DETECCIÓN AUTOMÁTICA DE ARTEFACTOS EN MAMOGRAFÍA DIGITAL

GERSON AFRICANO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2019

DETECCIÓN AUTOMÁTICA DE ARTEFACTOS EN MAMOGRAFÍA DIGITAL

GERSON AFRICANO

Trabajo de grado para optar al título de Ingeniero Electrónico

Director
Said Pertuz
Doctorado en Ingeniería Informática

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2019

Índice general

Índice de figuras

Índice de tablas

RESUMEN

TÍTULO: DETECCIÓN AUTOMÁTICA DE ARTEFACTOS EN MAMOGRAFÍA DIGITAL *

AUTOR: GERSON AFRICANO **

PALABRAS CLAVE: Artefacto, cáncer, mamografía.

DESCRIPCIÓN:

El diagnóstico temprano de cáncer de seno aumenta las posibilidades de supervivencia. Por ello, es importante la implementación de técnicas que permitan su detección temprana. A partir de la mamografía y con la ayuda de algoritmos específicos se pueden extraer características del seno, tales como la proporción de tejido denso, que es un factor de riesgo ya establecido. El problema con estos algoritmos es que no contemplan que la mamografía presente artefactos, los cuales pueden alterar las medidas tomadas de la imagen. En este documento se estudian dos tipos de artefactos que aparecen en algunas imágenes: dispositivos de ampliación e implantes de seno. El objetivo de este trabajo es estudiar estrategias para la detección automática de artefactos en imágenes mamográficas. Para ello, se estudiaron 3 tipos de métodos: métodos empíricos, métodos de sub-espacios, entre los que se consideraron 3 modelos diferentes (RL paso a paso, PCA+RL y LDA+RL), y métodos basados en *machine learning* (SVM y *boosting trees*). Los mejores desempeños con cada uno de los tipos de métodos en la clasificación de las imágenes fueron, 23,18 % de error para el método empírico, 5,63 % para el método de sub-espacios (LDA+RL) y 4,5 % con *machine learning* (*boosting tree*).

* Trabajo de grado

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Said Pertuz, Doctorado en Ingeniería Informática.

ABSTRACT

TITLE: AUTOMATIC DETECTION OF ARTIFACTS IN DIGITAL MAMOGRAPHY*

AUTHOR: GERSON AFRICANO**

KEYWORDS: Artifact, cancer, mammography.

DESCRIPTION:

Early diagnosis of breast cancer increases the chances of survival. Therefore, the implementation of techniques that allow early detection is important. From mammography and with the help of specific algorithms, breast characteristics can be extracted, such as the proportion of dense tissue, which is an established risk factor. The problem with these algorithms is that they do not contemplate that mammography presents artifacts, which can alter the measurements taken from the image. In this document we study two types of artifacts that appear in some images: enlargement devices and breast implants. The objective of this work is to study strategies for the automatic detection of artifacts in mammographic images. For this, 3 types of methods were studied: empirical methods, sub-space methods, among which 3 different models were considered (stepwise RL, PCA + RL and LDA + RL), and methods based on textit machine learning (SVM and textit boosting trees). The best performances with each of the types of methods in the classification of the images were, 23.18 % error for the empirical method, 5.63 % for the sub-spaces method (LDA + RL) and 4,5 % with textit machine learning (textit boosting tree).

* Bachelor Thesis

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Said Pertuz, Doctorado en Ingeniería Informática.

INTRODUCCIÓN

En Colombia el cáncer de mama se perfila como un problema de salud pública debido a las dificultades para el diagnóstico y la detección temprana. En el año 2014 la tasa de mortalidad fue de 2649 mujeres en todo el territorio nacional, presentándose una tendencia al incremento en los próximos años[1]. A la alta incidencia se suman los distintos problemas de atención y acceso a los procesos de diagnóstico por parte de la población afectada; por lo tanto, es prioritaria la implementación de estrategias de detección temprana de cáncer de seno. En el diagnóstico del cáncer de mama se han buscado constantemente nuevas estrategias que mejoren y faciliten su detección cuando aún no se presentan síntomas y el cáncer es más tratable[2].

Una de las principales herramientas para el estudio, análisis y diagnóstico de este tipo de cáncer es la mamografía. La mamografía es un tipo de imagen médica especializada que utiliza un sistema de dosis baja de rayos X para visualizar el interior del seno. El caso particular de la mamografía digital, también llamada mamografía digital de campo completo (FFDM *full field digital mammography*), hace referencia a un sistema mamográfico en el que la película de captura es reemplazada por sensores electrónicos que transforman los rayos X en imágenes. Estas imágenes se transfieren a una computadora para su revisión y almacenamiento a largo plazo. El análisis de las mamografías las realiza un radiólogo, quien da un diagnóstico subjetivo basado en su experiencia.

Con el avance de la tecnología, se han buscado soluciones que incrementen el valor agregado del radiólogo mediante el desarrollo de herramientas de soporte para la interpretación y el diagnóstico. Los sistemas de detección asistida por computadora (CAD *computer-aided detection*) son un ejemplo de herramienta que ayuda a los radiólogos para determinar regiones de interés en una mamografía, y para la detección y el diagnóstico de anomalías[3]. El surgimiento de la mamografía digital ha facilitado e incentivado

[1]social, Ministerio de salud y protección. *Cáncer de mama, una enfermedad en ascenso en Colombia*. https://www.minsalud.gov.co/Paginas/-Cancer-de-mama,-una-enfermedad-en-ascenso-en-Colombia.aspx. cosultado:11.05.2019. 2014.

[2]Xi, Xiaoming y col. «Breast tumor segmentation with prior knowledge learning». En: *Neurocomputing* 237 (2017), págs. 145-157.

[3]Dromain, C y col. «Computed-aided diagnosis CAD in the detection of breast cancer». En: *Euro-*

la implementación y desarrollo de los sistemas CAD[4].

