COMPARACIÓN DE MÉTODOS DE ANÁLISIS ESPECTRAL DE MAMOGRAFÍAS PROCESADAS PARA LA EVALUACIÓN DE RIESGO DE CÁNCER DE SENO

JUAN DAVID ANGARITA PINZÓN

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2020

COMPARACIÓN DE MÉTODOS DE ANÁLISIS ESPECTRAL DE MAMOGRAFÍAS PROCESADAS PARA LA EVALUACIÓN DE RIESGO DE CÁNCER DE SENO

JUAN DAVID ANGARITA PINZÓN

Trabajo de Grado para optar al título de
Ingeniero Electrónico

Director
Said Pertuz
Ph.D.

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2020

DEDICATORIA

Dedico este trabajo a Dios, que me ha permitido lograr esta meta.

A mis padres, Martha y Pedro, mi hermano y mis abuelas por su apoyo total e incondicional.

A Sandra, por ser un gran soporte para mí en todo momento.

A mis amigos y todas aquellas personas que me han apoyado y animado en este proceso.

CONTENIDO

pág.

INTRODUCCIÓN **10**

1. OBJETIVOS **15**

2. MATERIALES Y METODOLOGÍA **16**

2.1. BASE DE DATOS 17

2.2. MÉTODOS BASADOS EN DIMENSIÓN FRACTAL 19

2.2.1. Box-Counting 19

2.2.2. Dimensión global de Minkowski 21

2.3. METODOS BASADOS EN ENERGÍAS DE LAS BANDAS ESPECTRALES 22

2.3.1. Preprocesamiento de la imagen 23

2.3.2. Análisis de anillos 25

2.4. ANÁLISIS DISCRIMINANTE LINEAL 27

2.5. SELECCIÓN DE CARACTERÍSTICAS 29

2.5.1. Prueba T y valores p 30

2.5.2. Análisis de componentes vecinas 36

2.5.3. Selección secuencial de características 39

3. EXPERIMENTOS Y RESULTADOS **41**

3.1. MÉTODOS BASADOS EN DIMENSIÓN FRACTAL 41

3.2. MÉTODOS BASADOS EN ENERGÍA DE LAS BANDAS ESPECTRALES 43

3.3. COMPARACIÓN DE RENDIMIENTOS 48

4. CONCLUSIONES **52**

BIBLIOGRAFÍA **54**

LISTA DE FIGURAS

pág.

Figura 1. Método de Box-Counting. 20

Figura 2. Método de dimensión de Minkowski 21

Figura 3. ROIs utilizadas 24

Figura 4. Transformada de Fourier de una imagen 26

Figura 5. Curvas de distribución t 33

Figura 6. Regiones de la distribución t 34

Figura 7. Tabla de valores p 34

Figura 8. Curvas ROC de los métodos de Box-Counting y LDA - Box-Counting 42

Figura 9. Curvas ROC de los métodos de dimensión de Minkowski y LDA - Dimensión de Minkowski 43

Figura 10. Comparación de tendencias de valores p 46

Figura 11. Curvas ROC de los métodos de selección de características para las potencias P_r+p_r 49

Figura 12. Curvas ROC de los métodos de selección de características para las potencias P_r 49

Figura 13. Curvas ROC de los métodos de selección de características para las potencias p_r 50

LISTA DE TABLAS

pág.

Tabla 1. Rendimiento de los métodos basados en dimensión fractal 41

Tabla 2. Rendimiento de los métodos basados en dimensión fractal con LDA 42

Tabla 3. Rendimiento de los métodos basados en energías de las bandas espectrales sin selección de características 44

Tabla 4. Rendimiento de los métodos basados en energías de las bandas espectrales con selección de características 47

Tabla 5. Rendimiento de los métodos basados en energías de las bandas espectrales con NCA 47

Tabla 6. Rendimiento de los métodos basados en energías de las bandas espectrales con selección secuencial 48

Tabla 7. Características seleccionadas por lo métodos NCA y selección secuencial 48

Tabla 8. Recopilación del rendimiento de los métodos 51

RESUMEN

TÍTULO: COMPARACIÓN DE MÉTODOS DE ANÁLISIS ESPECTRAL DE MAMOGRAFÍAS PROCESADAS PARA LA EVALUACIÓN DE RIESGO DE CÁNCER DE SENO [*]

AUTOR: JUAN DAVID ANGARITA PINZÓN [**]

PALABRAS CLAVE: ANÁLISIS ESPECTRAL, CÁNCER DE SENO, EVALUACIÓN DE RIESGO, MAMOGRAFÍAS, PROCESAMIENTO DE IMÁGENES.

DESCRIPCIÓN:

El cáncer de seno es un problema que afecta a millones de mujeres alrededor del mundo. Para atacar este problema, muchas investigaciones apuntan a que el uso de procesamiento digital de imágenes (DIP por sus siglas en inglés, Digital Image Processing) mamográficas es una herramienta poderosa.

El análisis espectral, presentado como un método de DIP, ha mostrado, según algunas investigaciones, resultados importantes y relevantes en la evaluación de riesgo de cáncer de seno. Sin embargo, a pesar de que en la literatura existen diversos métodos de análisis espectral, no existen trabajos comparativos. En este proyecto de investigación se propone estudiar, implementar y comparar algunos de estos métodos para determinar cuál puede ofrecer un mejor resultado en el procesamiento de imágenes mamográficas para la evaluación de riesgo de cáncer de seno.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Said Pertuz, Ph.D.

ABSTRACT

TITLE: SPECTRAL ANALYSIS METHODS COMPARISON OF PROCESSED MAMMOGRAMS FOR BREAST CANCER RISK ASSESSMENT [*]

AUTHOR: JUAN DAVID ANGARITA PINZÓN [**]

KEYWORDS: BREAST CANCER, IMAGE PROCESSING, MAMMOGRAMS, RISK ASSESSMENT, SPECTRAL ANALYSIS.

DESCRIPTION:

Breast cancer is an issue that affects millions of women worldwide. In order to address this problem, much research suggests that the use of digital image processing (DIP) in mammograms is a powerful tool.

According to some research, the spectral analysis presented as a DIP method has shown important and relevant results in breast cancer risk assessment. Nevertheless, despite the fact that there are various spectral analysis methods in the literature, there are still no comparative works. In this research project, it is proposed to study, implement and compare some of these methods to determine which one can offer a better result in the mammographic image processing for breast cancer risk assessment.

---

[*] Bachelor Thesis

[**] Faculty of Physical-Mechanical Engineering; School of Electrical, Electronic and Telecommunications Engineering. Director: Said Pertuz, Ph.D.

INTRODUCCIÓN

El cáncer es la segunda mayor causa de muertes alrededor del mundo [1]. Entre los más conocidos tipos de cáncer se encuentra el cáncer de seno, el cual es el cáncer invasivo más común en mujeres [2]. En Colombia, el cáncer de seno fue el cáncer más diagnosticado y el tercero que más muertes produjo en el año 2018 [3]. Con el fin de reducir la tasa de mortalidad se han realizado estudios y se han propuesto investigaciones en dos importantes áreas: la detección temprana y la evaluación de riesgo.

El objetivo de la detección temprana es determinar si un paciente, que preferiblemente aún no muestra síntomas, tiene algún tumor. Adicionalmente, sirve para saber si el tumor es benigno o maligno, para que en cualquiera de los dos casos se le realice un adecuado seguimiento y en caso de ser maligno se lleve a cabo un tratamiento. Lo anterior es muy importante debido a que los tratamientos cuando se detecta un cáncer de manera temprana son más efectivos, además de ser menos

---

[1] WORLD HEALTH ORGANIZATION. *Cancer. Fact sheets*. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed: 27-07-2019.

[2] MEDICAL NEWS TODAY. *What to know about breast cancer.* https://www.medical- newstoday.com/articles/37136.php. Accessed: 27-07-2019.

[3] GLOBOCAN. “Fact-sheets Colombia 2018”. En: *International Agency For Research on Cancer* 380.Globocan (2019), págs. 2018-2019.

invasivos [4] [5] [6]. A lo largo de los últimos años, investigadores han encontrado diversos métodos o pruebas para poder conseguir una detección temprana para muchos tipos de cáncer [7].

La evaluación de riesgo consiste en establecer si el paciente sano puede llegar a padecer cáncer. A diferencia de la detección temprana, esta área no ha tenido muchas investigaciones que muestren métodos muy efectivos. Los modelos de riesgo más aceptados hasta la fecha, requieren de un especialista. Por ejemplo, el modelo de Gail [8] [9], el cual mediante información de la paciente tal como la edad, historial clínico, historial familiar, densidad del seno, entre otros [10], permite determinar a un especialista el nivel de riesgo de que la paciente desarrolle cáncer en el futuro. La

---

[4] UK TRIAL OF EARLY DETECTION OF BREAST CANCER GROUP. "First results on mortality reduction in the UK trial of early detection of breast cancer". En: *The Lancet* 332.8608 (1988), págs. 411-416. DOI: 10.1016/S0140-6736(88)90410-2.

[5] L. S. CAPLAN, B. L. WELLS y S. HAYNES. "Breast cancer screening among older racial/ethnic minorities and whites: barriers to early detection." En: *Journal of gerontology* 47 Spec No (1992), págs. 101-10.

[6] FE ALEXANDER y col. "14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening". En: *The Lancet* 353.9168 (1999), págs. 1903-1908. DOI: 10.1016/S0140-6736(98)07413-3.

[7] AMERICAN CANCER SOCIETY. *Breast Cancer Early Detection and Diagnosis* . https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html. Accessed: 27-07-2019.

[8] Mitchell H. GAIL y col. "Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer". En: *Journal of the National Cancer Institute* 91.21 (1999), págs. 1829-1846. DOI: 10.1093/jnci/91.21.1829.

[9] Joseph P. COSTANTINO y col. "Validation studies for models projecting the risk of invasive and total breast cancer incidence". En: *Journal of the National Cancer Institute* 91.18 (1999), págs. 1541-1548. DOI: 10.1093/jnci/91.18.1541.

[10] NATIONAL CANCER INSTITUTE. *Breast Cancer Risk Assessment Tool*. https://bcri- sktool.cancer.gov/. Accessed: 10-10-2019.

principal limitación de este método es su reducido poder predictivo. Esta limitación hace que el estudio y desarrollo de nuevos biomarcadores de cáncer de seno sea una importante necesidad.

Una alternativa para solucionar el problema previamente planteado es la evaluación de riesgo por medio del procesamiento digital de imágenes [11]. Para el caso específico del cáncer de seno, se analizan las mamografías para esta evaluación de riesgo. Aunque en este caso no existan muchas investigaciones al respecto, sí se pueden encontrar algunas alternativas que evalúan el riesgo de padecer cáncer de seno de una manera suficientemente validada y reproducible, tal como es el caso de OpenBreast v1.0 [12] [13]. En la literatura, diferentes trabajos sugieren que el análisis espectral es una herramienta muy poderosa en el procesamiento digital de imáge-

---

[11] Maryellen L. GIGER, Nico KARSSEMEIJER y Julia A. SCHNABEL. "Breast Image Analysis for Risk Assessment, Detection, Diagnosis, and Treatment of Cancer". En: *Annual Review of Biomedical Engineering* 15.1 (2013), págs. 327-357. DOI: 10.1146/annurev-bioeng-071812-152416.

[12] Said PERTUZ y col. "Open Framework for Mammography-based Breast Cancer Risk Assessment". En: *IEEE International Conference on Biomedical and Health Informatics* (2019).

[13] Said PERTUZ y col. "Clinical Evaluation of a Fully-automated Parenchymal Analysis Software for Breast Cancer Risk Assessment: a Pilot Study in a Finnish Sample". En: *European Journal of Radiology* (2019), pág. 108710. DOI: 10.1016/j.ejrad.2019.108710.

nes [14] [15] [16] [17] y más específicamente imágenes biomédicas [18] [19] [20] [21].

El análisis espectral es uno de los métodos de análisis de datos más usados en diversos campos de la ciencia, tales como la geofísica, astronomía, ingeniería, entre otros [22]. Este se puede entender como el barrido o examen que se realiza a través de todas las componentes de frecuencia de una señal con el fin de adquirir nueva información sobre esta [23]; para el caso específico de este proyecto de investigación,

---

[14] Andrea SILVETTI y Claudio DELRIEUX. "Análisis Multifractal Aplicado a Imágenes Médicas". En: (2010).

[15] Steven R. MEIKLE y col. "Parametric image reconstruction using spectral analysis of PET projection data". En: *Physics in Medicine and Biology* 43.3 (1998), págs. 651-666. DOI: 10.1088/0031-9155/43/3/016.

