriptores basados en las imágenes[7].

En la tabla 1 se muestra la información clínica y general que se tiene de las biopsias:

Tabla 1.- Información clínica y general

Feature	Description
Age	The age of the patient at the time of the study
Breast Density	The density of the breast at the time of the study according to the BI-RADS standard
Mammography Nodule	The lesion contains a mass
Mammography Calcification	Calcifications were detected in the lesion
Mammography Microcalcification	Microcalcifications were detected in the lesion
Mammography Axillary Adenopathy	Axillary adenopathy detected
Mammography Architectural Distortion	Signs of architectural distortion
Mammography Stroma Distortion	Signs of stroma distortion
Classification	Classification of lesion given by the biopsy result.
Image View	Type of image view (1-RCC, 2-LCC, 3-RO, 4-LO).
Mammography Type	Presence of abnormality.

En la tabla 2 se muestra un conjunto de descriptores de intensidad calculados directamente de los niveles de grises de los pixeles que están dentro de la lesión identificada por los radiólogos.

Tabla 2.- Conjuntos de descriptores de intensidad.

Feature	Description
Mean (i_mean)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, with n being the number of pixels inside the region delimited by the contour and x_i being the grey level intensity of the i[th] pixel inside the contour.
Standard Deviation (i_std)	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
Skewness (i_skewness)	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}$
Kurtosis (i_kurtosis)	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}-3$
Minimum (i_min)	The minimum intensity value in the region surrounded by the contour
Maximum (i_max)	The maximum intensity value in the region surrounded by the contour

En la tabla 3 se muestra un conjunto de descriptores de textura calculados apartir de la matriz de co-ocurrencia de nivel de gris relacionados con el cuadro delimitador del contorno de la lesión identificada por los radiólogos.

En la tabla 4 se muestra un conjunto de descriptores de forma y localización de la lesión identificada por el radiólogo.

Tabla 3. Conjunto de descriptores de textura.

Feature	Description
Energy (t_energ)	$\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)^2$ with L being the number of grey-levels, and p being the grey-level co-occurrence matrix and, thus, $p(i,j)$ is the probability of pixels with grey-level i occur together to pixels with grey-level j.
Contrast (t_contr)	$\sum_{i}\sum_{j}(i-j)^2 p(i,j)$
Correlation (t_corr)	$\frac{\sum_{i}\sum_{j}(ij)p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$ with μ_x, μ_y, σ_x and σ_y being the means and standard deviations of p_x and p_y, the partial probability density functions.
Sum of Squares: Variance (t_sosvh)	$\sum_{i}\sum_{j}(i-\mu)^2 p(i,j)$ with μ being the mean of $p(i,j)$ for all i and j.
Homogeneity (t_homo)	$\sum_{i}\sum_{j}\frac{1}{1+(i-j)^2} p(i,j)$
Sum Average (t_savgh)	$\sum_{i=2}^{2L} i p_{x+y}(i)$ with $p_{x+y}(i)$ being the probability of the co-occurrence matrix coordinates summing $i=x+y$
Sum Entropy (t_senth)	$se=-\sum_{i=2}^{2L} p_{x+y}(i)\log\left(p_{x+y}(i)\right)$
Sum Variance (t_svarh)	$\sum_{i=2}^{2L}(i-se)^2 p_{x+y}(i)$
Entropy (t_entro)	$-\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log(p(i,j))$
Difference Variance (t_dvarh)	$\sum_{i=0}^{L-1} i^2 p_{x-y}(i)$ with $p_{x-y}(i)$ being the probability of the co-occurrence matrix coordinates subtracting $i=\lvert x-y\rvert$
Difference Entropy (t_denth)	$-\sum_{i=0}^{L-1} p_{x-y}(i)\log\left(p_{x-y}(i)\right)$
Information Measure of Correlation 1 (t_inf1h)	$\frac{-\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log(p(i,j))+\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log\left(p_x(i)p_y(j)\right)}{\max\left(\sum_{i=1}^{L} p_x(i)\log(p_x(i)),\sum_{i=1}^{L} p_y(i)\log\left(p_y(i)\right)\right)}$
Information Measure of Correlation 2 (t_inf2h)	$\sqrt{1-\exp\left(2\left(\sum_{i=1}^{L}\sum_{j=1}^{L} p_x(i)p_y(j)\log\left(p_x(i)p_y(j)\right)-\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log(p(i,j))\right)\right)}$