En el cáncer de mama existen algunos factores de riesgo confirmados que determinan si una persona es propensa a padecer esta enfermedad. Entre los factores de riesgo más reconocidos se encuentran la edad, antecedentes familiares, y densidad del seno, entre otros[5].

Teniendo en cuenta que la densidad del seno ha mostrado ser un factor de riesgo relevante, es necesario encontrar técnicas que mejoren la forma en cómo se realiza la estimación del tejido denso del seno a partir de la mamografía. En la práctica clínica, uno de los métodos de estimación de tejido denso del seno lo lleva a cabo un radiólogo, el cual subjetivamente realiza un análisis de la mamografía para cuantificar la cantidad de tejido "denso" o clasifica el seno en diferentes categorías según la distribución del tejido denso (e.g., las categorías BI-RADS[6]). Con el objetivo de facilitar esta tarea y de eliminar la subjetividad en la lectura, en la actualidad, existen algoritmos que permiten realizar el cálculo del tejido denso de forma automática a partir de las mamografías[7].

Un paso previo para el cálculo automático de la densidad, es la detección del contorno del seno y del músculo pectoral[8], tal como se ilustra en la Fig. 1. Uno de los problemas del análisis automático de imágenes mamográficas es que al momento de la adquisición de las imágenes, en algunos casos se introducen artefactos para la ampliación de zonas específicas del seno o por la presencia de implantes (ver Fig. 2). Estos artefactos dificultan el proceso de análisis de imágenes mamográficas y tienen el potencial de alterar las medidas cuantitativas que se realizan de forma automática. En la Figura 3 se puede observar el resultado de la segmentación automática del tejido denso del seno en una imagen con artefactos. En esta imagen se puede observar que el artefacto está siendo

pean journal of radiology 82.3 (2013), págs. 417-423.

[4]Ibíd.

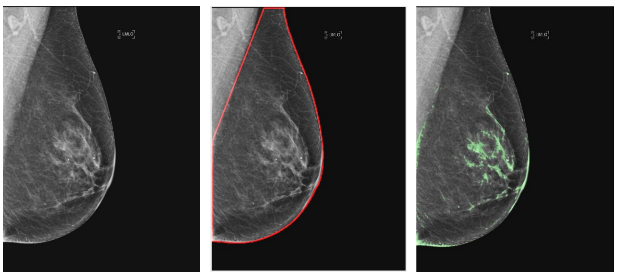
[5]Guerrero, Rachael T Leon y col. «Risk factors for breast cancer in the breast cancer risk model study of Guam and Saipan». En: *Cancer epidemiology* 50 (2017), págs. 221-233.

[6]Østerås, Bjørn Helge y col. «BI-RADS density classification from areometric and volumetric automatic breast density measurements». En: *Academic radiology* 23.4 (2016), págs. 468-478.

[7]Quintana, C; Redondo, M y Tirao, G. «Implementation of several mathematical algorithms to breast tissue density classification». En: *Radiation Physics and Chemistry* 95 (2014), págs. 261-263.

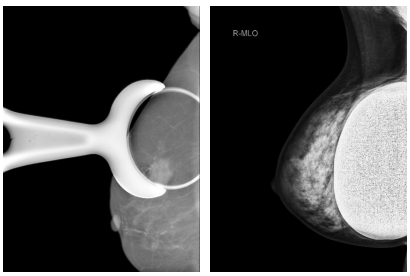
[8]Rampun, Andrik y col. «Fully automated breast boundary and pectoral muscle segmentation in mammograms». En: *Artificial intelligence in medicine* 79 (2017), págs. 28-41.

Figura 1: Imagen procesada sin artefactos. (a) Imagen de entrada. (b) Segmentación del seno. (c) Segmentación de tejido denso (resaltado en verde)



(a) (b) (c)

Figura 2: Artefactos que afectan el análisis automático de imágenes mamográficas. (a) Dispositivo de ampliación. (b) Implantes de seno

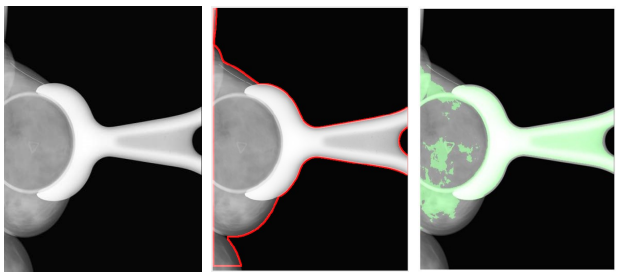


(a) (b)

tomado como tejido denso, alterando así la estimación de densidad.

En la literatura, los algoritmos desarrollados suelen validarse en mamografías que no presentan "artefactos". Sin embargo, en la práctica clínica, normalmente aparecen mamografías las cuales presentan implantes o dispositivos de ampliación. El propósito de este trabajo es desarrollar un algoritmo para la detección de imágenes mamográficas que presenten artefactos como los mostrados en la Fig. 2. El algoritmo a desarrollar debe identificar de forma totalmente automática las imágenes con dichos artefactos en una base de datos de pruebas. A su vez, la detección de artefactos tiene como objetivo servir como una etapa previa que permita guiar tareas de análisis posteriores.

Figura 3: Imagen procesada con artefacto. (a) Imagen de entrada. (b) Segmentación del seno. (c) Segmentación de tejido denso



(a) (b) (c)

1. METODOLOGÍA

Para el problema planteado se cuenta con una base de datos inicial de 15836 imágenes, la cual se clasificó manualmente identificando las imágenes que poseían artefactos, encontrando 399 de este tipo (2.51 %). Finalmente para el desarrollo del proyecto se creó una base de datos de 798 imágenes de las cuales 399 poseen artefactos resultantes de la clasificación manual y 399 imágenes normales seleccionadas aleatoriamente dentro de la base de datos completa.

Debido a que en la literatura no existen soluciones previas para este problema, se optó inicialmente por desarrollar una solución empírica con un conjunto de características a manera de análisis exploratorio (sección 1.1), la cual será nuestro punto de partida, para luego proponer un nuevo conjunto de características con las que se desea mejorar el desempeño inicial, utilizando métodos mas sofisticados, basados en sub-espacios vectoriales (sección 1.2) y técnicas de *machine learning* (sección 1.3).