[16] F. J. GARCIA, M. J. TAYLOR y M. C. KELLEY. "Two-dimensional spectral analysis of mesospheric airglow image data". En: *Applied Optics* 36.29 (1997), pág. 7374. DOI: 10.1364/ao.36.007374.

[17] Cha ZHANG y Tsuhan CHEN. "Spectral analysis for sampling image-based rendering data". En: *IEEE Transactions on Circuits and Systems for Video Technology* 13.11 (2003), págs. 1038-1050. DOI: 10.1109/TCSVT.2003.817350.

[18] M. TSURUOKA, R. SHIBASAKI y S. MURAI. "Spectral analysis of standing balance using medical stereo images". En: *International Conference of the IEEE Engineering in Medicine and Biology Society. (Cat. No.97CH36136)*. Vol. 4. IEEE, págs. 1671-1674. DOI: 10.1109/IEMBS.1997.757041.

[19] P. ALJABAR, D. RUECKERT y W. R. CRUM. "Automated morphological analysis of magnetic resonance brain imaging using spectral analysis". En: *NeuroImage* 43.2 (2008), págs. 225-235. DOI: 10.1016/j.neuroimage.2008.07.055.

[20] J. Gallo VILLEGAS y J. FARBIARZ. "Análisis espectral de la variabilidad de la frecuencia cardiaca". En: *Iatreia* 12.2 (1999), págs. 61-71.

[21] Robert ALFANO. "Method and apparatus for detecting cancerous tissue using luminescence excitation spectra". En: ().

[22] Hongbin LI. *Spectral Analysis of Signals [Book Review]*. Vol. 24. 1. 2008, págs. 148-150. DOI: 10.1109/msp.2007.273066.

[23] Tomáš SVOBODA. "Frequency analysis in images 2D Fourier Transform". En: (2008).

señales bidimensionales, las imágenes mamográficas. Además, este análisis considera el inconveniente de determinar el contenido espectral de una serie finita de mediciones temporales [24] [25].

En la literatura, se han considerado diferentes métodos basados en el análisis espectral para la evaluación de riesgo. Estos métodos incluyen el cálculo de diferentes medidas a partir del espectro, tales como la dimensión fractal [26], o la energía de anillos concéntricos en el dominio de la frecuencia [27] [28]. Vale la pena aclarar que aunque los resultados de estos trabajos son prometedores, no hay trabajos previos que permitan comparar de forma objetiva y sistemática los diferentes métodos de análisis espectral. Por lo tanto, el objetivo de este trabajo comparativo es identificar las ventajas y desventajas de cada método.

---

[24] Don PERCIVAL. "Introduction to spectral analysis". En: *Applied Physics* (2003), págs. 1-8.

[25] J. N. RAYNER. "Spectral Analysis". En: *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2001, págs. 14861-14864. DOI: 10.1016/B0-08-043076-7/02514-6.

[26] Hui LI y col. "Fractal Analysis of Mammographic Parenchymal Patterns in Breast Cancer Risk Assessment". En: *Academic Radiology* 14.5 (2007), págs. 513-521. DOI: 10.1016/j.acra.2007.02.003.

[27] Erin E.E. FOWLER y col. "Generalized breast density metrics". En: *Physics in Medicine and Biology* 64.1 (2019). DOI: 10.1088/1361-6560/aaf307.

[28] John J. HEINE y Robert P. VELTHUIZEN. "Spectral analysis of full field digital mammography data". En: *Medical Physics* 29.5 (2002), págs. 647-661. DOI: 10.1118/1.1445410.

1. OBJETIVOS

Objetivo general

- Implementar y comparar el rendimiento de diferentes métodos de análisis espectral de imágenes mamográficas procesadas para la evaluación de riesgo de cáncer de seno.

Objetivos específicos

- Implementar dos métodos diferentes de análisis espectral para el análisis de imágenes mamográficas procesadas: el método de la dimensión fractal [26], y el método de la energía de las bandas espectrales [27].
- Diseñar un conjunto de experimentos para comparar diferentes técnicas de análisis espectral para la evaluación de riesgo de cáncer de seno.
- Evaluar el desempeño de las técnicas de análisis espectral implementadas.

2. MATERIALES Y METODOLOGÍA

En el presente proyecto de investigación se busca comparar el rendimiento (enfocado en la evaluación de riesgo de cáncer de seno) de dos metodologías de análisis espectral: los métodos basados en dimensión fractal [26] (sección 2.2) y los métodos basados en la energía de las bandas espectrales [27] (sección 2.3).

Para lo anterior, se utiliza la base de datos TAYS-I, la cual está conformada por 7710 imágenes mamográficas de 277 pacientes con cáncer de seno y 330 pacientes sanas (sección 2.1). Estas imágenes serán utilizadas como herramienta de comparación para medir diferencias espectrales de dichas mamografías entre pacientes sanas y enfermas. Además, en el presente trabajo, se utilizarán algunas técnicas con el objetivo de mejorar el rendimiento de los métodos expuestos. Estas técnicas son: el análisis discriminante lineal (sección 2.4) y la selección de características (sección 2.5).

La primera de las anteriores técnicas, se aplica en los métodos basados en la dimensión fractal (como se indica en [26]) con el fin de mejorar la clasificación y obtener un mejor rendimiento. Por otro lado, la segunda se aplica sobre los métodos basados en la energía de las bandas espectrales, debido a que de este método se extrae una gran cantidad de variables (factor que puede afectar el rendimiento [ver sección 2.5]), sugiriendo el uso de métodos de selección de características para mejorar el rendimiento. Los métodos de selección de características que se utilizan son: los valores p, el análisis de componentes vecinas y la selección secuencial de características.

2.1. BASE DE DATOS

Para abordar el problema en cuestión, se cuenta con una base de datos compuesta por 7710 imágenes mamográficas las cuales corresponden a 607 pacientes comprendidas por: 330 pacientes sanas (controles) y 277 pacientes diagnosticadas con cáncer de seno (casos). La base de datos fue reducida mediante un proceso de selección, donde se descartaron los casos que cumplían al menos uno de los siguientes criterios de exclusión [13]:

1. **Intervención o tumores malignos previos (n=50):** Estos casos fueron desestimados debido a que dichas situaciones suelen afectar el tejido parenquimatoso del seno, y con esto el actual estudio y sus resultados.

2. **Síntomas en el momento del cribado (n=30):** Estos casos fueron desestimados porque no se tiene claridad sobre la condición de la paciente. Además, se sugiere un manejo más profundo al respecto.

3. **Lesión en el seno contralateral (n=6):** Estos casos fueron desestimados debido a que el seno contralateral es frecuentemente utilizado en el análisis de imágenes mamográficas.

4. **No disponibilidad de la anterior ronda de cribado (n=41):** Estos casos fueron desestimados debido a que, al ser un problema de evaluación de riesgo, se utilizará la anterior ronda de cribado para la comparación entre casos y controles.

5. **Sistema mamográfico con poca disponibilidad (n=7):** En la base de datos original se encuentran imágenes provenientes de diferentes sistemas mamográficos. Sin embargo, debido a la poca disponibilidad de la mayoría de estos,

solo se tuvieron en cuenta dos: el MicroDose SI (Philips Healthcare) y el Senographe Essential (General Electrics).

El número total de casos desestimados fue N=134, con lo cual queda un total de 143 casos en estudio, de los cuales 112 son del sistema mamográfico de General Electrics y 31 de Philips. Cada uno de estos 143 casos fue emparejado con un único control con el fin de hacer las comparaciones entre las características de un paciente sano y uno enfermo. Este emparejamiento se realizó con base en los siguientes criterios:

- Mismo año de nacimiento.
- Mismo sistema mamográfico.
- Mismo año de cribado y día de cribado lo más cercano posible.

El proceso de selección anterior dio como resultado 286 pacientes conformadas por 143 casos y 143 controles con 4 vistas mamográficas cada una (medio-lateral oblicua [MLO] y cráneo-caudal [CC] de ambos senos). Sin embargo, en el actual estudio se utilizará solo una vista con el fin de eliminar el factor de variabilidad entre vistas. La vista seleccionada para cada uno de los casos corresponde a la vista CC del seno afectado, mientras que en los controles será la vista CC del mismo seno de su pareja. La elección de la vista CC se hace para evitar el gasto computacional que ejerce la separación de la pared pectoral en la vista MLO [29]. La región de interés (ROI) usada en los métodos es el cuadrado más grande posible dentro del seno [12].

---

[29] Oscar ARAQUE y col. “Selecting the mammographic-view for the parenchymal analysis-based breast cancer risk assessment”. En: *2019 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2019 - Proceedings* (2019). DOI: 10.1109/BHI.2019.8834461.

A continuación, se discutirán brevemente dos metodologías que han sido consideradas en la literatura para el análisis espectral de imágenes mamográficas para la evaluación de riesgo de cáncer de seno: los métodos basados en dimensión fractal, y los métodos basados en la energía de bandas espectrales.

2.2. MÉTODOS BASADOS EN DIMENSIÓN FRACTAL

En la literatura se conocen diversos métodos que utilizan la dimensión fractal para realizar el análisis espectral de imágenes médicas [26] [14]. La dimensión fractal es una medida que da idea sobre la manera en la que un fractal llena un espacio conforme este se va haciendo más fino. Lo anterior ha sido muy utilizado en el análisis de imágenes médicas y en el caso especial de las imágenes mamográficas para determinar características relacionadas a la textura del seno [30].

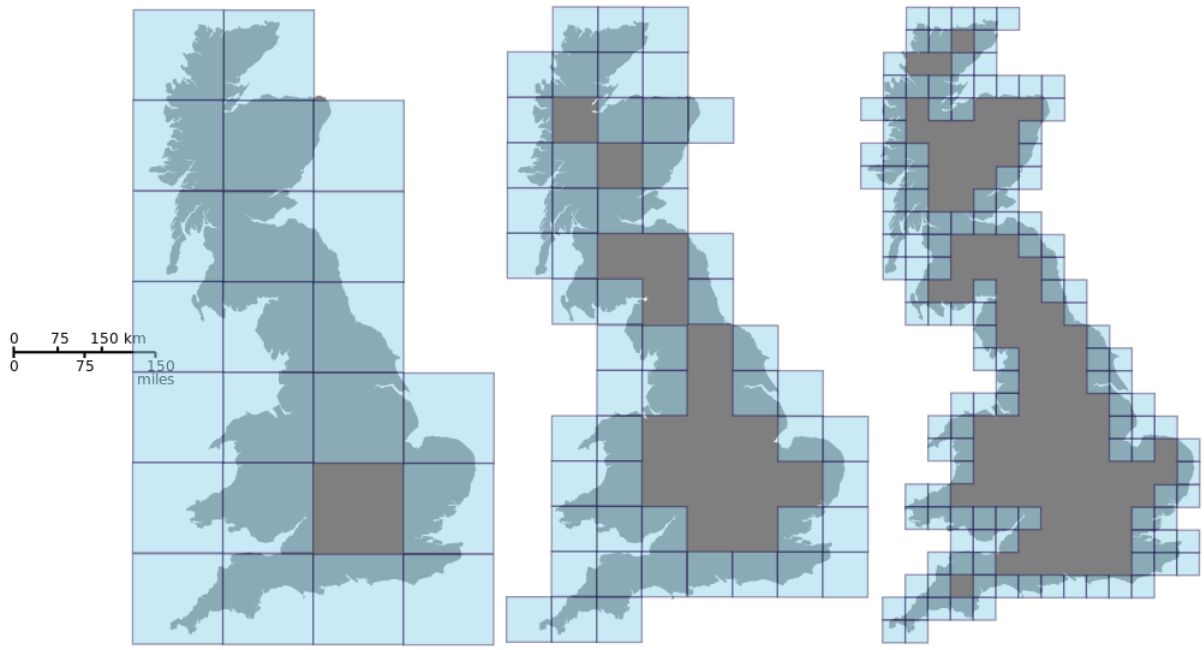
Este proyecto se fundamenta en el trabajo de investigación realizado por Li *et al* [26], el cual hace uso de la dimensión fractal, por medio de diferentes modelos, para hacer análisis de texturas en un seno. En este trabajo, se estudia el cálculo de la dimensión fractal por medio de dos diferentes métodos: Box-Counting y Dimensión global de Minkowski.

2.2.1. Box-Counting La técnica del Box-Counting consiste en distribuir 'cajas' de un tamaño determinado a lo largo y ancho de la imagen y contar el número de 'cajas' totales que la abarca en su totalidad, para después variar su respectivo

---

[30] A.N. MARANA y col. "Estimating crowd density with Minkowski fractal dimension". En: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. IEEE, 1999, 3521-3524 vol.6. DOI: 10.1109/ICASSP.1999.757602.