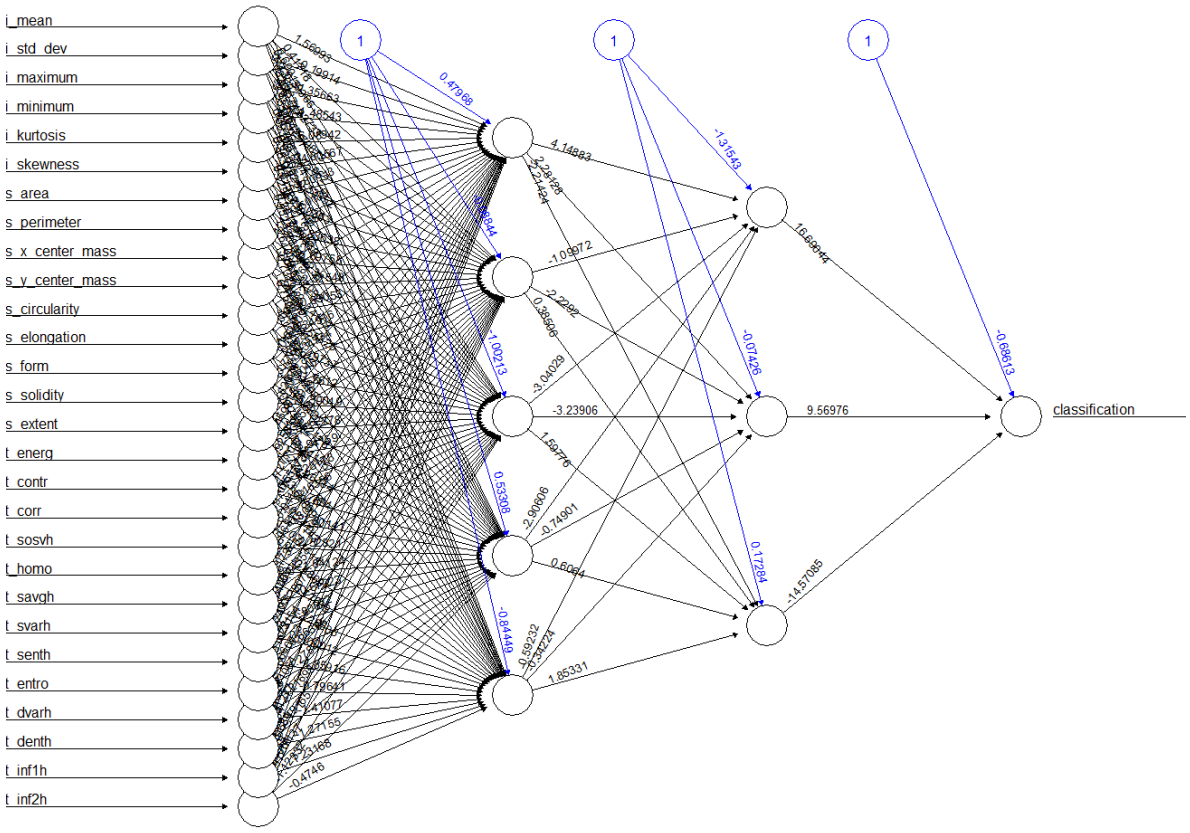
Tabla 4. Conjunto de descriptores de forma y localización.