1.1. MÉTODO EMPÍRICO

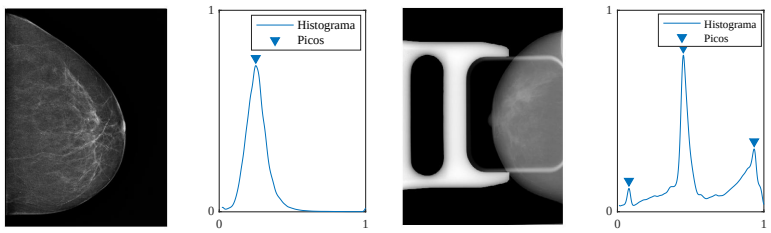
Como se puede observar en la Fig.4, una imagen que posee artefactos puede presentar diferencias en el histograma, especialmente en las intensidades altas donde se espera apreciar el efecto de los artefactos. Por este motivo, se propone un método *ad-hoc* que consiste en extraer características del histograma para intentar clasificar las imágenes.

A continuación se describe la metodología propuesta para este primer planteamiento, que se ilustra en la Fig. 5.

1.1.1. Pre-procesamiento

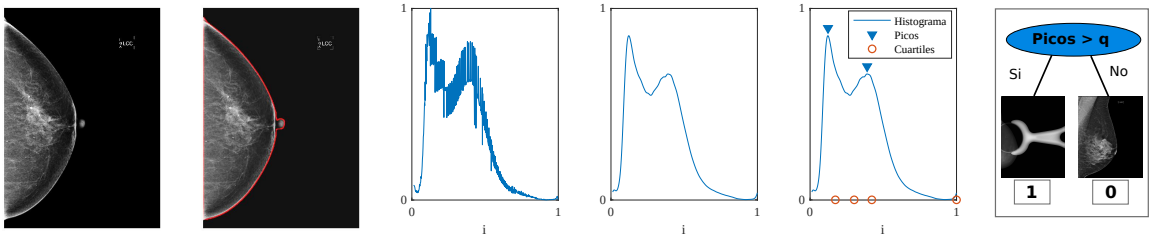
- Adquisición de la imagen: El formato de la imagen es DICOM, la cual se en-

Figura 4: Histograma. (a) Mamografía sin artefacto. (b) Histograma de imagen en (a). (c) Mamografía con artefacto. (d) Histograma de la imagen en (c) en la que el pico cercano a las intensidades altas corresponde al artefacto.



(a) (b) (c) (d)

Figura 5: Método empírico para detección de artefactos. (a) Imagen de entrada. (b) Pre-Segmentación del seno. (c) Histograma del seno. (d) Filtrado del histograma. (e) Extracción de características (conteo de picos y cuartiles del histograma). (f) Clasificación de la imagen según el número de picos dominantes del histograma.



(a) (b) (c) (d) (e) (f)

cuentra en escala de grises con una profundidad de 12-16bits según el sistema mamográfico y sus respectivos valores se almacenan en una matriz $I(x, y)$, donde las dimensiones x y y dependerán del tamaño de la imagen.

- Pre-segmentación de la imagen: Lo que se busca en esta etapa es la eliminación del fondo que no aporta información relevante para el análisis de la imagen. Para ello se toma el histograma de $I(x, y)$ denominado $H(i)$. Específicamente, para una imagen de tamaño $m \times n$, $H(i)$ se define como:

$$H(i) = \sum_{y=0}^{n} \sum_{x=0}^{m} s((I(x, y) - i)) \tag{1.1}$$

$$s(x) = \begin{cases} \text{si } 0 \leq x < \frac{1}{b} & 1 \\ \text{otros casos} & 0 \end{cases} \tag{1.2}$$

  donde b es el número de niveles del histograma y $i \in [0, 1]$ representa el i-ésimo nivel de gris del histograma.

  Para la segmentación, el fondo se detecta simplemente mediante el umbral i_{th} del nivel de gris en función del modo más alto del histograma de intensidad[1].

- Adquisición del histograma de la imagen sin fondo: Una vez obtenido i_{th}, se genera un nuevo histograma de intensidades de la imagen a partir de i_{th} (ver Fig. 6c).

- Filtrado del histograma: Con el objetivo de eliminar picos espurios y máximos locales en el histograma es necesario aplicar un filtrado. Para filtrar el histograma se toma la transformada de Fourier de H y se multiplica con un filtro gaussiano en el dominio de la frecuencia (ver ecuación 3) con valor de $\rho = 10{,}1$, el cual fue escogido de forma que los experimentos tuvieran el mejor desempeño. Posteriormente a este resultado se le toma la transformada inversa de Fourier, obteniendo así el histograma filtrado H_f (ver ecuación 4).

$$G(\omega) = e^{-\rho^2 \omega^2} \tag{1.3}$$

[1]Pertuz, Said y col. «Open Framework for Mammography-based Breast Cancer Risk Assessment». En: *IEEE-EMBS International Conference on Biomedical and Health Informatics* (2019). accepted, págs. 1-4.

Tabla 1: Características extraídas del histograma filtrado y descripción de cada una de ellas

Característica	Descripción
Picos	Número de picos presentes en el histograma del seno segmentado.
Amplitud	Amplitud de los picos del histograma.
Energía	Energía del histograma.
Cuartiles	Posición de los cuartiles del histograma.

$$H_f = \mathcal{F}^{-1}\{G(\omega)\mathcal{F}\{H\}\} \tag{1.4}$$

En la Fig. 6d se ilustra el efecto del filtrado del histograma de la imagen.

