UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

DESARROLLO DE UN SISTEMA DE EXTRACCIÓN AVANZADA DE CARACTERÍSTICAS EN IMAGEN HISTOLÓGICA PARA LA IDENTIFICACIÓN AUTOMÁTICA DEL CÁNCER DE PRÓSTATA.

AUTORA: ELENA PAYÁ BOSCH

TUTORA: VALERY NARANJO ORNEDO

COTUTOR: JOSÉ GABRIEL GARCÍA PARDO

Curso Académico: 2017-18

Agradecimientos

“ A mi familia
A mis amigos
A mi tutora y mi cotutor
A mis compañeros
y a toda la gente que me ha acompañado en esta etapa.

Gracias. ”

Resumen

El presente trabajo fin de grado pretende abordar una de las fases intermedias de un proyecto nacional de mayor embergadura cuyo último objetivo es proporcionar una herramienta a los patólogos que actúe a modo de sistema de ayuda para el diagnóstico temprano del cáncer de próstata. En este TFG se lleva a cabo el diseño y desarrollo de un método de extracción avanzada de características de glándulas prostáticas de imágenes histológicas para la detección del cáncer de grado 3.

Para ello se lleva a cabo una exhaustiva revisión del estado del arte y se desarrolla un método que permite realizar una profunda extracción de características basado en descriptores de forma, de textura y de color. Una vez realizado, se selecciona el espacio de características que proporciona el mayor poder discriminatorio. A continuación, se implementa un algoritmo para abordar una clasificación supervisada que distinga con la mayor precisión posible las glándulas sanas de las patológicas. De esta forma, se consigue un modelo de clasificación capaz de discernir entre ambas clases con un 97,4 % de precisión.

Con respecto a la siguiente etapa, se evalúan los resultados obtenidos del mejor clasificador, la validación del mismo y la predicción con nuevas muestras. Con esta información, se pretende observar las ventajas y las limitaciones obtenidas con las técnicas implementadas, a fin de desarrollar un modelo robusto de clasificación para distinguir entre muestras histológicas de próstata benignas y con cáncer de grado 3. En base a las conclusiones, se expone una serie de líneas futuras de investigación para ayudar a la consecución del objetivo que persigue el proyecto general.

Palabras Clave: cáncer de próstata, imágenes histológicas, extracción y selección de características, clasificación supervisada, *machine learning*.

Resum

El present treball fi de grau pretén abordar una de les fases intermèdies d'un projecte nacional de major embergadura l'últim objectiu de la qual és proporcionar una ferramenta als patòlegs que actue a manera de sistema d'ajuda per al diagnòstic enjorn del càncer de pròstata. En este TFG es du a terme el disseny i desenrotllament d'un mètode d'extracció avançada de característiques de glàndules prostàtiques d'imatges histològiques per a la detecció del càncer de grau 3.

Per a això es du a terme una exhaustiva revisió de l'estat de l'art i es desenrotlla un mètode que permet realitzar una profunda extracció de característiques basat en descriptors de forma, de textura i de color. Una vegada realitzat, se selecciona l'espai de característiques que proporciona el major poder discriminatori. A continuació, s'implementa un algoritme per a abordar una classificació supervisada que distingisca amb la més precisió possible les glàndules sanes de les patològiques. D'esta manera, s'aconseguix un model de classificació capaç de discernir entre ambdós classes amb un 97,4 % de precisió.

Respecte a la següent etapa, s'avaluen els resultats obtinguts del millor classificador, la validació del mateix i la predicció amb noves mostres. Amb esta informació, es pretén observar els avantatges i les limitacions obtingudes amb les tècniques implementades, a fi de desenrotllar un model robust de classificació per a distingir entre mostres histològiques de pròstata benignes i amb càncer de grau 3. Basant-se en les conclusions, s'exposa una sèrie de línies futures d'investigació per a ajudar a la consecució de l'objectiu que perseguix el projecte general.

Paraules clau: càncer de pròstata, imatges histològiques, extracció i selecció de característiques, classificació supervisada, *machine learning*.

Abstract

The current TFG aims to approach one of the intermediate phases of a larger national project whose ultimate goal is to provide a tool to pathologists that acts as an aid-system for the early diagnosis of prostate cancer. In this TFG, the design and development of an advanced extraction method of prostate gland characteristics of histological images for the detection of grade 3 cancer is executed.

For this, an exhaustive review of the state of the art is accomplished and a method that allows a deep extraction of characteristics based on descriptors of shape, texture and color is developed. Once done, the space of characteristics that provides the greatest discriminatory power is selected. Next, an algorithm that allows a supervised classification that distinguishes, as accurately as possible, the healthy glands from the pathological glands is implemented. In this way, a classiication model capable of discerning between both classes with a 97.4 % accuracy is achieved.

Regarding the next stage, the obtained results from the best classifier, the validation and the prediction with new samples are evaluated. With this information, it is intended to observe the advantages and limitations obtained with the techniques implemented, in order to develop a robust classification model to separate between benign prostate histological samples and grade 3 cancer. Based on the conclusions, a lines of future research is exposed in order to help the global project to achive its goals.

Keywords: prostate cancer, histological images, feature extraction and selection, supervised classification, machine learning.