Feature	Description
Area (s_area)	$area = \|O\|$ with O being the set of pixels that belong to the segmented lesion
Perimeter (s_perimeter)	$perimeter = length(E)$ with $E \subset O$ being the edge pixels
Center of mass (s_x_center_mass, s_y_center_mass)	Normalized coordinates of the center of mass of O
Circularity (s_circularity)	$4\pi \frac{area}{perimeter^2}$
Elongation (s_elongation)	$elongation = \frac{m}{M}$ with m being the minor axis and M the major axis of the ellipse that has the same normalized second central moments as the region surrounded by the contour
Form (s_form)	$\frac{perimeter \times elongation}{8 \times area}$
Solidity (s_solidity)	$\frac{area}{\|H\|}$ with H being the set of pixels that belong to the convex hull of the segmented region
Extent (s_extent)	$\frac{area}{\|B\|}$ with B being the set of pixels that belong to the bounding box of the segmented region

Como forma de obtener un modelo de clasificación que permitiría distinguir entre lesiones malignas y lesiones benignas se construyó una red neuronal de propagación hacia atrás la cual tomaron como datos de entrada a la red neuronal los conjuntos de descriptores de intensidad, textura, forma y localización.

La red neuronal se formó por la capa de entrada con las 28 características, con 2 capas ocultas, la primera de 5 y la segunda de 3 y la capa de salida binaria con la clasificación de cáncer maligno o benigno, como se muestra en la imagen 2.

Imagen 2.- Red Neuronal



La red neuronal se probó con los datos normalizados y sin normalizar, se realizó validación cruzada con datos para entrenamiento 75% y para prueba 25% y se corrieron diferentes configuraciones de la(s) capa(s) oculta(s).

3.- RESULTADOS

Como se puede ver en la Imagen 2 y 3 utilizando los datos sin normalizados se obtuvo un AUC = 0.9797 y AUC = 0.8509 para train y test respectivamente, y en las imágenes 4 y 5 con datos normalizados se obtuvo un AUC = 0.7435 y AUC = 0.7078 para train y test respectivamente.