Figura 1. Método de Box-Counting aplicado sobre la costa de la isla de Gran Bretaña.



Fuente: Tomada de 31.

tamaño definido y repetir el conteo tal como se observa en la Fig. 1 (tomada de [31]) [32]. Este conteo de 'cajas' es la variable que se conoce como dimensión fractal. Su definición matemática viene dada por (1).

$$D_{BC} = 2 - \lim_{\varepsilon \to 0} \frac{log[A(\varepsilon)]}{log(\varepsilon)} \tag{1}$$

donde D_{BC} es la medida de la dimensión fractal y $A(\varepsilon)$ es el área superficial de la región de interés (ROI) en función del tamaño ε del pixel. Esta última variable se

---

[31] la enciclopedia libre WIKIPEDIA. *File: Great Britain Box.svg - Wikimedia Commons*. https://commons.wikimedia.org/wiki/File:Great_Britain_Box.svg. Accessed: 2020-07-03.

[32] Victor Hugo SALGADO LOAIZA. *Cálculo de la Dimensión Fractal mediante Box-Counting*.

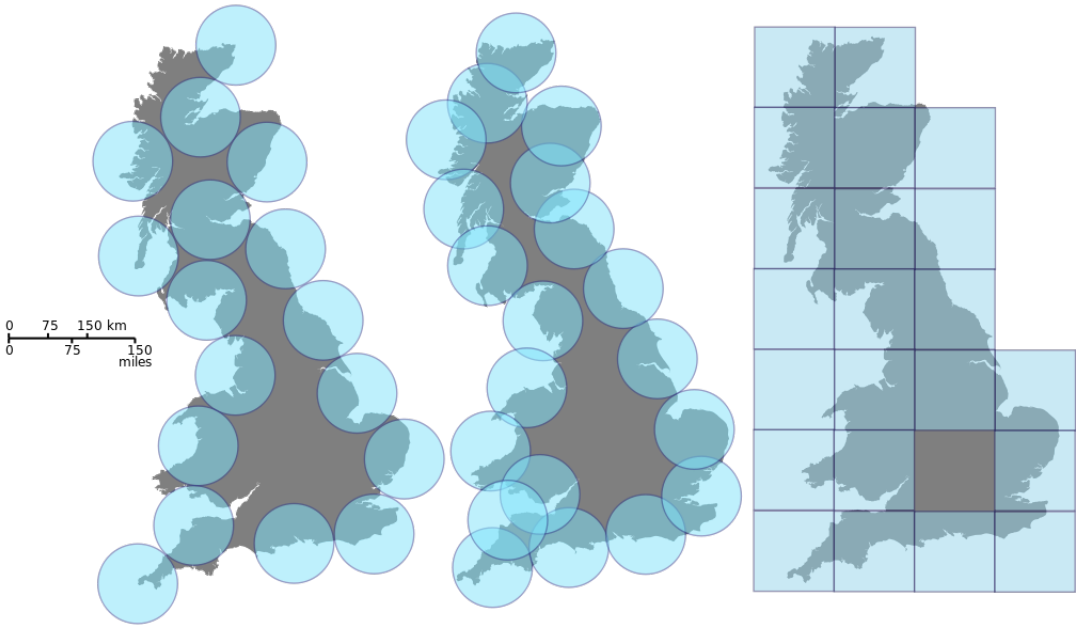
define además por medio de (2):

$$A(\varepsilon) = \sum_{x,y} \varepsilon[|i_\varepsilon(x,y) - i_\varepsilon(x,y+1)| + |i_\varepsilon(x,y) - i_\varepsilon(x+1,y)|] \quad (2)$$

donde $i_\varepsilon(x,y)$ es la intensidad en niveles de gris del pixel ubicado en (x,y).

2.2.2. Dimensión global de Minkowski La técnica de dimensión global de Minkowski es una generalidad de la técnica de Box-Counting, y de hecho tiene el mismo fundamento de calcular espacios llenados con objetos de un tamaño definido para luego variar su tamaño tal como se observa en la Fig. 2 (tomada de [33]).

Figura 2. Método de dimensión de Minkowski aplicado sobre la costa de la isla de Gran Bretaña.



Fuente: Tomada de 33

[33] la enciclopedia libre. WIKIPEDIA. *File: Great Britain coverings.svg - Wikimedia Commons*. https://commons.wikimedia.org/wiki/File:Great_Britain_coverings.svg. Accessed: 11-10-2019.

Su formulación matemática está definida por (3).

$$D_M(f) = \lim_{\epsilon \to 0} \frac{log[V_g(\epsilon)/\epsilon^3]}{log[1/\epsilon]} \tag{3}$$

donde $V_g(\epsilon)$ es el $"volumen"$ entre 2 versiones procesadas de la ROI en una imagen f usando operadores morfológicos de imágenes a una escala ϵ. Este se define matemáticamente por medio de (4).

$$V_g(\epsilon) = \sum_{i=1}^{256} \sum_{j=1}^{256} [(f \oplus \epsilon g) - (f \otimes \epsilon g)] \tag{4}$$

donde g es un elemento estructural usado por los operadores morfológicos y ϵ es la escala a la que se trabaja. Los operadores $\otimes$ y $\oplus$ corresponden a la erosión y dilatación morfológica en imágenes.

2.3. METODOS BASADOS EN ENERGÍAS DE LAS BANDAS ESPECTRALES

Al igual que en el caso de la dimensión fractal, existen investigaciones que dirigen a las energías de las bandas espectrales como factores asociados al riesgo de cáncer de seno. Una de las cuales fue desarrollada recientemente por Fowler *et al* [27]. En este método se plantea determinar el espectro de la imagen mamográfica y analizarlo por medio de regiones anulares y cada una de estas regiones va a representar una banda definida de frecuencias. Para después obtener la energía que un específico anillo posee y con base en esta realizar análisis de textura en el seno estudiado.

El método se segmenta en dos principales fases:

1. Preprocesamiento de la imagen.
2. Análisis de anillos.

2.3.1. Preprocesamiento de la imagen En esta primera fase del método, se busca obtener la representación de la energía en regiones anulares de la imagen mamográfica (por medio de bandas espectrales) para que esta última sea utilizada en el análisis de anillos, la cual es la parte vital de este método, con el fin de extraer el grupo total de características que se usarán en la evaluación de riesgo. Para realizar este preprocesamiento se deben seguir los siguientes pasos:

1. **Segmentación del seno:** Con el fin de mejorar el análisis, se debe realizar una segmentación donde la única área a estudiar pertenezca al seno. Eliminando así, mediante diversos algoritmos [12] [34], los demás elementos dentro de la imagen, por ejemplo el fondo, marcadores de vista, la pared pectoral (en caso de que la imagen sea MLO), entre otros.

2. **Extracción del área de interés (ROI):** Estudios sugieren que el uso de una ROI sobre la cual se realice el análisis (en lugar de la totalidad del seno) puede proveer mejoras significativas en la evaluación de riesgo de cáncer de seno [12].

   Como bien se comentó previamente, la ROI implementada en todos los métodos es la del cuadrado con el área más grande que pueda inscribirse en el seno.

3. **Reducción de márgenes:** Debido a que el factor humano puede afectar la toma de la mamografía, y más específicamente, la posición y forma en la que se coloca el seno, se realiza una reducción en las márgenes de la ROI seleccionada. Esto, con el fin de que las zonas donde el seno puede no estar

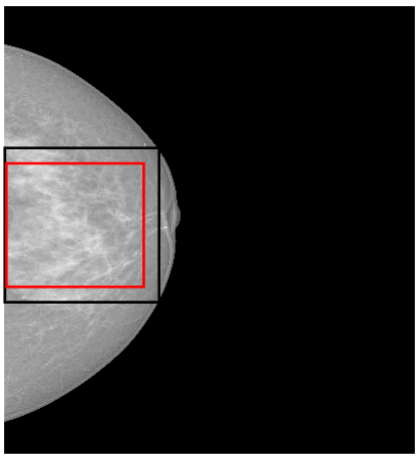
---

[34] German F. TORRES y col. "Morphological Area Gradient: System-independent Dense Tissue Segmentation in Mammography Images". En: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (2019), págs. 4855-4858. DOI: 10.1109/EMBC.2019.8857320.

uniformemente comprimido no se tengan en cuenta en el análisis posterior.

Estas reducciones corresponden al 10 % de ambas regiones laterales y la región anterior (la cara exterior del seno) de la ROI y un 1 % de la región posterior (la cara interior). Para un mejor entendimiento del resultado ver la Fig. 3.

Figura 3. ROIs utilizadas. La primer ROI obtenida (en negro) corresponde al cuadrado más grande inscrito en el seno. La ROI final (en rojo) evidencia la reducción donde las regiones laterales corresponden a la zona superior e inferior de la imagen, y las regiones anterior y posterior a la zona derecha e izquierda del seno respectivamente.



4. **Eliminación de la componente continua:** Una vez se tiene la imagen después de la reducción de márgenes en la ROI, se procede a eliminar la componente continua de la imagen. Lo anterior, reduciendo en toda la imagen el valor equivalente a la intensidad media de la imagen [35]. Esto último, tiene como intención que se resalten aquellos detalles más finos (como los patrones

[35] E. Oran BRIGHAM. *The fast Fourier transform and its applications*. 1988.

en el tejido parenquimatoso) y por lo tanto se mejore el análisis [36].

5. **Enventanado Hanning:** Buscando que la energía de la señal no se pierda en las frecuencias más lejanas, se le aplica a la señal un enventanado usando la ventana Hanning para evitar la fuga espectral [35] [37].

2.3.2. Análisis de anillos En esta fase se procede a realizar la transformada de Fourier de la imagen, la cual se descompondrá en una serie de anillos concéntricos de radio definido, cuyo ancho representa el rango de frecuencias que cada anillo aborda, tal como se observa en la Fig. 4 (replicada de [27]).

En esta se puede ver que el ancho de banda de los anillos está definido y viene dado por la ecuación (5).

$$\varepsilon = \frac{f_c}{n} \tag{5}$$

donde n es el número de anillos con los que se va a trabajar y f_c es la frecuencia espacial máxima del espectro, la cual viene descrita matemáticamente por (6)

$$f_c = \frac{1}{2\Delta} \tag{6}$$

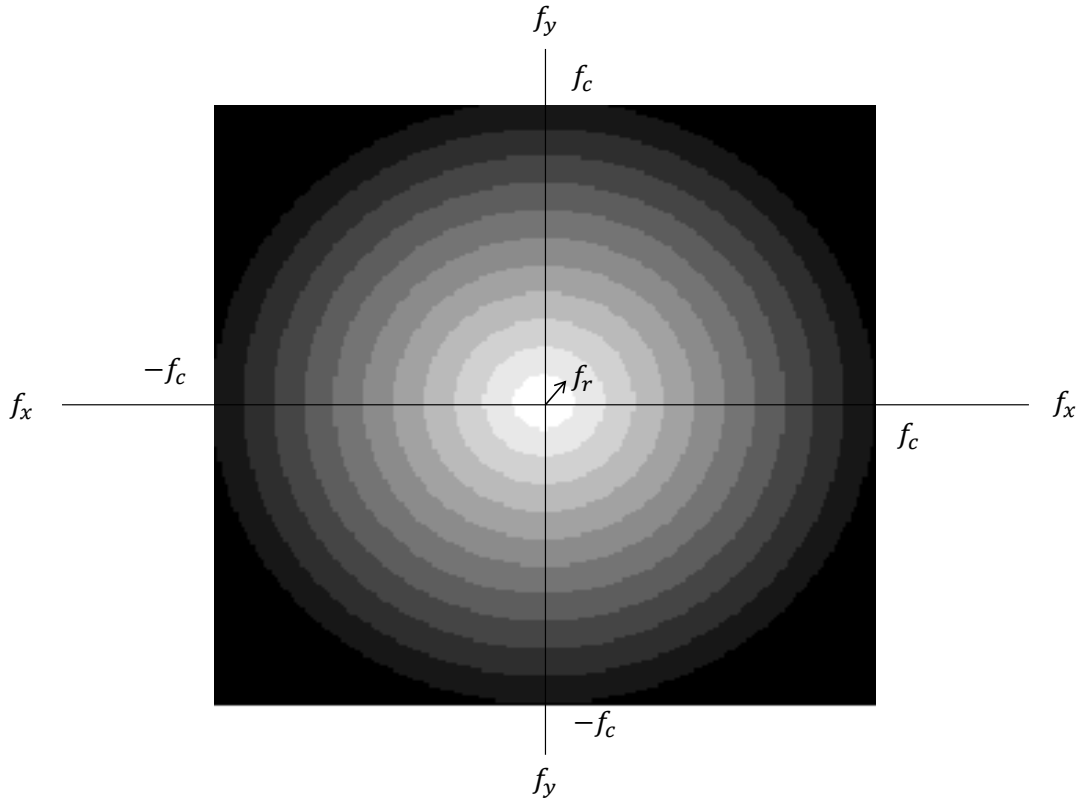
donde Δ es el tamaño del cuadro que cada pixel representa, es decir, la resolución, la cual depende completamente del sistema mamográfico que se esté usando.