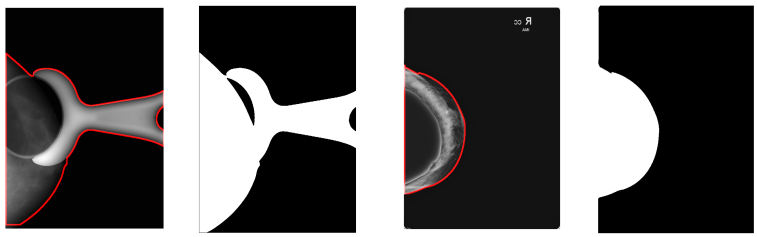
1.1.2. Extracción de características y clasificación

- Extracción de características: Para el método propuesto las características son extraídas del histograma (ver Fig. 6e), las cuales se resumen en la tabla 1.
- Clasificación de las imágenes: En esta etapa a partir de las características extraídas del histograma se hace un estudio experimental usando arboles de decisión de un nivel (ver Fig. 6f), con el objetivo de identificar la(s) característica(s) que permitan la clasificación de las imágenes dentro de las dos categorías posibles: sin artefacto (0) y con artefacto (1),

1.2. MÉTODO DE SUB-ESPACIOS

En este trabajo, las imágenes mamográficas pueden ser agrupadas en dos categorías: cuando el seno es normal, y cuando posee dispositivos de ampliación o implantes. Cuando se presentan dispositivos de ampliación en la imagen, al segmentar y obtener la máscara (ver Fig. 7b) se presentan irregularidades evidentes que se van a aprovechar

Figura 6: Regiones donde fueron extraídas las características para la sección 1.2 y 1.3. (a) Seno segmentado con dispositivo de ampliación. (b) Máscara binaria de la mamografía de la Fig. (a). (c) Seno segmentado con implante. (d) Máscara binaria de la mamografía de la Fig. (c).



(a) (b) (c) (d)

para extraer el nuevo conjunto de características que permitan la clasificación de las imágenes. En el caso de las imágenes con implantes, la máscara no aporta información útil debido a la segmentación aparenta estar correcta (ver Fig. 7d) y no brinda información discriminante de una imagen normal. Por esta razón, se decide también extraer características del seno segmentado en escala de grises (ver Fig. 7a y 7c) para poder clasificar los dos tipos de anomalías. En la tabla 2 se describe el nuevo conjunto de características, las cuales fueron extraídas a partir del seno segmentado y de su máscara binaria.

Para esta sección se realizaron análisis basados en los siguiente métodos: Análisis de componentes principales (PCA)[2] y Análisis discriminante lineal (LDA)[3]. A continuación se presenta como están definidos y cómo se realiza el cálculo de los parámetros asociados a cada uno de ellos.

1.2.1. PCA Este método es comúnmente utilizado para la reducción de la cantidad de variables haciendo uso de la varianza de los datos[4]. Considere el conjunto de datos $\mathbf{x}_n$ donde $n = 1, 2, \ldots, N$ (N=número de imágenes a analizar) y $\mathbf{x}_n \in \mathbb{R}^P$ (P=número

[2]Shang, Han Lin. «A survey of functional principal component analysis». En: *AStA Advances in Statistical Analysis* 98.2 (2014), págs. 121-142.

[3]Yan, Shuicheng y col. «Graph embedding and extensions: A general framework for dimensionality reduction». En: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1 (2007), págs. 40-51.

[4]Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.

Tabla 2: Nuevo conjunto de características para la clasificación de las imágenes

Característica	Descripción
Área	Área de los huecos de la máscara de la imagen.
Área convexa	Área de la región convexa que contiene la imagen.
Centro x	Posición en x del centro de masa de la máscara.
Centro y	Posición en y del centro de masa de la máscara.
Desviación estándar	Desviación estándar del seno segmentado.
Media	Media del seno segmentado.
Excentricidad	Excentricidad de la elipse que tiene los mismos segundos momentos que la máscara de la imagen.
Kmedia	Media de la curvatura de la máscara.
Kstd	Desviación estándar de la curvatura de la máscara.
Diámetro	Diámetro de un circulo con el mismo área de la máscara.
Relación	Relación entre el área de la máscara y el área del rectángulo mas pequeño que contiene la máscara.
Perímetro	Perímetro de la máscara de la imagen.
solidez	Relación de píxeles en el casco convexo que también se encuentran en la máscara.
Eje menor	Longitud en píxeles del eje menor de la elipse que tiene los mismos segundos momentos centrales normalizados como la máscara.
Eje mayor	Longitud en píxeles del eje principal de la elipse que tiene los mismo segundos momentos centrales normalizados como la máscara.
Kurtosis	Kurtosis del seno segmentado.
Skewness	Skewness del seno segmentado.

de características extraídas), la idea detrás de este método es proyectar los datos a un espacio con dimensión $k < P$ maximizando la varianza de los datos proyectados[5]. Inicialmente se considera la proyección a una dimensión para luego extenderse a k dimensiones.

Se define la dirección del nuevo espacio $\mathbf{u}_1 \in \mathbb{R}^P$, el cual cumple la condición $\mathbf{u}_1^T\mathbf{u}_1 = 1$. Cada $\mathbf{x}_n$ es proyectado de la forma.

$$y_1 = \mathbf{u}_1^T\mathbf{x}_n \tag{1.5}$$

Ahora, la varianza de los datos proyectados está dada por $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$, donde $\mathbf{S}$ es la matriz de covarianzas de los datos originales[6].

Como se mencionó anteriormente en este método se busca maximizar la varianza de los datos proyectados, por ende se debe maximizar la expresión $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$, sujeta a la restricción $\mathbf{u}_1^T\mathbf{u}_1 = 1$. Este problema puede ser resuelto mediante multiplicadores de Lagrange, obteniendo como resultado, que $\mathbf{u}_1$ es el vector propio asociado al máximo valor propio de la matriz de covarianzas de los datos $\mathbf{S}$. Generalizando para k-dimensiones se tiene que la proyección lineal óptima para la cual la varianza es maximizada, está definida por los k vectores propios $\mathbf{u}_1,\ldots,\mathbf{u}_k$ de la matriz de covarianzas $\mathbf{S}$ correspondientes a los k mayores valores propios $\lambda_1, \ldots, \lambda_k$[7].