Imagen 2.- Train – Datos sin normalización

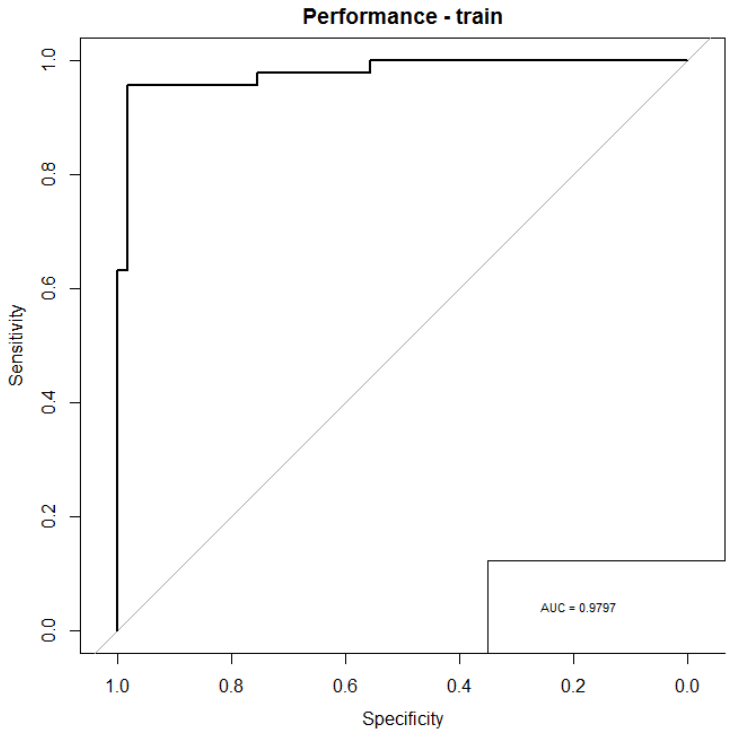


Imagen 3.- Test – Datos sin normalización

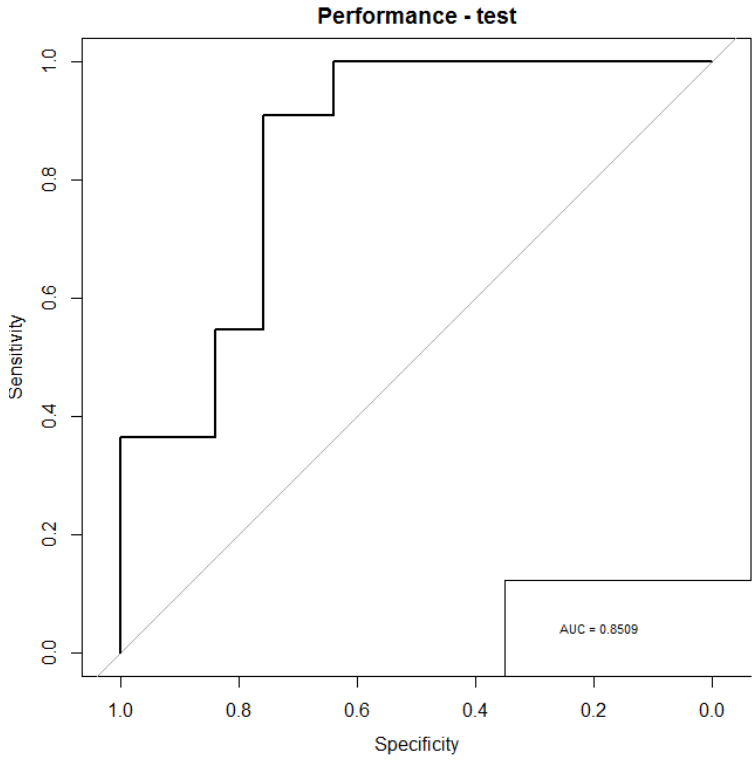


Imagen 4.- Train – Datos normalizados

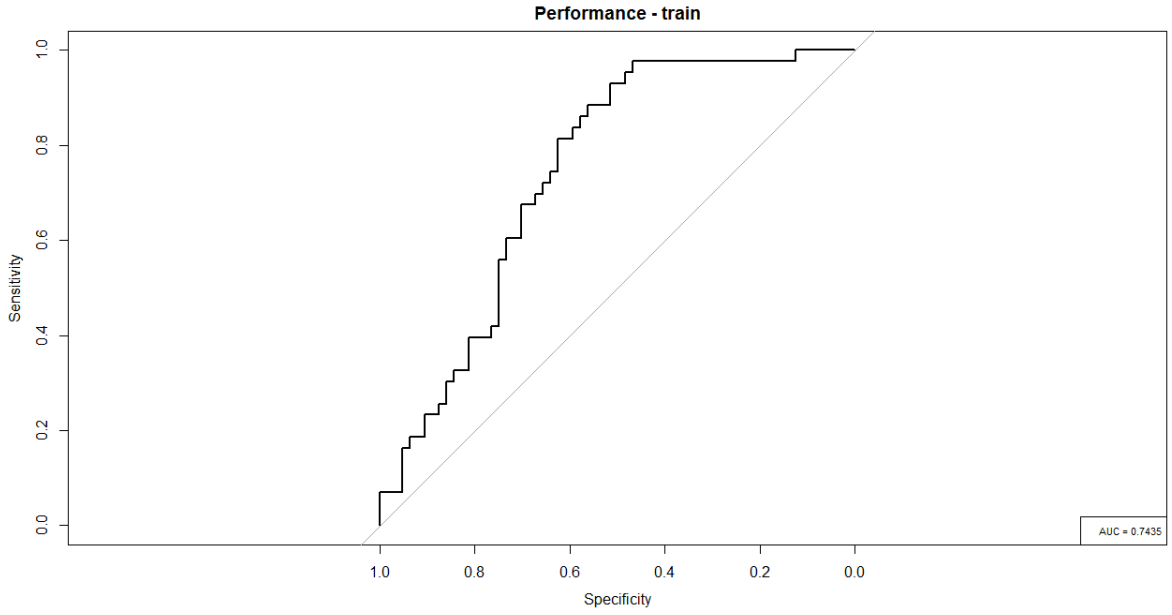
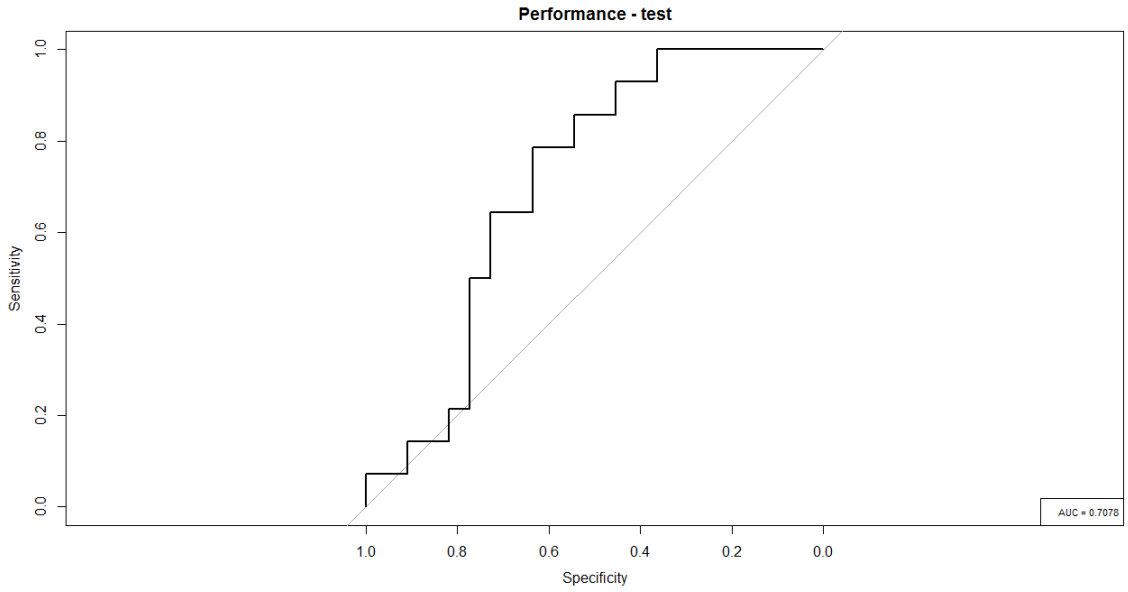