---

[36] Rafael C. GONZALES y Richard E. WOODS. “Digital Image Processing”. En: *Computer* 7.5 (1974), págs. 17-19. DOI: 10.1109/MC.1974.6323522.

[37] UNIVERSIDAD DE LA REPÚBLICA DE URUGUAY. *Introducción a la Teoría del Procesamiento Digital de Señales de Audio*. https://www.eumus.edu.uy/eme/ensenanza/electivas/dsp/presentaciones/clase06.pdf. Accessed: 2020-07-03.

Figura 4. Transformada de Fourier de una imagen.



Fuente: Replicada de 27

Una vez se obtiene el espectro de la imagen con las características previamente mencionadas, se determina que según los estudios en [27] los anillos de menor frecuencia están relacionados con las texturas más toscas y gruesas del seno, las cuales son las texturas asociadas al riesgo de cáncer de seno.

Según lo anterior, se obtienen la potencia promedio del primer anillo y el promedio de las potencias promedio de los anillos 16 a 34. Además también se obtienen una potencia normalizada a la información del paciente, esta con el fin de que, en el caso de que haya imágenes de distintas fuentes, la medida no depende de la fuente

o sistema sino solo del paciente. Esta potencia normalizada viene determinada por (7).

$$p_r = \frac{P_r}{P_T + P_c} \tag{7}$$

donde p_r es la potencia normalizada del anillo r, P_r es la potencia promedio del anillo r, P_T es la potencia promedio total en todos los anillos excepto el central y P_c es la potencia promedio de las esquinas.

De esta potencia normalizada se obtiene su correspondiente al primer anillo y el promedio de los anillos 5 a 60. Con estas mediciones de potencias se busca medir la asociación entre la energía de determinadas bandas de frecuencia y el riesgo de cáncer de seno [27].

2.4. ANÁLISIS DISCRIMINANTE LINEAL

El análisis discriminante lineal (LDA por sus siglas en inglés Linear Discriminant Analysis) es una técnica de clasificación supervisada que se utiliza comúnmente en $machine\ learning$ [38] [39]. Esta técnica es utilizada con el fin de maximizar la covarianza intergrupal y minimizar la covarianza intragrupal [39] [40].

Para realizar este procedimiento, se tiene el vector de características A y un vector

---

[38] Ming LI y Baozong YUAN. "2D-LDA: A statistical linear discriminant analysis for image matrix". En: (2004). DOI: 10.1016/j.patrec.2004.09.007.

[39] the free encyclopedia WIKIPEDIA. *Linear discriminant analysis*. https://en.wikipedia.org/ wiki/Linear_discriminant_analysis. Accessed: 2020-07-03.

[40] S. BALAKRISHNAMA y A. GANAPATHIRAJU. "Linear Discriminant Analysis - a Brief Tutorial". En: *Compute* October (2015).

columna x (este es el usado para hacer las variaciones comentadas previamente sobre las covarianzas) [38] [41]. Así, A es proyectado sobre x según la transformación lineal mostrada en (8).

$$y = Ax \tag{8}$$

Las covarianzas, tanto intergrupal (S_B) como intragrupal (S_W), se representan mediante (9) y (10).

$$S_B = \sum_{c=1}^{2} n_c(\overline{A^c} - \overline{A})(\overline{A^c} - \overline{A})^T \tag{9}$$

$$S_W = \sum_{i=1}^{L}(A_i - A^{ci})(A_i - A^{ci})^T \tag{10}$$

donde n_c es el número de imágenes en la clase c, $\overline{A^c}$ el vector de medias que corresponden a la clase c, A^{ci} el vector de medias que corresponde a la clase a la que pertenece A_i y L el número de imágenes [42] [38].

Una vez se tienen estas covarianzas, se busca escoger el vector x que optimice el análisis discriminante. Así, se introduce la función criterio mostrada en (11).

$$J(x) = \frac{x^T S_B x}{x^T S_W x} \tag{11}$$

El x óptimo se encuentra cuando la covarianza intergrupal toma un valor grande y la intragrupal uno bajo. Según esto, el valor óptimo para lograr el objetivo del análisis

---

[41] Peter N. BELHUMEUR, Joao P. HESPANHA y David J. KRIEGMAN. “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”. En: 19.7 (1997), pág. 711.

[42] Joaquín Amat RODRIGO. “Análisis discriminante lineal (LDA) y Análisis discriminante cuadrático (QDA)”. En: (2016).

se logra en el valor máximo de (11) [43]. Para esto, se deriva la función con respecto a x y se iguala a 0 (cero). Dando como resultado la función mostrada en (12).

$$S_B x_{opt} = \lambda S_W x_{opt} \tag{12}$$

donde λ es el máximo eigenvalor de $S_W^{-1} S_B$.

En el trabajo realizado en [26] se utiliza esta técnica en las características extraídas del método de Box-Counting, donde utiliza las 8 áreas superficiales encontradas como variables de entrada, y del método de la dimensión de Minkowski, donde las variables de entrada corresponden a los 10 volúmenes calculados.

2.5. SELECCIÓN DE CARACTERÍSTICAS

La selección de características es el proceso mediante el cual se realiza una reducción de la cantidad de variables con el fin de construir modelos capaces de predecir de manera más efectiva [44] [45].

Esta reducción en las características es utilizada, frecuentemente, en los algoritmos de aprendizaje automático, minería de datos, entre otros. Lo anterior, debido a que los conjuntos de datos suelen dar mucha más información que la necesaria para generar el modelo, lo cual puede degradar la calidad de los patrones a detectar por

---

[43] Richard O. DUDA, Peter E. HART y David G. STORK. *Pattern classification. Second Edition*. 2000.

[44] Augusto PEREIRA GONZÁLEZ. “Selección de características para el reconocimiento de patrones con datos de alta dimensionalidad en fusión nuclear”. En: (2015).

[45] Radostina SPASOVA DIMITROVA. “Desarrollo y evaluación de métodos de selección de características para la predicción de eventos adversos en pacientes polimedicados”. En: (2017).

medio de variables redundantes, irrelevantes, ruidosas o correlacionadas [46] [47].

Algunas de las más comunes razones para realizar la selección de características a un conjunto de datos son las siguientes [46] [45]:

- Tiempo de entrenamiento más corto.
- Simplificación de los modelos.
- Evitar el *overfitting*, que es la consecuencia por entrenar un modelo con una gran cantidad de variables, generando que este pierda capacidad predictiva ante una nueva entrada.

Existe una gran diversidad de métodos de selección de características. Sin embargo, en este documento se presentarán solo tres de ellos, los cuales se basan en los siguientes conceptos: Valores p, análisis de componentes vecinas y selección secuencial de características.

2.5.1. Prueba T y valores p Toda investigación requiere previamente de una hipótesis, la cual es la declaración de algo que el investigador sospecha [48]. Ahora, uno de los métodos más utilizados para probar si estas hipótesis son correctas en toda la población de estudio, y no es solo una mera acción del azar en la muestra,

---

[46] Ryan J. URBANOWICZ y col. *Relief-based feature selection: Introduction and review*. 2018. DOI: 10.1016/j.jbi.2018.07.014. arXiv: 1711.08421.

[47] María Lucía VIOLINI. "Selección de Características. Su aplicación a Clasificación de Texturas." En: (2014).

[48] Richard ROYALL. *The Nature of Scientific Evidence*. 2004, págs. 119-152. DOI: 10.7208/chicago/9780226789583.003.0005.

es el valor p [49] [50].

Para esto, se definen las dos "$clasificaciones$" de hipótesis: la hipótesis de investigación o alternativa (H_1), que es de la cual se habló previamente, y la hipótesis nula (H_O), la cual es aquella que el investigador busca refutar o negar [51] [52]. Siguiendo con lo anterior, el valor p es la probabilidad de error de que al aceptar la hipótesis como verdadera esta sea falsa [50] [51]. Para hallar esta probabilidad se debe elegir un nivel de significancia α, que es el umbral a partir del cual un valor p acepta o rechaza una hipótesis.

Para realizar este procedimiento, se utiliza la prueba T (de la cual se hablará más adelante) para hallar el valor del t-stat y a partir de este hallar el valor p. Una vez con el valor p calculado correctamente se compara con el α y se realiza una de las siguientes afirmaciones [50]:

- Valor p <= α ; Se rechaza H_0, es decir, la probabilidad de error de que al aceptar la hipótesis H_1 como cierta esta sea falsa es pequeña respecto al umbral, y por lo tanto esta hipótesis sea válida.
- Valor p>α ; Se acepta H_0, es decir, que la probabilidad de error mencionada

---

[49] Enrico RIPAMONTI. "The use of p-values in applied research: Interpretation and new trends". En: *Statistica* 76.4 (2016), págs. 315-325. DOI: 10.6092/issn.1973-2201/6439.

[50] Juliana CARVALHO FERREIRA y Cecilia Maria PATINO. "What does the p value really mean?" En: *J Bras Pneumol* 41.5 (2015), págs. 485-485. DOI: 10.1590/S1806-37132015000000215.

[51] H. M. HUNG y col. "The Behavior of the P-Value When the Alternative Hypothesis is True". En: *Biometrics* 53.1 (1997), pág. 11. DOI: 10.2307/2533093.

[52] Jason C. TRAVERS, Bryan G. COOK y Lysandra COOK. "Null Hypothesis Significance Testing and p Values". En: *Learning Disabilities Research and Practice* 32.4 (2017), págs. 208-215. DOI: 10.1111/ldrp.12147.

anteriormente es grande con respecto a α, y por lo tanto la hipótesis H_1 tenga que ser descartada.

Como se comentó anteriormente, uno de los principales pasos para hallar los valores p es el uso de la pruebas T, la cual es una prueba estadística que ayuda a determinar si existen diferencias significativas entre las medias de dos muestras que presenten una distribución normal pero que el tamaño muestral sea pequeño [53] [54].

Esta prueba, para muestras independientes, se aplica siguiendo la ecuación (13).

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_{x1}^2}{n_1} + \frac{S_{x2}^2}{n_2}}} \tag{13}$$

donde $\overline{x_1}$, S_{x1} y n_1 son , respectivamente, el valor medio, la desviación estándar y el tamaño muestral de la muestra 1, mientras que $\overline{x_2}$, S_{x2} y n_2 sus respectivas contrapartes para la muestra 2.

Como se observa en (13) el valor de t (conocido como t-stat) depende de las variaciones entre las medias y las desviaciones de cada muestras. Así, el t-stat será menor entre menor sea la variación de las medias de las muestras y será mayor entre mayor sean las desviaciones estándar.

Este valor de t-stat se ubicará en su respectiva curva de distribución t (dependiendo de los grados de libertad de la prueba) y se calculará, a partir de este punto en la

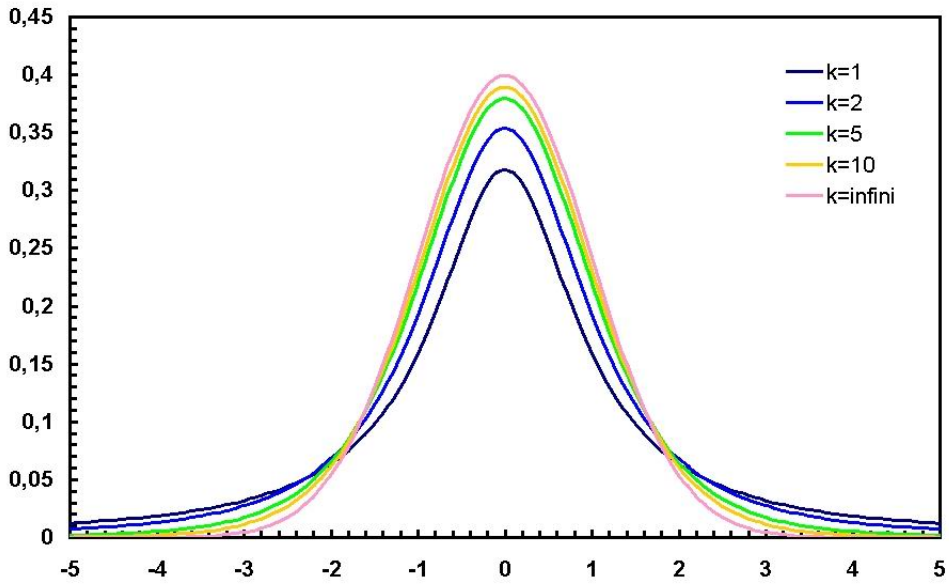
---

[53] Reinaldo Alberto SÁNCHEZ TURCIOS. “t-Student. Usos y abusos”. En: (2015).