1.2.2. LDA El Análisis Discriminante lineal es una técnica estadística tradicional utilizada en aprendizaje de máquinas, la cual ha sido empleada para diferentes tareas

[5]Ibíd.
[6]Ibíd.
[7]Ibíd.

de clasificación como reconocimiento de la acción humana[8], reconocimiento facial[9] y reconocimiento de personas[10], debido a su efectividad en la reducción de variables y obtención de nuevas variables discriminantes. Sabiendo que cada imagen se puede clasificar dentro de dos categorías, LDA proyecta el vector de características $\mathbf{x} \in \mathbb{R}^P$ a una dimensión de la forma

$$y = \mathbf{w}^T \mathbf{x} \tag{1.6}$$

Donde $\mathbf{w}$ es un vector que es calculado de forma que maximice la varianza entre grupos y minimice la varianza intra-grupos[11].

la covarianza de $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ viene dada por la expresión:

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{1.7}$$

Donde N es el número de imágenes a analizar y $\bar{\mathbf{x}}$ es el vector de medias de cada una de las variables. La covarianza $\mathbf{S}$ se puede expresar como la suma de la covarianza entre grupos ($\mathbf{S}_B$) y la covarianza intra-grupos ($\mathbf{S}_W$)[12].

$$\begin{aligned} \mathbf{S} &= \mathbf{S}_W + \mathbf{S}_B \\ \mathbf{S}_W &= \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{x}^{c_i})(\mathbf{x}_i - \mathbf{x}^{c_i})^T, \\ \mathbf{S}_B &= \sum_{c=1}^{2} n_c (\overline{\mathbf{x}}^c - \overline{\mathbf{x}})(\overline{\mathbf{x}}^c - \overline{\mathbf{x}})^T = N\mathbf{S} - \mathbf{S}_W \end{aligned} \tag{1.8}$$

[8]Iosifidis, Alexandros y col. «Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis». En: *Computer Vision and Image Understanding* 116.3 (2012), págs. 347-360. ISSN: 10773142. DOI: 10.1016/j.cviu.2011.08.008. URL: http://dx.doi.org/10.1016/j.cviu.2011.08.008.

[9]Belhumeur, Peter N; Hespanha, Joao P y Kriegman, David J. *Recognition Using Class Specific Linear Projection*. 1997.

[10]Iosifidis, Alexandros; Tefas, Anastasios y Pitas, Ioannis. «Activity-based person identification using fuzzy representation and discriminant learning». En: *IEEE Transactions on Information Forensics and Security* 7.2 (2012), págs. 530-542.

[11]Belhumeur; Hespanha y Kriegman, *Recognition Using Class Specific Linear Projection*, óp.cit.

[12]Duda, Richard O; Hart, Peter E y Stork, David G. *Pattern classification*. John Wiley & Sons, 2012.

Donde $\mathbf{x}^{c_i}$ es el vector de medias correspondiente a la clase que pertenece $\mathbf{x}_i$, $\mathbf{x}^c$ es el vector de medias correspondiente a la clase c y n_c es el numero de imágenes que pertenece a la clase c[13].

$\mathbf{w}$ es calculado de manera que se maximice la siguiente expresión[14]:

$$J(w) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{1.9}$$

Con este criterio se trata de determinar el eje discriminante de forma que las distribuciones proyectadas sobre el mismo estén lo más separadas posible entre sí (mayor variabilidad entre-grupos) y, al mismo tiempo, que cada una de las distribuciones esté lo menos dispersa (menor variabilidad intra-grupos). Para maximizar la ecuación 6 se deriva respecto $\mathbf{w}$ y se iguala a cero, obteniendo

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w} \tag{1.10}$$

La obtención del vector $\mathbf{w}$ se traduce en tomar el mayor valor propio de $\mathbf{S}_W^{-1}\mathbf{S}_B$ y hallar su respectivo vector propio o simplemente $\mathbf{w} = \mathbf{S}_w^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$[15].

1.3. MACHINE LEARNING

El objetivo principal de esta sección es evaluar algunos algoritmos clásicos de *machine learnig* (Máquina de soporte vectorial (SVM *support vector machine*) y *boosting trees*), para finalmente compararlos con los métodos propuestos anteriormente. Para los métodos propuestos se define el conjunto de datos $T = \{\mathbf{x}_i, y_i\}_1^N$, donde $\mathbf{x}_i \in \mathbb{R}^P$ denota las muestras de entrada, $y_i \in \{-1, 1\}$ es la clase a la que pertenece cada imagen y N el número de imágenes.

[13] Yan y col., «Graph embedding and extensions: A general framework for dimensionality reduction», óp.cit.

[14] Duda; Hart y Stork, *Pattern classification*, óp.cit.

[15] Ibíd.

1.3.1. Máquina de soporte vectorial SVM, propuesto por Vapnik, es un método de aprendizaje automático desarrollado a partir de la teoría estadística. SVM adopta el principio de minimización del riesgo estructural para clasificar los datos mediante la construcción del hiperplano óptimo[16]. Puede resolver eficazmente problemas de clasificación de muestras pequeñas, no lineales y de alta dimensión[17].

SVM busca el plano de clasificación óptimo ajustando **w** y b:

$$f(x) = sign(\mathbf{x}^T\mathbf{w} + b) \tag{1.11}$$

Donde **w** es el vector de pesos y b una constante. El problema de buscar el hiperplano de clasificación óptimo puede transformarse en un problema de optimización cuadrática restringida al considerar exhaustivamente los criterios de minimización del riesgo estructural, los términos de regularización y los errores de ajuste[18]. El hiperplano que separa los datos se puede hallar al minimizar la función de costo $J(\mathbf{w}, \zeta)$:

$$\text{mín } J(\mathbf{w}, \zeta) = \frac{1}{2}\mathbf{w}^2 + C\sum_{i=1}^{N}\zeta_i \tag{1.12}$$

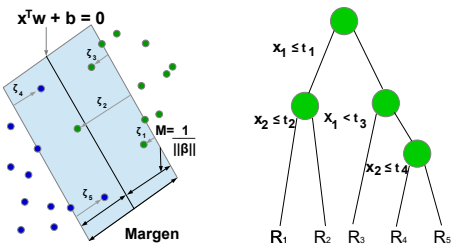
donde C es el factor de penalización y ζ_i es el factor de holgura. Se introducen variables de holgura positivas ζ_i para medir la distancia entre el margen y los vectores $\mathbf{x}_i$ que se encuentran en el lado equivocado del margen (ver Fig. 8a).