Imagen 5.- Test – Datos normalizados



En la imagen 6 se muestra el resultado de la tabla de confusión con los datos de los casos benignos y malignos acertados por el algoritmo, así como los falsos negativos y positivos.

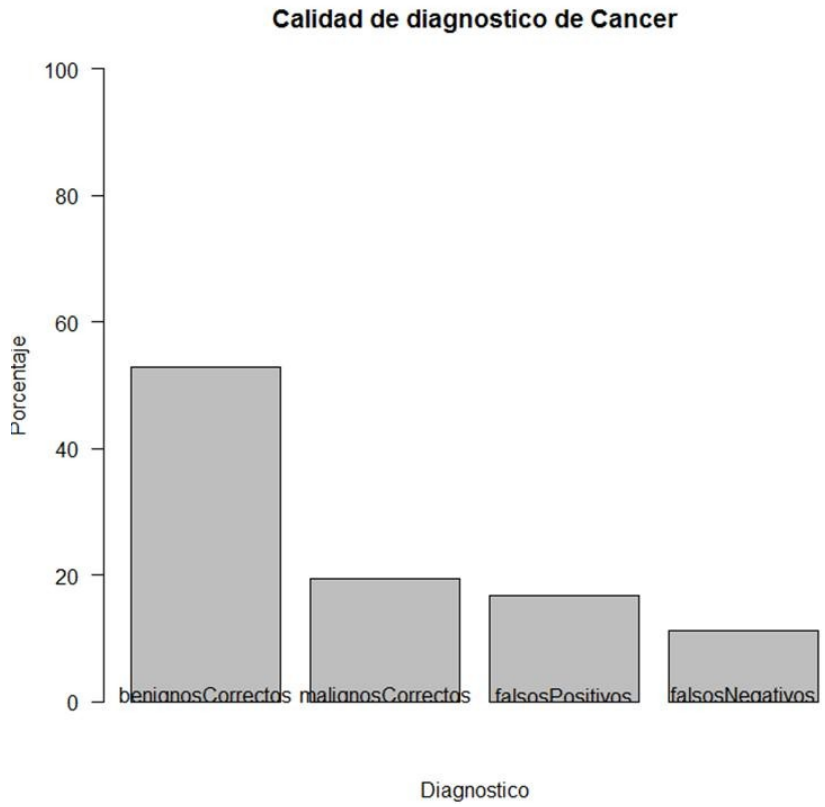


Imagen 6.- Calidad de diagnóstico.