[54] SCIENTIFIC EUROPEAN FEDERATION OF OSTEOPATHS. “Prueba "t"de Student”. En: (2019).

curva, el valor p correspondiente [55] [56].

Figura 5. Curvas de distribución t con 1, 2, 5, 10 e infinitos grados de libertad.



Fuente: Tomada de 57.

En la Fig. 5 (tomada de [57]) se observan algunas curvas de la distribución t, las cuales (como se mencionó anteriormente) dependen del grado de libertad. En la curva correspondiente se ubica el valor del t-stat con el fin de saber la región de dicha curva en la que se encuentra este. Regiones que se muestran en la Fig. 6 (replicada

---

[55] Zexun CHEN, Bo WANG y Alexander N. GORBAN. "Multivariate Gaussian and Student-t process regression for multi-output prediction". En: *Neural Computing and Applications* (2019). DOI: 10.1007/s00521-019-04687-8. arXiv: 1703.04455.

[56] Rodolfo RIVAS RUIZ, Marcela PEREZ RODRÍGUEZ y Juan O. TALAVERA. "Clinical research XV. From the clinical judgment to the statistical model. Difference between means. Student's t test." En: *Revista médica del Instituto Mexicano del Seguro Social* 51.3 (2013), págs. 300-303.

[57] la enciclopedia libre WIKIPEDIA. *File:Student densite best.JPG - Wikimedia Commons*. https://commons.wikimedia.org/wiki/File:Student_densite_best.JPG. Accessed: 2020-07-03.

Figura 6. Regiones de la distribución t.

Fuente: Replicada de 58.

de [58]), si el t-stat se ubica en la región de aceptación la hipótesis de investigación se validará como cierta, caso contrario a cuando se ubica en la región de rechazo.

Figura 7. Tabla de valores p hasta 10 grados de libertad y un t-stat hasta 0.02.

gl	Cola de probabilidad p						
	0.25	0.2	0.15	0.1	0.05	0.025	0.02
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359

Fuente: Replicada de 59.

También, se observa en la Fig. 7 (replicada de [59]) la tabla donde, a partir del t-stat

---

[58] José Gerardo MARTÍNEZ TOVAR. *Distribución "T" de Student*. https://estadisticaen-investigacion.wordpress.com/distribucion-t-de-student/. Accessed: 2020-07-03.

[59] KHAN ACADEMY. *El uso de una tabla para estimar el valor p del estadístico*

(columnas) y los grados de libertad (filas) se determina (o aproxima) el valor p del que se ha hablado previamente.

Una vez explicados estos conceptos, es necesario aclarar que los valores p pueden ser utilizados como método de selección de características. Lo anterior, hallando los valores p de un grupo de características que comparten dos muestras, y así las características cuyos valores p entren en la región de aceptación, serán las características seleccionadas para utilizar en el estudio.

Por ejemplo, imagine que a un grupo A (conformado por personas con antecedentes cardíacos clínicos) y un grupo B (conformado por personas sanas) se les extraen 20 características que, se sospecha, permiten analizar diferencias entre los dos grupos. Entonces, se hallan los valores p para cada una de las 20 características y se encuentra que solo las características 1, 5, 7 y 10 entraron en la zona de aceptación. Así, dichas características fueron seleccionadas para generar el modelo que se vaya a utilizar más adelante.

En el trabajo realizado en [27] se utilizó este método de selección de características. En este caso se utilizaron las 120 características extraídas (60 potencias y 60 potencias normalizadas) en cada paciente, donde para cada una de estas se les aplicaba la prueba t y se hallaban los valores p siendo el umbral $\alpha \sim 0.02$, el cual fue obtenido por ellos basados en la experiencia de trabajar con esta clase de problemas.

Otro concepto importante, y que será utilizado posteriormente, de este método de selección es la potencia estadística. Esta es la probabilidad de que al rechazar la

---

t. https://es.khanacademy.org/math/statistics-probability/significance-tests-one-sample/ tests-about-population-mean/v/calculating-p-value-from-t-statistic. Accessed: 2020-07-03.

hipótesis H_0 esta sea falsa, es decir, la probabilidad de no cometer falsos negativos [60] [61]. Este concepto suele ser muy utilizado para calcular el tamaño mínimo de la muestra o el tamaño del efecto necesario para que se detecte en un análisis [62] [63]. Normalmente, los investigadores consideran un buen valor de potencia estadística a uno que sea mayor o igual a 0.8 [64] [65]. Sin embargo, en muchos casos este valor no se alcanza y se estudian los factores que lo puedan afectar, los cuales son: El tamaño de la muestra, el nivel de significancia o la magnitud del efecto.

2.5.2. Análisis de componentes vecinas El análisis de componentes vecinas (NCA por sus siglas en inglés Neighborhood Component Analysis) es un método de selección de características que tiene como objetivo maximizar la precisión en los algoritmos de predicción. Lo anterior lo hace mediante el aprendizaje de un vector de pesos para cada característica, de tal manera que se maximice la exactitud de la clasificación leave-one-out (dejando uno afuera) [66].

---

[60] Martyn SHUTTLEWORTH y Lindsay WILSON. *Type I Error and Type II Error* . https://explorable.com/type-i-error. Accessed: 2020-07-03.

[61] J. NEYMAN y E. S. PEARSON. “The testing of statistical hypotheses in relation to probabilities a priori”. En: *Mathematical Proceedings of the Cambridge Philosophical Society* 29.4 (1933), págs. 492-510. DOI: 10.1017/S030500410001152X.

[62] Paul ELLIS. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, 2010, pág. 52.

[63] Brian S. EVERITT. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 2002, pág. 321.

[64] C.L. ABERSON. *Applied Power Analysis for the Behavioral Science*. 2010.

[65] the free encyclopedia WIKIPEDIA. *Power of a test*. https://en.wikipedia.org/wiki/Power_ of_a_test. Accessed: 2020-07-03.

[66] the free encyclopedia. WIKIPEDIA. *Neighbourhood components analysis*. https://en.wiki- pedia.org/wiki/Neighbourhood_components_analysis. Accessed: 2020-07-03.

Para esto, se aplicarán los conceptos mostrados en [67]. En primer lugar, se debe definir un conjunto de entrenamiento $T = \{(x_i, y_i), i = 1, 2, ..., n\}$, donde x_i corresponde al vector de características d-dimensional y y_i a sus respectivas etiquetas de clase.

Para llevar a cabo este procedimiento, se escoge aleatoriamente un punto de referencia, el cual será usado como el vecino más cercano de un nuevo punto x, dentro de T de tal manera que todos los puntos tengan cierta probabilidad de ser escogidos. La probabilidad de que un punto sea escogido como referencia es más alta si este está más cerca al nuevo punto.

La distancia, previamente mencionada, entre dos muestras x_i y x_j está definida por (14).

$$D_w(x_i, x_j) = \sum_{l=1}^{d} w_l^2 |x_{il} - x_{jl}| \tag{14}$$

donde w_l es el peso asociado a la característica número l.

Con el fin de mejorar la exactitud de la clasificación leave-one-out se utiliza una aproximación efectiva que dice que la elección del punto de referencia está determinada por una distribución de probabilidad. Así, la probabilidad de que un punto x_i seleccione a x_j como su punto de referencia está dado por (15).

$$p_{ij} = \frac{K(D_w(x_i, x_j))}{\sum_{k \neq i} K(D_w(x_i, x_k))} \tag{15}$$

---

[67] Wei YANG, Kuanquan WANG y Wangmeng ZUO. "Neighborhood component feature selection for high-dimensional data". En: *Journal of Computers* 7.1 (2012), págs. 162-168. DOI: 10.4304/jcp.7.1.161-168.

donde K es una función núcleo que asume valores grandes cuando D_w es pequeño.

Ahora, la probabilidad de que el punto x_i haya sido bien clasificado (según la clasificación leave-one-out y (15)) está dada por (16).

$$p_i = \sum_j y_{ij} p_{ij} \tag{16}$$

donde

$$y_{ij} = \begin{cases} 1 & si \quad y_i = y_j \\ 0 & si \quad y_i \neq y_j \end{cases}$$

Una vez teniendo la probabilidad p_i, se puede definir la exactitud de la clasificación leave-one-out según (17).

$$F(w) = \frac{1}{n} \sum_i p_i = \frac{1}{n} \sum_i \sum_j yijp_{ij} \tag{17}$$

Con la finalidad de mejorar el rendimiento de la selección de características se agrega un término adicional a (17), el cual introduce el parámetro de regulación λ. Así $F(w)$ se define finalmente como se muestra en (18).

$$F(w) = \frac{1}{n} \sum_i p_i = \frac{1}{n} \sum_i \sum_j yijp_{ij} - \lambda \sum_{l=1}^{d} w_l^2 \tag{18}$$

donde el parámetro de regulación λ se sintoniza mediante validación cruzada y hace que muchos pesos en w sean 0.

Debido a que la función definida en (18) es diferenciable, esta se deriva con la finalidad de buscar el vector w que maximice la función.

Finalmente, el vector resultante es aquel que se utiliza para la selección de carac-

terísticas. En este se tienen los pesos de cada una de ellas, y aquellas variables cuyos pesos tengan un nivel de significancia por encima de un umbral establecido serán las características seleccionadas.

2.5.3. Selección secuencial de características La selección secuencial de características es un método comúnmente utilizado para la reducción de variables [68]. Este se compone de dos principales aspectos: el primero es el criterio con el que se busca reducir la cantidad de características en el subconjunto de selección, y el segundo es el algoritmo de búsqueda secuencial que agrega o elimina características de un subconjunto mientras se está evaluando el criterio [69].

Este método posee dos diferentes variantes [70], las cuales son:

- **Selección secuencial directa:** Es aquella en la que se añaden características a un subconjunto candidato vacío hasta que el criterio disminuya.
- **Selección secuencial hacia atrás:** Es aquella en la que se eliminan características de un subconjunto candidato lleno hasta que el criterio incremente.

Una de las principales técnicas de selección secuencial de características es la regresión escalonada que consiste en una combinación entre la selección secuencial directa y la selección secuencial hacia atrás. Así, después de cada paso donde una

---

[68] Thomas RÜCKSTIESS, Christian OSENDORFER y Patrick VAN DER SMAGT. "Sequential Feature Selection for Classification". En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7106 LNAI. 2011, págs. 132-141. DOI: 10.1007/978-3-642-25832-9_14.

[69] F J FERRI y col. *Comparative Study of Techniques for Large-Scale Feature Selection*. Inf. téc.

[70] NCSS STATISCAL SOFTWARE. *Stepwise Regression*. https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf. Accessed: 2020-07-03.

variable se haya añadido al subconjunto candidato, este se analiza buscando si alguna de las características ha reducido su significancia en el modelo por debajo del criterio especificado [71].

Siguiendo con lo anterior, este proceso requiere de dos criterios: el primero para añadir variables y segundo para eliminarlas, donde el primero debe ser menor que el segundo para que el proceso no entre en un ciclo infinito.

---

[71] Leland WILKINSON. "Tests of significance in stepwise regression". En: *Psychological Bulletin* 86.1 (1979), págs. 168-174. DOI: 10.1037/0033-2909.86.1.168.

3. EXPERIMENTOS Y RESULTADOS

En esta sección, se mostrarán los experimentos utilizados para implementar los métodos basados en dimensión fractal y basados en energías de las bandas espectrales. Además, se mostrará la comparación entre los rendimientos de estos métodos. Para lo anterior, se utilizó la técnica de validación cruzada de 5 iteraciones y el desempeño de cada método fue medido en términos del área bajo la curva (AUC) ROC.

3.1. MÉTODOS BASADOS EN DIMENSIÓN FRACTAL

Con las características de cada una de las dos técnicas (Box-Counting y dimensión global de Minkowski) se realizaron 2 experimentos para medir sus rendimientos: el primero fue medir sus respectivos rendimientos utilizando un modelo de regresión logística, mientras que el segundo fue realizar el LDA junto a la regresión logística.

El primer experimento, después de replicar cada una de las dos técnicas de [26], consistió en pasar los datos por un modelo de regresión logística del cual se obtenía un vector de puntajes de predicción utilizando la validación cruzada, la cual divide los datos con un 80 % para entrenamiento y un 20 % para validación. Así, este vector de puntajes de predicción se comparó con la clasificación real de cada una de las variables y se obtuvo el AUC, además del intervalo de confianza del 95 %, de estas dos técnicas. En la tabla 1 se muestran los rendimientos de estas.