1.3.2. Boosting trees La idea detrás de los métodos *boosting* es combinar la salida de algunos clasificadores 'débiles', para generar un clasificador mas poderoso[19].

El método de árbol de decisión divide el conjunto de valores de las variables predictoras

[16]Yuan, Fang y col. «A Transformer Fault Diagnosis Model Based on Chemical Reaction Optimization and Twin Support Vector Machine». En: *Energies* 12.5 (2019), pág. 960.
[17]Ibíd.
[18]Ibíd.
[19]Bishop, *Pattern recognition and machine learning*, óp.cit.

Figura 7: Machine learning. (a) Se ilustra el caso en el cual se produce solapamiento de las muestras y los ζ_i que son tomados en cuenta al momento de la formulación del modelo. (b) Se muestra el ejemplo del método de árbol el cual a partir de las variables predictoras se divide en 5 regiones R_j



(a) (b)

en regiones desunidas R_j, j=1, 2, ..., J que representan los nodos terminales del árbol[20] (ver Fig.8b). Una constante γ_j es asignada a cada región y la regla predictiva es:

$$\mathbf{x} \in R_j \Rightarrow f(\mathbf{x}) = \gamma_j \tag{1.13}$$

Un arbol de desición puede ser formalmente formulado como:

$$T(\mathbf{x}; \Theta) = \sum_{j=1}^{J} \gamma_j I(\mathbf{x} \in R_j) \tag{1.14}$$

con parámetros $\Theta = \{R_j, \gamma_i\}_1^J$.

Ahora el modelo del boosting trees es la combinación de varios clasificadores como lo describe la ecuación 1.15.

$$f_m(\mathbf{x}) = \sum_{i=1}^{m} T(\mathbf{x}; \Theta) \tag{1.15}$$

[20]Ibíd.

2. EXPERIMENTOS Y RESULTADOS

En esta sección se mostrará los experimentos realizados para los tres métodos propuestos, método empírico, método basado en sub-espacios y *machine learning*, con los que se dio solución al problema de clasificación de las imágenes con artefactos. Para poder realizar una comparación entre los métodos, todos los experimentos fueron realizados haciendo uso de una validación cruzada de 5 *folds*. El desempeño de los métodos se midió en términos del área bajo la curva ROC[1] y el error de clasificación. A continuación se muestran los experimentos dentro de cada método.

2.1. MÉTODO EMPÍRICO

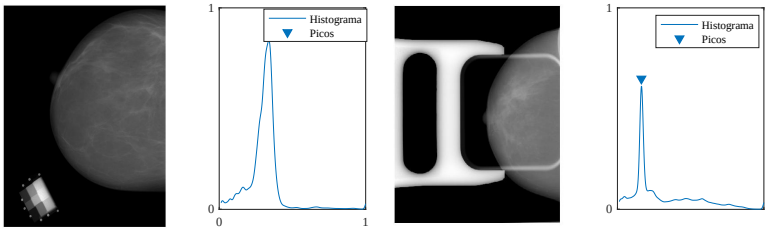
Con las características tomadas del histograma de las imágenes las cuales están descritas en la tabla 1, se hicieron algunos experimentos con el fin de encontrar las que permitan discriminar entre imágenes. Encontrando que, haciendo uso únicamente del número de picos se obtiene el mejor desempeño para este primer método, debido a que las demás características al momento de la clasificación no aportaban información discriminante. A continuación se realiza la descripción del experimento con el que se obtuvo el mejor desempeño.

Como se ilustra en la Fig. 2 los artefactos parecen tener tonalidades mas brillantes. Además como se puede observar en la Fig. 4 se esperaría una cantidad mayor de picos en el histograma de la imagen con artefactos. Basado en esto se realiza el experimento para corroborar si el número de picos que presente el histograma es un parámetro relevante para la clasificación de las imágenes. La forma como se realiza la clasificación se muestra en la siguiente ecuación:

$$\text{Clasificación}(\mathrm{x}_1) = \begin{cases} \text{si } x_1 > q & 1 \\ \text{otros casos} & 0 \end{cases} \tag{2.1}$$

[1] Hanley, James A y McNeil, Barbara J. «The meaning and use of the area under a receiver operating characteristic ROC curve.» En: *Radiology* 143.1 (1982), págs. 29-36.

Figura 8: Imágenes con y sin artefacto que presenta la misma cantidad de picos. (a) Mamografía FFDM sin artefacto. (b)Histograma de la imagen de la Fig. (a) la cual presenta un pico caracterizado (c) Mamografía FFDM con artefacto. (d) Histograma de la imagen de la Fig (c) con la misma cantidad de picos que la imagen sin artefacto de la Fig. (a).



(a) (b) (c) (d)

Donde q toma valores enteros de 0-12 y x_1 es el número de picos que contiene la imagen. Con los datos de entrenamiento y haciendo uso de la ecuación 2.1, se encuentra el valor q que presenta el menor error, el cual será el clasificador de los datos de prueba. Se encontró que para cada *fold* el valor de q con el mínimo error siempre fue de 2, lo que quiere decir que si una imagen presenta más de dos picos, será clasificada como imagen con artefacto.

Con el método empírico se logró obtener un error del 23,18 % en la clasificación de las imágenes y un área bajo la curva ROC (AUC) de 0.8.

Una muestra del por qué el algoritmo falla en la clasificación se ilustra en la Fig. 8, en la que se presentan dos imágenes, una con artefacto y otra sin artefacto con la misma cantidad de picos en el histograma.