En la tabla 5 se muestra el área bajo la curva resultante de modificar la cantidad de neuronas en las capas ocultas para poder comprobar que tanto afecta la distribución de la red neuronal.

Tabla 5.- Capas Ocultas

1ra oculta	2da oculta	Train - AUC	Test - AUC
1	1	0.9608	0.8765
2	1	0.9605	0.8971
3	1	0.9746	0.8477
4	1	0.9894	0.8642
5	1	1.0000	0.8354
6	1	0.9915	0.8601
7	1	0.9936	0.8642
8	1	0.9944	0.8354
1	2	0.9753	0.8519
2	2	0.9892	0.8333
3	2	0.9852	0.8765
4	2	0.9693	0.8642
5	2	1.0000	0.7613
6	2	1.0000	0.7860
7	2	0.9703	0.8889
8	2	1.0000	0.8066
1	3	0.9718	0.8765
2	3	0.9848	0.8642
3	3	0.9552	0.8848
4	3	1.0000	0.8189
5	3	0.9799	0.8724
6	3	1.0000	0.7860
7	3	1.0000	0.8107
8	3	1.0000	0.8148
1	4	0.9594	0.8477
2	4	0.9693	0.8580
3	4	0.9929	0.8724
4	4	0.9809	0.8807
5	4	0.9908	0.8025
6	4	1.0000	0.7407
7	4	1.0000	0.8560
8	4	1.0000	0.8272
1	5	0.9746	0.8436
2	5	0.9926	0.8189
3	5	0.9898	0.8519
4	5	1.0000	0.8148
5	5	0.9936	0.8642
6	5	1.0000	0.8560
7	5	1.0000	0.7860
8	5	1.0000	0.8519

4.- DISCUSION

Mediante el uso de una red neuronal de propagación hacia atrás para la clasificación de pacientes con alto y bajo riesgo de padecer cáncer de mama con los datos sin normalizar y utilizando todas las características con un área bajo la curva de 0.9018. Los resultados de este trabajo se limitaron a un escáner de la base de datos pública BCDR y como se demostró en la sección de resultados en la tabla 5, sin importar el tipo de arquitectura de la red neuronal se obtienen resultados similares y con un excelente desempeño.

5.- CONCLUSIONES

En esta investigación se demostró que se puede clasificar un paciente con riesgo de padecer cáncer de mama con la extracción de características de las mamografías y la clasificación que proporcionan las redes neuronales y en este caso con la propagación hacia atrás.

Acknowledgments

Este trabajo fue apoyado económicamente por beca CONACYT con el número de apoyo 439502 y número de registro becario 601498 para el grado de MAESTRÍA.

REFERENCIAS

[1] N. Jazdi, "Cyber physical systems in the context of Industry 4.0," in *Automation, Quality and Testing, Robotics, 2014 IEEE International Conference on*, 2014, pp. 1-4: IEEE.

[2] H.-D. Cheng, X. Cai, X. Chen, L. Hu, and X. Lou, "Computer-aided detection and classification of microcalcifications in mammograms: a survey," *Pattern recognition,* vol. 36, no. 12, pp. 2967-2991, 2003.

[3] R. M. Rangayyan, F. J. Ayres, and J. L. Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," *Journal of the Franklin Institute,* vol. 344, no. 3, pp. 312-348, 2007.

[4] L. H. Eadie, P. Taylor, and A. P. Gibson, "A systematic review of computer-assisted diagnosis in diagnostic cancer imaging," *European journal of radiology,* vol. 81, no. 1, pp. e70-e76, 2012.

[5] N. Andrew, "Machine Learning," in *Machine Learning*, C. Stanford University, Ed., ed: Coursera, 2016.

[6] D. C. Moura and M. A. G. López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *International journal of computer assisted radiology and surgery,* vol. 8, no. 4, pp. 561-574, 2013.

[7] M. A. G. López, N. Posada, D. C. Moura, R. Polln, J. M. F. Valiente, and C. S. Ortega, "BCDR: a breast cancer digital repository," in *15th International Conference on Experimental Mechanics*, 2012.