Tabla 1. Rendimiento de las técnicas propuestas por Li y col.

Método	AUC	95 % CI
Box-Counting	0.486	[0.42, 0.55]
Dimensión de Minkowski	0.500	[0.43, 0.57]

En el segundo experimento, el proceso consistió en aplicarle el LDA a los dos métodos. De modo que, para el método de Box-Counting, se aplica un LDA de ocho entradas (las ocho áreas superficiales extraídas de cada imagen). Mientras que, para el método de la dimensión de Minkowski, se aplica uno de 10 entradas (los 10 volúmenes extraídos de cada imagen). Donde en ambos casos, como solamente existen 2 clases, el discriminante sería proyectado en una dimensión, es decir, una única variable de salida. Los rendimientos de cada uno de los dos métodos con el LDA aplicado se muestran en la tabla 2.

Tabla 2. Rendimiento de las técnicas propuestas por Li y col. con LDA.

Método	AUC	95 % CI
LDA - Box-Counting	0.438	[0.37, 0.5]
LDA - Dimensión de Minkowski	0.454	[0.39, 0.52]

En las Fig. 8 y 9 se muestran las curvas ROC para el caso del Box-Counting y su variante de LDA (Fig. 8) y el caso de la dimensión de Minkowski con su variante de LDA (Fig. 9).

Figura 8. Curvas ROC de los métodos de Box-Counting y LDA - Box-Counting.

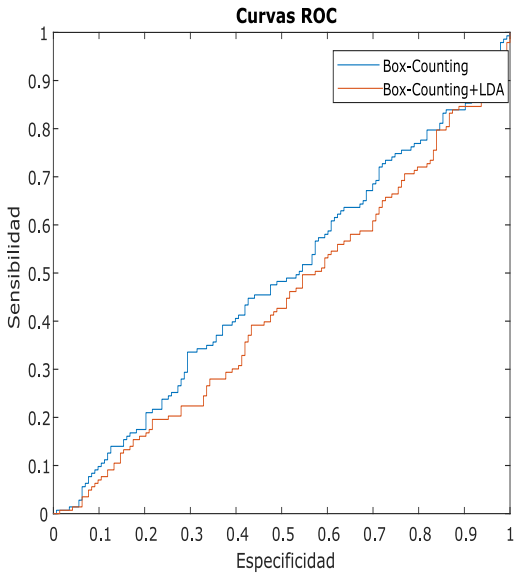
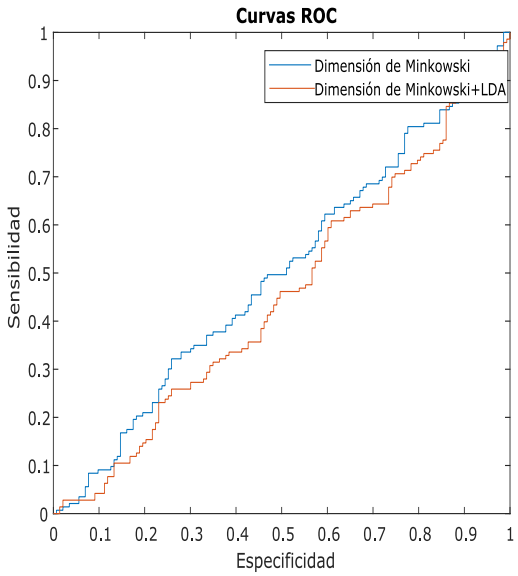


Figura 9. Curvas ROC de los métodos de dimensión de Minkowski y LDA - Dimensión de Minkowski.



En el caso de ambos métodos se evidencia, mediante la información de las Fig. 8 y 9 y las tablas 1 y 2, que el análisis discriminante lineal 'aleja' los rendimientos del peor de los casos ($AUC = 0.5$). Sin embargo, es necesario comprobar si el LDA ha tenido un efecto estadísticamente significativo en los rendimientos, de lo cual se hablará más adelante.

3.2. MÉTODOS BASADOS EN ENERGÍA DE LAS BANDAS ESPECTRALES

Con la finalidad de calcular el rendimiento del método propuesto por [27], del cual se obtenía un total de 120 características, se propusieron cuatro experimentos: el primero consistió en medir el rendimiento mediante un modelo de regresión logística utilizando todas las características extraídas. Para los siguientes experimentos se utilizaron los conceptos mostrados en la sección 2.5. Así, para el segundo experimento se realizó una selección de características por el método de los valores p antes de utilizar el modelo de regresión logística, para el tercer experimento se

utilizó la técnica de NCA junto con la regresión logística y finalmente, en el cuarto experimento se utilizó el método de selección secuencial ajustando los datos con un modelo de regresión logística escalonada.

Es necesario aclarar que, para todos los experimentos previamente mencionados se calculó el rendimiento para tres casos: el caso 1 donde se utilizaron solo las 60 potencias promedio (P_r) por imagen, el caso 2 donde se utilizaron las 60 potencias normalizadas (p_r) por imagen y el caso 3 donde se utilizaron las 120 potencias (P_r+p_r) por imagen.

En el primer experimento, los datos fueron ajustados mediante un modelo de regresión logística para cada uno de los 3 casos mencionados anteriormente. Los resultados de los rendimientos medidos en términos de AUC y sus respectivos intervalos de confianza se muestran en la tabla 3.

Tabla 3. Rendimiento de las técnicas propuestas por Fowler y col. sin selección de características.

Caso	AUC	95 % CI
Potencias P_r	0.528	[0.46, 0.6]
Potencias p_r	0.518	[0.45, 0.58]
Potencias $P_r + p_r$	0.519	[0.45, 0.59]

Para el caso del segundo experimento, se utilizó el método de valores p para seleccionar las características que utilizadas en el modelo de regresión logística. Para este método, se utilizó un umbral $\alpha \sim 0.02$, el cual fue escogido en base a la experiencia mostrada en [27].

En este experimento se presentó que, para la base de datos con la que se está realizando el estudio, los valores p no presentaron la suficiente significancia y ninguna

de las características entró en la región de aceptación. Esto se debe a la potencia estadística del análisis, ya que cuando existen tendencias débiles, estas requieren de una alta potencia estadística para obtener el tamaño del efecto necesario para que este sea evidenciado [64].

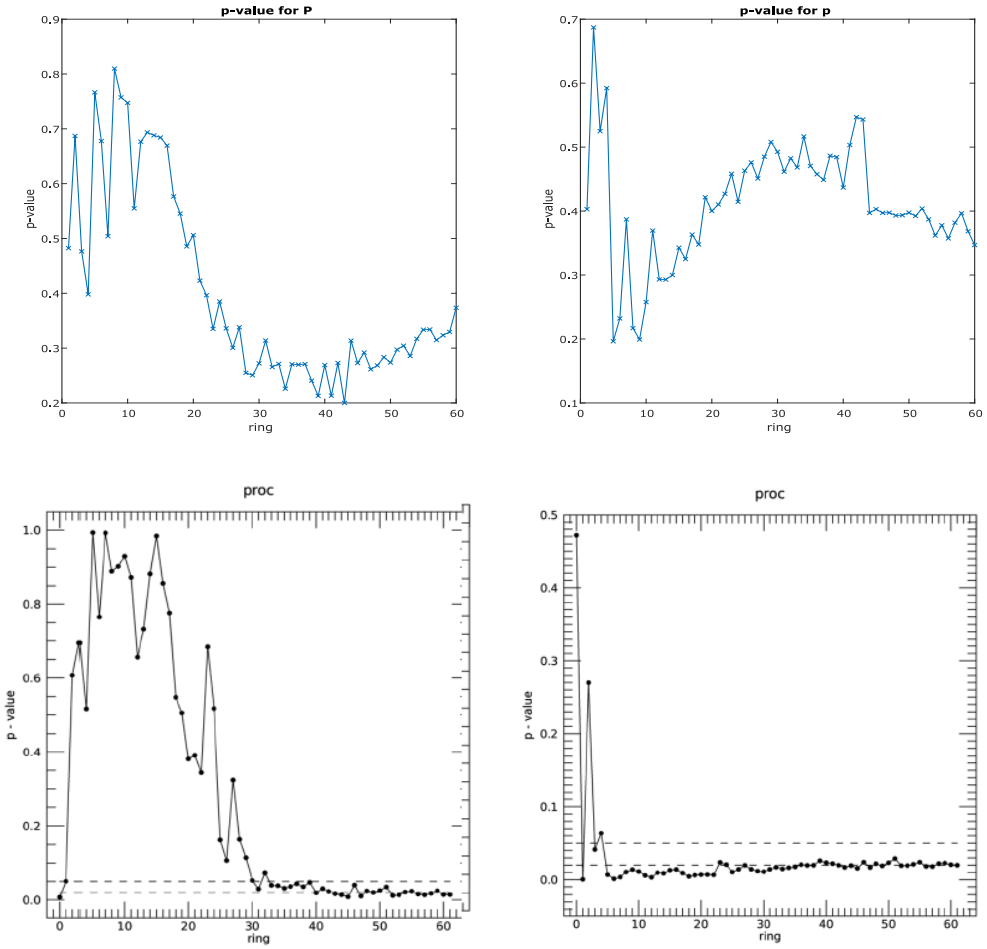
Como se comentó anteriormente, dos de los factores que pueden aumentar esta potencia estadística son el tamaño de la muestra y el nivel de significancia, los cuales no se pueden modificar para el presente estudio. El primero porque es un factor del cual no se tiene disponibilidad y el segundo porque modificar el nivel de significancia generaría que se pierda sentido en el análisis práctico de la predicción de cáncer de seno.

A pesar de lo anterior, se puede afirmar que las tendencias que presentan en [27] están presentes en el presente estudio. Esto se evidencia en la Fig. 10, donde muestra que las curvas de la zona izquierda (referentes a las potencias promedio P_r) poseen tendencias similares, mismo caso que en las curvas de la zona derecha (referentes a las potencias normalizadas p_r).

Ahora, con la finalidad de obtener características producto de este método, se utiliza la recomendación de [27], que se muestra en la subsección 2.3.2, donde se utilizan la potencia promedio correspondiente al primer anillo (P_1), el promedio de las potencias promedio de los anillos 16 a 34 (P_{16-34}), la potencia normalizada correspondiente al primer anillo (p_1) y el promedio de las potencias normalizadas de los anillos 5 a 60 (p_{5-60}). Los rendimientos, después de ajustar los datos con un modelo de regresión logística, se muestran en la tabla 4.

En el caso del tercer experimento, se utilizó la técnica de NCA como método de

Figura 10. Comparación de tendencias de valores p entre el presente estudio (zona superior de la imagen) y el estudio realizado en Fowler y col. (zona inferior de la imagen).



selección de características. Con la finalidad de encontrar el valor óptimo del parámetro de regulación λ se utilizó la validación cruzada de 5 iteraciones. Este valor óptimo fue aplicado en la ecuación 18 para obtener el vector de pesos definitivo con el cual realizar la selección de características.

Después de realizar la selección de características, se utiliza un modelo de regresión logística para ajustar los datos y se obtiene el rendimiento para cada uno de los tres

Tabla 4. Rendimiento de las técnicas propuestas por Fowler y col. con selección de características recomendada.

Caso	AUC	95 % CI
Selección recomendada - Potencias P_r	0.498	[0.43, 0.57]
Selección recomendada - Potencias p_r	0.466	[0.48, 0.53]
Selección recomendada - Potencias $P_r + p_r$	0.508	[0.44, 0.5]

casos que ya se mencionaron. En la tabla 5 se muestran los rendimientos con esta técnica aplicada.

Tabla 5. Rendimiento de las técnicas propuestas por Fowler y col. con NCA.

Caso	AUC	95 % CI
NCA - Potencias P_r	0.484	[0.42, 0.55]
NCA - Potencias p_r	0.446	[0.38, 0.51]
NCA - Potencias $P_r + p_r$	0.524	[0.46, 0.59]

En el caso del cuarto y último experimento, se utilizó el método de selección secuencial de características, mediante el cual se realizó la reducción de características y el ajuste de los datos con el modelo de regresión logística escalonada.

Para realizar esta técnica, se utilizó como criterio los cambios en el Criterio de Información de Akaike (AIC), el cual es un parámetro que compara la capacidad predictiva de un modelo con su complejidad, de tal manera que favorezca con una alta capacidad predictiva y castigue con una alta complejidad [72].

Después de ajustar los datos mediante un modelo de regresión logística escalonada, se calcula su respectivo rendimiento acompañado de su intervalo de confianza. Estos resultados se muestran en la tabla 6. Además, en la tabla 7 se muestran las

[72] Félix Francisco CABALLERO DÍAZ. *Selección de modelos mediante criterios de información en análisis factorial. Aspectos teóricos y computacionales*. 2011.