2.2. MÉTODO DE SUB-ESPACIOS

Con la características extraídas del seno segmentado y de la máscara de cada una de las imágenes, se realizaron 3 experimentos en los cuales se implementaron los métodos enunciados para esta sección, junto con el modelo de regresión logística. Los experimen-

tos realizados fueron, regresión logística paso a paso, PCA junto con regresión logística (PCA+RL) y LDA junto con regresión logística (LDA+RL).

Para una visualización rápida de los resultados obtenidos en cada uno de los experimentos, en la tabla 3 se puede encontrar el porcentaje de error para cada uno de los experimentos y en la Fig. 11 sus respectivas curvas ROC y AUC. Además en la Fig. 10 se muestran las características más relevantes para los modelos RL paso a paso y LDA.

A Continuación se describe cada uno de los experimentos: RL paso a paso: Se implemento el modelo de regresión logística paso a paso el cual busca que el modelo contenga las características mas relevantes y con la menor redundancia con las que se pueda dar solución al problema. La selección de características paso a paso sobre todas las características posibles, se realiza comparando el rendimiento de la regresión con y sin la característica usando t-test[2]. En la Fig. 11a se muestran las variables que se mantuvieron y su respectiva relevancia para el modelo. Finalmente con este modelo se logró obtener un error de 6,63 % y con una AUC de 0.97.

PCA + RL: Para este experimento, inicialmente se hallan los componentes principales con los que se ajustará el modelo de regresión logística. Como se indicó en el planteamiento de PCA, se pueden obtener la misma cantidad de componentes principales como de variables iniciales. El reto ahora es determinar la cantidad de componentes que se van a tener en cuenta en el modelo de regresión logística. Como solución en cada iteración de la validación cruzada, se dividió los datos de entrenamiento que representan el 80 % de los datos totales en : entrenamiento (75 %) y validación (25 %).

La lógica del proceso fue la siguiente: Obtener los componentes principales, ajustar el modelo con un componente y probarlo en los datos de validación, seguidamente se agrega otro componente y se vuelve a probar en validación, continuando así hasta llegar al último componente. Finalmente la manera como se eligió la cantidad de componentes que tendría el modelo que se utilizará en los datos de prueba, fue revisando la gráfica Error vs Componentes y tomando como umbral, el número de componentes donde se encuentre el codo del gráfico. En la Fig. 9 se muestra una de las 5 gráficas que se

[2]Draper, Norman R y Smith, Harry. *Applied regression analysis*. Vol. 326. John Wiley & Sons, 1998.

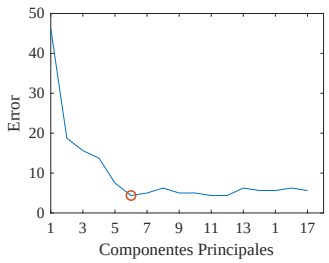
Tabla 3: Errores obtenidos para los modelos basados en sub-espacios. RL = regresión logística.

Método	Error
RL paso a paso	6.63 %
PCA + RL	7.39 %
LDA + RL	5.63 %

obtienen al realizar la validación cruzada, para escoger la cantidad de componentes.

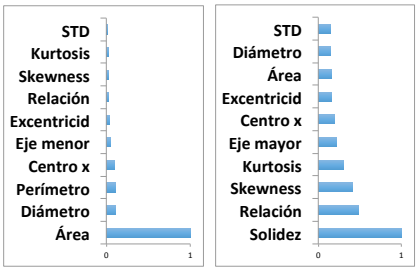
Para este método no se muestran las características más relevantes, debido a que para cada componente principal varían las características más importantes, lo cual es de esperarse por la forma como funciona el método. Para el modelo PCA+RL se obtuvo un error del 7.3 % con una AUC=0.95.

Figura 9: Selección de componentes principales



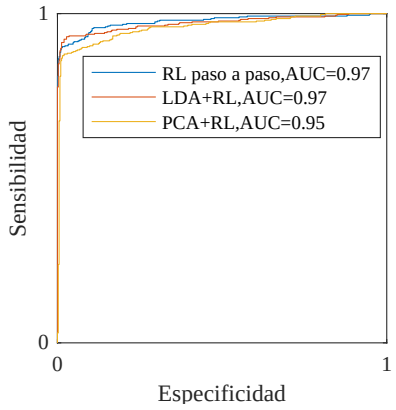
LDA+RL:Para este modelo como se mencionó en el planteamiento del método (LDA),

Figura 10: Características más relevantes. (a) Características que se mantuvieron en el modelo de regresión logística paso a paso. (b) Características mas relevantes al implementar LDA



(a) (b)

Figura 11: Curvas ROC para los métodos basados en sub-espacios



como se tienen dos clases se obtendrá un solo eje discriminante, con lo cual se pasa de un espacio de dimensión $\mathbb{R}^P$ a una dimensión, en donde para nuestro caso P=17. Una vez calculado el eje discriminante, este es utilizado en el modelo de regresión logística, obteniendo así el modelo de clasificación final.

En la Fig. 11b se muestra cuales son las características mas relevantes al implementar LDA, notando que existe diferencia entre las características más importantes dentro de cada método. Para este modelo se obtuvo un error del 5,63 % y con una AUC=0.97. Siendo este el modelo con el mejor desempeño dentro de la sección 2.2.

2.3. MACHINE LEARNING

Para esta sección una vez extraídas la características a usar (ver tabla 2), se ajustan los modelos SVM y *boosting trees*, los cuales fueron presentados en la sección 1.3.

En la tabla 4, se presentan los resultados con cada modelo. Para el caso de *boosting trees* fue necesario dividir los datos de la misma manera que se hizo en la sección 2.2 para el método PCA+RL con el objetivo de hallar la cantidad de clasificadores 'débiles' (árboles de decisión) que tendrá el modelo para cada *fold*. Finalmente con este modelo se logró obtener el menor error en la clasificación de las imágenes.

Tabla 4: Errores obtenidos con cada uno de los modelos estudiados de *machine learning*.

Método	Error
SVM	12.9 %
Boosting trees	4,5 %

3. CONCLUSIONES

En este documento se presentaron dos tipos de artefactos que tienen el potencial de alterar las medidas tomadas sobre el seno. Por ello, se propusieron 3 métodos para identificar dichas imágenes de forma automática.

Partiendo de un método empírico, se obtuvo un error en la clasificación del 23,18 %, haciendo uso del número de picos presentes en histograma de la imagen. Posteriormente se presentó un nuevo conjunto de características y se estudiaron dos metodologías (subespacios y *machine learning*), para mejorar el desempeño inicial. Para el método de subespacios, se presentaron 3 modelos diferentes (RL paso a paso, PCA+RL y LDA+RL), obteniendo el mejor desempeño con el modelo LDA+RL, el cual logró un error en la clasificación del 5,63 %. Finalmente para métodos basados en *machine learning*, se implementaron dos modelos (SVM y *Bossting trees*), obteniendo el mejor desempeño con *boosting trees*, con un error del 4,5 %, que fue el mejor resultado obtenido al momento de clasificar las imágenes.

Para futuros trabajos, una vez realizada la clasificación automática de las mamografías que contengan artefactos, es necesario adecuar o proponer métodos para la correcta segmentación de dichas imágenes.

BIBLIOGRAFÍA

BELHUMEUR, Peter N; HESPANHA, Joao P y KRIEGMAN, David J. *Recognition Using Class Specific Linear Projection*. 1997.

BISHOP, Christopher M. *Pattern recognition and machine learning*. springer, 2006.

DRAPER, Norman R y SMITH, Harry. *Applied regression analysis*. Vol. 326. John Wiley & Sons, 1998.

DROMAIN, C; BOYER, B; FERRE, R; CANALE, S; DELALOGE, S y BALLEYGUIER, C. «Computed-aided diagnosis CAD in the detection of breast cancer». En: *European journal of radiology* 82.3 (2013), págs. 417-423.

DUDA, Richard O; HART, Peter E y STORK, David G. *Pattern classification*. John Wiley & Sons, 2012.

GUERRERO, Rachael T Leon; NOVOTNY, Rachel; WILKENS, Lynne R; CHONG, Marie; WHITE, Kami K; SHVETSOV, Yurii B; BUYUM, Arielle; BADOWSKI, Grazyna y BLAS-LAGUAÑA, Michelle. «Risk factors for breast cancer in the breast cancer risk model study of Guam and Saipan». En: *Cancer epidemiology* 50 (2017), págs. 221-233.

HANLEY, James A y MCNEIL, Barbara J. «The meaning and use of the area under a receiver operating characteristic ROC curve.» En: *Radiology* 143.1 (1982), págs. 29-36.

IOSIFIDIS, Alexandros; TEFAS, Anastasios; NIKOLAIDIS, Nikolaos y PITAS, Ioannis. «Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis». En: *Computer Vision and Image Understanding* 116.3 (2012),

págs. 347-360. ISSN: 10773142. DOI: 10.1016/j.cviu.2011.08.008. URL: http://dx.doi.org/10.1016/j.cviu.2011.08.008.

IOSIFIDIS, Alexandros; TEFAS, Anastasios y PITAS, Ioannis. «Activity-based person identification using fuzzy representation and discriminant learning». En: *IEEE Transactions on Information Forensics and Security* 7.2 (2012), págs. 530-542.

ØSTERÅS, Bjørn Helge; MARTINSEN, Anne Catrine T; BRANDAL, Siri Helene B; CHAUDHRY, Khalida Nasreen; EBEN, Ellen; HAAKENAASEN, Unni; FALK, Ragnhild Sørum y SKAANE, Per. «BI-RADS density classification from areometric and volumetric automatic breast density measurements». En: *Academic radiology* 23.4 (2016), págs. 468-478.

PERTUZ, Said; TORRES, German F; TAMIMI, Rulla y KAMARAINEN, Joni. «Open Framework for Mammography-based Breast Cancer Risk Assessment». En: *IEEE-EMBS International Conference on Biomedical and Health Informatics* (2019). accepted, págs. 1-4.

QUINTANA, C; REDONDO, M y TIRAO, G. «Implementation of several mathematical algorithms to breast tissue density classification». En: *Radiation Physics and Chemistry* 95 (2014), págs. 261-263.

RAMPUN, Andrik; MORROW, Philip J; SCOTNEY, Bryan W y WINDER, John. «Fully automated breast boundary and pectoral muscle segmentation in mammograms». En: *Artificial intelligence in medicine* 79 (2017), págs. 28-41.

SHANG, Han Lin. «A survey of functional principal component analysis». En: *AStA Advances in Statistical Analysis* 98.2 (2014), págs. 121-142.

SOCIAL, Ministerio de salud y protección. *Cáncer de mama, una enfermedad en ascenso en Colombia*. https://www.minsalud.gov.co/Paginas/-Cancer-de-mama,-una-enfermedad-en-ascenso-en-Colombia.aspx. cosultado:11.05.2019. 2014.

XI, Xiaoming; SHI, Hao; HAN, Lingyan; WANG, Tingwen; DING, Hong Yu; ZHANG, Guang; TANG, Yuchun y YIN, Yilong. «Breast tumor segmentation with prior knowledge learning». En: *Neurocomputing* 237 (2017), págs. 145-157.

YAN, Shuicheng; XU, Dong; ZHANG, Benyu; ZHANG, Hong-Jiang; YANG, Qiang y LIN, Stephen. «Graph embedding and extensions: A general framework for dimensionality reduction». En: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1 (2007), págs. 40-51.

YUAN, Fang; GUO, Jiang; XIAO, Zhihuai; ZENG, Bing; ZHU, Wenqiang y HUANG, Sixu. «A Transformer Fault Diagnosis Model Based on Chemical Reaction Optimization and Twin Support Vector Machine». En: *Energies* 12.5 (2019), pág. 960.