Tabla 6. Rendimiento de las técnicas propuestas por Fowler y col. con selección secuencial.

Caso	AUC	95 % CI
Secuencial - Potencias P_r	0.475	[0.41, 0.54]
Secuencial - Potencias p_r	0.503	[0.44, 0.57]
Secuencial - Potencias $P_r + p_r$	0.502	[0.44, 0.57]

características que, tanto el método de NCA como el de selección secuencial, se utilizaron para generar sus respectivos modelos.

Tabla 7. Características seleccionadas por lo métodos NCA y selección secuencial.

Caso	Características NCA	Características selección secuencial
Potencias P_r	P_1, P_5, P_9, P_{24}, P_{36}, P_{43}, P_{52}, P_{57}, P_{59}	P_{59}, P_{60}
Potencias p_r	p_{54}	p_1
Potencias $P_r + p_r$	P_3, P_5, P_{36}, P_{42}, p_7, p_{40}	P_{59}, P_{60}, p_1

Buscando mostrar cómo se comportan los rendimientos de los métodos para cada uno de los tres casos (potencias P_r, p_r y P_r+p_r), se muestra, mediante las Fig. 11, 12 y 13, las curvas ROC extraídas de los cuatro experimentos. Donde la Fig. 11 corresponde a las potencias P_r+p_r, la 12 a las potencias P_r y la Fig. 13 a las potencias p_r.

3.3. COMPARACIÓN DE RENDIMIENTOS

Una vez explicados cada uno de los experimentos aplicados para medir los rendimientos, tanto en los métodos basados en dimensión fractal como los basados en energía de las bandas espectrales, se busca hacer una comparación de cada uno de ellos. Para esto, en la tabla 8 se recopilan todos los rendimientos medidos en

Figura 11. Curvas ROC de los métodos de selección de características para las potencias P_r+p_r

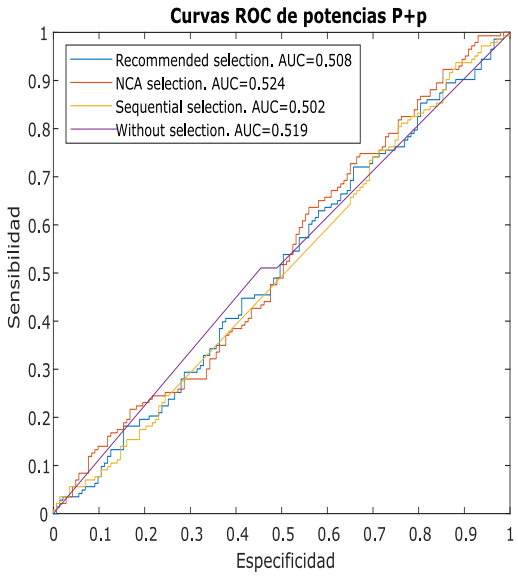
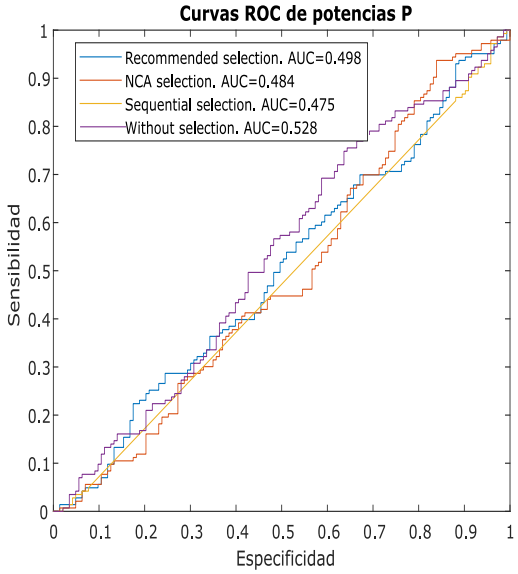


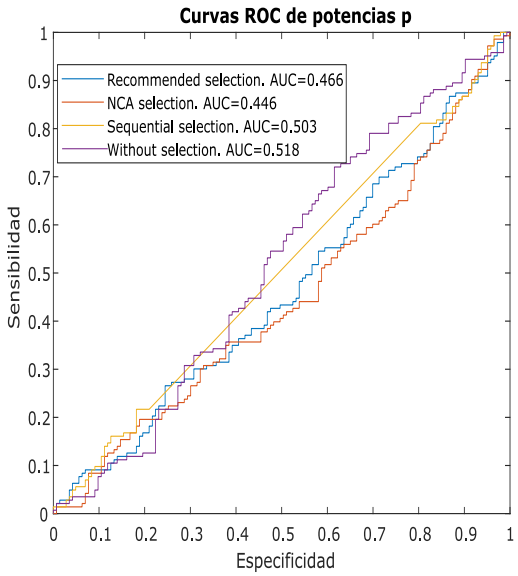
Figura 12. Curvas ROC de los métodos de selección de características para las potencias P_r



términos de AUC y cada uno de los intervalos de confianza de estos.

Con el fin de interpretar los resultados correctamente, es necesario aclarar que en

Figura 13. Curvas ROC de los métodos de selección de características para las potencias p_r



el caso de los valores de AUC menores a 0.5, las variables que intervienen en este proceso no poseen la suficiente fuerza predictiva. Sin embargo, se afirma que incluso estos rendimientos "mejoran" a medida que se encuentran alejados del valor de 0.5 (el peor de los casos). Con el fin de demostrar que tan significantes son estas supuestas mejoras y, en general, las diferencias entre los rendimientos, se realiza una comparación estadística entre estos.

Para comparar estos rendimientos, se utilizó la prueba de DeLong [73], de la cual se obtenía el valor p de cada par de métodos. En este caso, se utilizó un umbral α comúnmente utilizado en estos análisis estadísticos, el cual fue $\alpha = 0.05$.

---

[73] Elizabeth R. DELONG, David M. DELONG y Daniel L. CLARKE-PEARSON. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach". En: *Biometrics* 44.3 (1988), pág. 837. DOI: 10.2307/2531595.

Tabla 8. Recopilación del rendimiento de los métodos.

Método	AUC	95 % CI
Box-Counting	0.486	[0.42, 0.55]
LDA - Box-Counting	0.438	[0.37, 0.5]
Dimensión de Minkowski	0.500	[0.43, 0.57]
LDA - Dimensión de Minkowski	0.454	[0.39, 0.52]
Potencias P_r+p_r	0.519	[0.45, 0.59]
Potencias P_r	0.528	[0.46, 0.6]
Potencias p_r	0.518	[0.45, 0.58]
Recomendado - Potencias P_r+p_r	0.508	[0.44, 0.58]
Recomendado - Potencias P_r	0.498	[0.43, 0.57]
Recomendado - Potencias p_r	0.466	[0.4, 0.53]
NCA - Potencias P_r+p_r	0.524	[0.46, 0.59]
NCA - Potencias P_r	0.484	[0.42, 0.55]
NCA - Potencias p_r	0.446	[0.38, 0.51]
Secuencial - Potencias P_r+p_r	0.502	[0.44, 0.57]
Secuencial - Potencias P_r	0.475	[0.41, 0.54]
Secuencial - Potencias p_r	0.503	[0.44, 0.57]

Como resultado se obtuvo que solamente se presenta una diferencia estadísticamente significativa en la comparación del método de LDA - Box-Counting con el método sin selección de características de las potencias promedio P_r, con un $p = 0.0416$. Sin embargo, esto pierde relevancia debido a que no se presenta una diferencia significativa de este último método con respecto a los demás. Lo anterior, también se observa en la tabla 8, ya que los intervalos de confianza muestran que no existe un diferencia significativa y marcada para ningún método. Sugiriendo así, que no existe un método significativamente mejor que los demás.

4. CONCLUSIONES

En este trabajo, se presentó la comparación entre métodos de análisis espectral en el enfoque de la evaluación de riesgo de cáncer de seno. Estos métodos fueron divididos en dos principales clasificaciones, donde cada una de estas fue implementada mediante diferentes técnicas.

En primer lugar, para el caso de los métodos basados en dimensión fractal se observa que, si bien la aplicación del LDA supone una mejoría en el rendimiento, esta mejoría no es estadísticamente significativa y no propone un cambio real en la evaluación de riesgo de cáncer de seno mediante el análisis espectral.

En segundo lugar, para los métodos basados en energía de las bandas espectrales se analiza que para recrear los resultados obtenidos en [27] es necesario hacer uso de una base de datos más amplia, puesto que las débiles tendencias del método requieren de una cantidad alta de muestras (imágenes mamográficas). Además, se evidencia que a pesar del efecto positivo en el rendimiento que supondría el uso de técnicas de selección de características, estas no generan una mejora significativa.

Finalmente, es posible apreciar que, en el uso de únicamente las variables espectrales de la imagen, no se presenta variación significativa en la escogencia de un método u otro. Además, la presencia de únicamente variables espectrales hace que la evaluación de riesgo pierda fuerza (evidenciado en la tabla 8) debido a que en estos análisis suelen intervenir una cantidad mucho más alta de variables.

Por lo anterior, se sugiere, para futuros trabajos al respecto, realizar un estudio de métodos de análisis espectral teniendo en cuenta otros biomarcadores de imagen

enfocados en la evaluación del riesgo de cáncer de seno. Así mismo, en caso de ser posible, utilizar una base de datos más amplia que la mostrada en el presente trabajo.

BIBLIOGRAFÍA

ABERSON, C.L. *Applied Power Analysis for the Behavioral Science*. 2010 (vid. págs. 36, 45).

ALEXANDER, FE y col. “14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening”. En: *The Lancet* 353.9168 (1999), págs. 1903-1908. DOI: 10.1016/S0140-6736(98)07413-3 (vid. pág. 11).

ALFANO, Robert. “Method and apparatus for detecting cancerous tissue using luminescence excitation spectra”. En: () (vid. pág. 13).

ALJABAR, P., D. RUECKERT y W. R. CRUM. “Automated morphological analysis of magnetic resonance brain imaging using spectral analysis”. En: *NeuroImage* 43.2 (2008), págs. 225-235. DOI: 10.1016/j.neuroimage.2008.07.055 (vid. pág. 13).

AMERICAN CANCER SOCIETY. *Breast Cancer Early Detection and Diagnosis .* https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html. Accessed: 27-07-2019 (vid. pág. 11).

ARAQUE, Oscar y col. “Selecting the mammographic-view for the parenchymal analysis-based breast cancer risk assessment”. En: *2019 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2019 - Proceedings* (2019). DOI: 10.1109/BHI.2019.8834461 (vid. pág. 18).

BALAKRISHNAMA, S. y A. GANAPATHIRAJU. “Linear Discriminant Analysis - a Brief Tutorial”. En: *Compute* October (2015) (vid. pág. 27).

BELHUMEUR, Peter N., Joao P. HESPANHA y David J. KRIEGMAN. “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”. En: 19.7 (1997), pág. 711 (vid. pág. 28).

BRIGHAM, E. Oran. *The fast Fourier transform and its applications*. 1988 (vid. págs. 24, 25).

CABALLERO DÍAZ, Félix Francisco. *Selección de modelos mediante criterios de información en análisis factorial. Aspectos teóricos y computacionales*. 2011 (vid. pág. 47).

CAPLAN, L. S., B. L. WELLS y S. HAYNES. “Breast cancer screening among older racial/ethnic minorities and whites: barriers to early detection.” En: *Journal of gerontology* 47 Spec No (1992), págs. 101-10 (vid. pág. 11).

CARVALHO FERREIRA, Juliana y Cecilia Maria PATINO. “What does the p value really mean?” En: *J Bras Pneumol* 41.5 (2015), págs. 485-485. DOI: 10.1590/S1806-37132015000000215 (vid. pág. 31).

CHEN, Zexun, Bo WANG y Alexander N. GORBAN. “Multivariate Gaussian and Student-t process regression for multi-output prediction”. En: *Neural Computing and Applications* (2019). DOI: 10.1007/s00521-019-04687-8. arXiv: 1703.04455 (vid. pág. 33).

COSTANTINO, Joseph P. y col. “Validation studies for models projecting the risk of invasive and total breast cancer incidence”. En: *Journal of the National Cancer Institute* 91.18 (1999), págs. 1541-1548. DOI: 10.1093/jnci/91.18.1541 (vid. pág. 11).

DELONG, Elizabeth R., David M. DELONG y Daniel L. CLARKE-PEARSON. “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach”. En: *Biometrics* 44.3 (1988), pág. 837. DOI: 10.2307/2531595 (vid. pág. 50).

DUDA, Richard O., Peter E. HART y David G. STORK. *Pattern classification. Second Edition*. 2000 (vid. pág. 29).

ELLIS, Paul. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, 2010, pág. 52 (vid. pág. 36).

EVERITT, Brian S. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 2002, pág. 321 (vid. pág. 36).

FERRI, F J y col. *Comparative Study of Techniques for Large-Scale Feature Selection*. Inf. téc. (vid. pág. 39).

FOWLER, Erin E.E. y col. “Generalized breast density metrics”. En: *Physics in Medicine and Biology* 64.1 (2019). DOI: 10.1088/1361-6560/aaf307 (vid. págs. 14-16, 22, 25-27, 35, 43-45, 52).

GAIL, Mitchell H. y col. “Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer”. En: *Journal of the National Cancer Institute* 91.21 (1999), págs. 1829-1846. DOI: 10.1093/jnci/91.21.1829 (vid. pág. 11).

GARCIA, F. J., M. J. TAYLOR y M. C. KELLEY. “Two-dimensional spectral analysis of mesospheric airglow image data”. En: *Applied Optics* 36.29 (1997), pág. 7374. DOI: 10.1364/ao.36.007374 (vid. pág. 13).

GIGER, Maryellen L., Nico KARSSEMEIJER y Julia A. SCHNABEL. “Breast Image Analysis for Risk Assessment, Detection, Diagnosis, and Treatment of Cancer”. En: *Annual Review of Biomedical Engineering* 15.1 (2013), págs. 327-357. DOI: 10.1146/annurev-bioeng-071812-152416 (vid. pág. 12).

GLOBOCAN. “Fact-sheets Colombia 2018”. En: *International Agency For Research on Cancer* 380.Globocan (2019), págs. 2018-2019 (vid. pág. 10).

GONZALES, Rafael C. y Richard E. WOODS. “Digital Image Processing”. En: *Computer* 7.5 (1974), págs. 17-19. DOI: 10.1109/MC.1974.6323522 (vid. pág. 25).

HEINE, John J. y Robert P. VELTHUIZEN. “Spectral analysis of full field digital mammography data”. En: *Medical Physics* 29.5 (2002), págs. 647-661. DOI: 10.1118/1.1445410 (vid. pág. 14).

HUNG, H. M. y col. “The Behavior of the P-Value When the Alternative Hypothesis is True”. En: *Biometrics* 53.1 (1997), pág. 11. DOI: 10.2307/2533093 (vid. pág. 31).

KHAN ACADEMY. *El uso de una tabla para estimar el valor p del estadístico t*. https://es.khanacademy.org/math/statistics-probability/significance-tests-one-sample/tests-about-population-mean/v/calculating-p-value-from-t-statistic. Accessed: 2020-07-03 (vid. pág. 34).

LI, Hongbin. *Spectral Analysis of Signals [Book Review]*. Vol. 24. 1. 2008, págs. 148-150. DOI: 10.1109/msp.2007.273066 (vid. pág. 13).

LI, Hui y col. “Fractal Analysis of Mammographic Parenchymal Patterns in Breast Cancer Risk Assessment”. En: *Academic Radiology* 14.5 (2007), págs. 513-521. DOI: 10.1016/j.acra.2007.02.003 (vid. págs. 14-16, 19, 29, 41).

LI, Ming y Baozong YUAN. “2D-LDA: A statistical linear discriminant analysis for image matrix”. En: (2004). DOI: 10.1016/j.patrec.2004.09.007 (vid. págs. 27, 28).

MARANA, A.N. y col. “Estimating crowd density with Minkowski fractal dimension”. En: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. IEEE, 1999, 3521-3524 vol.6. DOI: 10.1109/ICASSP.1999.757602 (vid. pág. 19).

MARTÍNEZ TOVAR, José Gerardo. *Distribución “T” de Student*. https://estadisticaen-investigacion.wordpress.com/distribucion-t-de-student/. Accessed: 2020-07-03 (vid. pág. 34).

MEDICAL NEWS TODAY. *What to know about breast cancer.* https://www.medical-newstoday.com/articles/37136.php. Accessed: 27-07-2019 (vid. pág. 10).

MEIKLE, Steven R. y col. “Parametric image reconstruction using spectral analysis of PET projection data”. En: *Physics in Medicine and Biology* 43.3 (1998), págs. 651-666. DOI: 10.1088/0031-9155/43/3/016 (vid. pág. 13).

NATIONAL CANCER INSTITUTE. *Breast Cancer Risk Assessment Tool*. https://bcri-sktool.cancer.gov/. Accessed: 10-10-2019 (vid. pág. 11).

NCSS STATISCAL SOFTWARE. *Stepwise Regression*. https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf. Accessed: 2020-07-03 (vid. pág. 39).

NEYMAN, J. y E. S. PEARSON. “The testing of statistical hypotheses in relation to probabilities a priori”. En: *Mathematical Proceedings of the Cambridge Philosop-*

hical Society 29.4 (1933), págs. 492-510. DOI: 10.1017/S030500410001152X (vid. pág. 36).

PERCIVAL, Don. “Introduction to spectral analysis”. En: *Applied Physics* (2003), págs. 1-8 (vid. pág. 14).

PEREIRA GONZÁLEZ, Augusto. “Selección de características para el reconocimiento de patrones con datos de alta dimensionalidad en fusión nuclear”. En: (2015) (vid. pág. 29).

PERTUZ, Said y col. “Clinical Evaluation of a Fully-automated Parenchymal Analysis Software for Breast Cancer Risk Assessment: a Pilot Study in a Finnish Sample”. En: *European Journal of Radiology* (2019), pág. 108710. DOI: 10.1016/j.ejrad.2019.108710 (vid. págs. 12, 17).

PERTUZ, Said y col. “Open Framework for Mammography-based Breast Cancer Risk Assessment”. En: *IEEE International Conference on Biomedical and Health Informatics* (2019) (vid. págs. 12, 18, 23).

RAYNER, J. N. “Spectral Analysis”. En: *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2001, págs. 14861-14864. DOI: 10.1016/B0-08-043076-7/02514-6 (vid. pág. 14).

RIPAMONTI, Enrico. “The use of p-values in applied research: Interpretation and new trends”. En: *Statistica* 76.4 (2016), págs. 315-325. DOI: 10.6092/issn.1973-2201/6439 (vid. pág. 31).

RIVAS RUIZ, Rodolfo, Marcela PEREZ RODRÍGUEZ y Juan O. TALAVERA. “Clinical research XV. From the clinical judgment to the statistical model. Difference

between means. Student's t test." En: *Revista médica del Instituto Mexicano del Seguro Social* 51.3 (2013), págs. 300-303 (vid. pág. 33).

RODRIGO, Joaquín Amat. "Análisis discriminante lineal (LDA) y Análisis discriminante cuadrático (QDA)". En: (2016) (vid. pág. 28).

ROYALL, Richard. *The Nature of Scientific Evidence*. 2004, págs. 119-152. DOI: 10.7208/chicago/9780226789583.003.0005 (vid. pág. 30).

RÜCKSTIESS, Thomas, Christian OSENDORFER y Patrick VAN DER SMAGT. "Sequential Feature Selection for Classification". En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7106 LNAI. 2011, págs. 132-141. DOI: 10.1007/978-3-642-25832-9_14 (vid. pág. 39).

SALGADO LOAIZA, Victor Hugo. *Cálculo de la Dimensión Fractal mediante Box-Counting* (vid. pág. 20).

SCIENTIFIC EUROPEAN FEDERATION OF OSTEOPATHS. "Prueba "t"de Student". En: (2019) (vid. pág. 32).

SHUTTLEWORTH, Martyn y Lindsay WILSON. *Type I Error and Type II Error .* https://explorable.com/type-i-error. Accessed: 2020-07-03 (vid. pág. 36).

SILVETTI, Andrea y Claudio DELRIEUX. "Análisis Multifractal Aplicado a Imágenes Médicas". En: (2010) (vid. págs. 13, 19).

SPASOVA DIMITROVA, Radostina. "Desarrollo y evaluación de métodos de selección de características para la predicción de eventos adversos en pacientes polimedicados". En: (2017) (vid. págs. 29, 30).

SVOBODA, Tomáš. “Frequency analysis in images 2D Fourier Transform”. En: (2008) (vid. pág. 13).

SÁNCHEZ TURCIOS, Reinaldo Alberto. “t-Student. Usos y abusos”. En: (2015) (vid. pág. 32).

TORRES, German F. y col. “Morphological Area Gradient: System-independent Dense Tissue Segmentation in Mammography Images”. En: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (2019), págs. 4855-4858. DOI: 10.1109/EMBC.2019.8857320 (vid. pág. 23).

TRAVERS, Jason C., Bryan G. COOK y Lysandra COOK. “Null Hypothesis Significance Testing and p Values”. En: *Learning Disabilities Research and Practice* 32.4 (2017), págs. 208-215. DOI: 10.1111/ldrp.12147 (vid. pág. 31).

TSURUOKA, M., R. SHIBASAKI y S. MURAI. “Spectral analysis of standing balance using medical stereo images”. En: *International Conference of the IEEE Engineering in Medicine and Biology Society. (Cat. No.97CH36136)*. Vol. 4. IEEE, págs. 1671-1674. DOI: 10.1109/IEMBS.1997.757041 (vid. pág. 13).

UK TRIAL OF EARLY DETECTION OF BREAST CANCER GROUP. “First results on mortality reduction in the UK trial of early detection of breast cancer”. En: *The Lancet* 332.8608 (1988), págs. 411-416. DOI: 10.1016/S0140-6736(88)90410-2 (vid. pág. 11).

UNIVERSIDAD DE LA REPÚBLICA DE URUGUAY. *Introducción a la Teoría del Procesamiento Digital de Señales de Audio*. https://www.eumus.edu.uy/eme/ensenanza/electivas/dsp/presentaciones/clase06.pdf. Accessed: 2020-07-03 (vid. pág. 25).

URBANOWICZ, Ryan J. y col. *Relief-based feature selection: Introduction and review*. 2018. DOI: 10.1016/j.jbi.2018.07.014. arXiv: 1711.08421 (vid. pág. 30).

VILLEGAS, J. Gallo y J. FARBIARZ. "Análisis espectral de la variabilidad de la frecuencia cardiaca". En: *Iatreia* 12.2 (1999), págs. 61-71 (vid. pág. 13).

VIOLINI, María Lucía. "Selección de Características. Su aplicación a Clasificación de Texturas." En: (2014) (vid. pág. 30).

WIKIPEDIA, la enciclopedia libre. *File: Great Britain Box.svg - Wikimedia Commons*. https://commons.wikimedia.org/wiki/File:Great_Britain_Box.svg. Accessed: 2020-07-03 (vid. pág. 20).

WIKIPEDIA, la enciclopedia libre. *File: Great Britain coverings.svg - Wikimedia Commons*. https://commons.wikimedia.org/wiki/File:Great_Britain_coverings.svg. Accessed: 11-10-2019 (vid. pág. 21).

WIKIPEDIA, la enciclopedia libre. *File:Student densite best.JPG - Wikimedia Commons*. https://commons.wikimedia.org/wiki/File:Student_densite_best.JPG. Accessed: 2020-07-03 (vid. pág. 33).

WIKIPEDIA, the free encyclopedia. *Linear discriminant analysis*. https://en.wikipedia.org/wiki/Linear_discriminant_analysis. Accessed: 2020-07-03 (vid. pág. 27).

WIKIPEDIA, the free encyclopedia. *Neighbourhood components analysis*. https://en.wikipedia.org/wiki/Neighbourhood_components_analysis. Accessed: 2020-07-03 (vid. pág. 36).

WIKIPEDIA, the free encyclopedia. *Power of a test*. https://en.wikipedia.org/wiki/Power_of_a_test. Accessed: 2020-07-03 (vid. pág. 36).

WILKINSON, Leland. “Tests of significance in stepwise regression”. En: *Psychological Bulletin* 86.1 (1979), págs. 168-174. DOI: 10.1037/0033-2909.86.1.168 (vid. pág. 40).

WORLD HEALTH ORGANIZATION. *Cancer. Fact sheets*. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed: 27-07-2019 (vid. pág. 10).

YANG, Wei, Kuanquan WANG y Wangmeng ZUO. “Neighborhood component feature selection for high-dimensional data”. En: *Journal of Computers* 7.1 (2012), págs. 162-168. DOI: 10.4304/jcp.7.1.161-168 (vid. pág. 37).

ZHANG, Cha y Tsuhan CHEN. “Spectral analysis for sampling image-based rendering data”. En: *IEEE Transactions on Circuits and Systems for Video Technology* 13.11 (2003), págs. 1038-1050. DOI: 10.1109/TCSVT.2003.817350 (vid. pág. 13).