Índice general

Parte I

Memoria

Capítulo 1

Introducción

1.1 Motivación y descripción del problema

1.1.1 Cáncer y cáncer de próstata

1.1.1.1 Cáncer

La Organización Mundial de la Salud (OMS) se refiere al término genérico "cáncer" como un amplio grupo de enfermedades que pueden afectar a cualquier parte del organismo. El cáncer se define como una multiplicación rápida de células anormales que se extienden más allá de sus límites habituales y pueden invadir partes adyacentes del cuerpo o propagarse a otros órganos, dando lugar a metástasis [1]. La transformación de células normales a células tumorales es un proceso progresivo que parte de una lesión precancerosa. Las alteraciones son debidas a:

- Factores genéticos
- Agentes externos:
  - Carcinógenos físicos
  - Carcinógenos químicos
  - Carcinógenos biológicos

La Sociedad Española de Oncología Médica (SEOM) presenta en su último informe epidemiológico de la enfermedad en España [2] datos sobrecogedores que indican un incremento del número de casos y estimaciones de un aumento del 70 % en las próximas décadas, alcanzando aproximadamente los 24 millones de casos en el 2035. Las posibilidades de desarrollar un tumor maligno varían en función de la edad y del sexo, aumentando el riesgo en los hombres de mayor edad. Por tanto, el envejecimiento es otro factor influyente, probablemente debido a la pérdida de eficacia en los mecanismos de reparación celular. Teniendo en cuenta que la esperanza de vida es cada vez mayor, la incidencia de la enfermedad también lo es, por lo que se estima que aproximadamente

uno de cada dos hombres y una de cada tres mujeres padecerá cáncer en algún momento de su vida, como se observa en la figura 1.1.

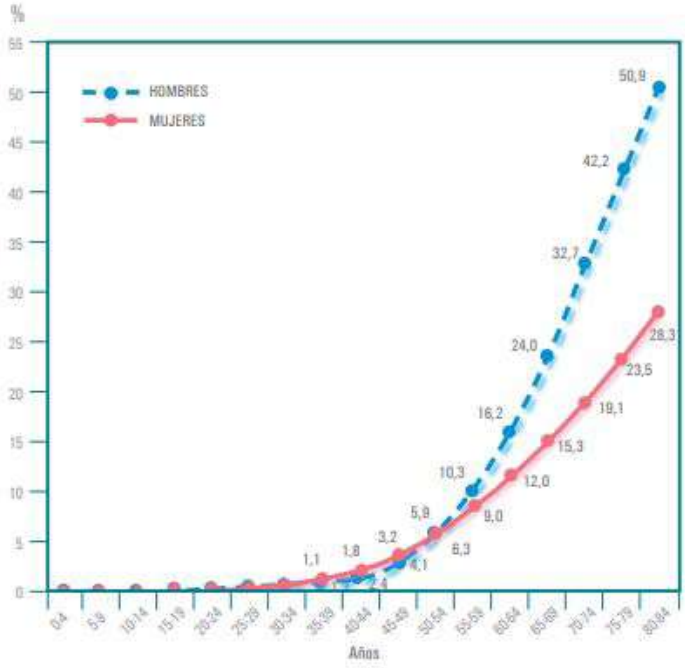


Figura 1.1: Probabilidad (%) de desarrollar un cáncer en España durante el periodo 2003-2007. [2]

Se conoce una gran cantidad de factores de riesgo, pero los principales son el consumo de tabaco y de alcohol, la falta de actividad física y la mala alimentación. Según cifras proporcionadas por GBD 2015 Risk Factors Collaborators, el tabaquismo es el principal factor de riesgo y ocasiona aproximadamente el 22 % de las muertes por cáncer [3]. La OMS propone la reducción de los factores de riesgo junto con la aplicación de estrategias de prevención para reducir entre un 30 % y un 50 % la incidencia del cáncer. No obstante, el hecho de evitar los factores de riesgo no asegura con certeza no padecerlo en un futuro. La prevención del cáncer, en muchas ocasiones pasa por una detección precoz del mismo y el consecuente tratamiento, ya que las posibilidades de una buena recuperación aumentan considerablemente. Existe una relación directa entre un diagnóstico temprano y un tratamiento más eficaz.

Además, cabe destacar que cada tipo de cáncer requiere un protocolo específico; por una parte se tiene que definir el objetivo principal y diseñar el tratamiento considerando las diferentes alternativas y, por otra parte, se debe tener en consideración la calidad de vida del enfermo, ofreciendo cuidados paliativos o apoyo psicosocial si es necesario [1]. Como se ha comentado anteriormente, el tratamiento dependerá del tipo de tumor. La SEOM ofrece estimaciones de los tumores más prevalentes en la población mundial como se puede observar en la Figura 1.2.

Figura 1.2: Estimación de la prevalencia a 5 años de tumores en el mundo para el año 2012. Fuente: GLOBOCAM 2012.[2]

1.1.1.2 Cáncer de próstata

En este trabajo, se estudiará el cáncer de próstata que, como se muestra en la figura 1.2, se encuentra entre los más frecuentes.

La próstata es un órgano glandular del aparato reproductor masculino, con tamaño de nuez, que se encuentra situado en la pelvis, detrás del pubis, delante del recto y por debajo de la vejiga. El tamaño de la próstata varía con la edad, pero se consideran normales valores de 4×3 cm. La función principal de la próstata es producir líquido seminal, el cual protege, mantiene y ayuda a transportar el esperma [4]. El cáncer se origina cuando las células sanas de la próstata cambian, proliferan sin control y forman un tumor. Los síntomas del cáncer aparecen en estadios posteriores e incluyen algunos como [5]:

- Micción frecuente.
- Sangre en la orina o en el líquido seminal.
- Dolor o ardor al orinar.
- Molestias debido al aumento del tamaño de la próstata.

La detección del cáncer se basa en descubrirlo antes de que los síntomas aparezcan. Se usan comúnmente dos pruebas: tacto rectal (DRE)[1] y análisis de sangre del PSA[2]. A pesar de que existen métodos de detección, no es sencillo predecir cuáles crecerán y se diseminarán lentamente y cuáles, rápidamente.

[1] Exploración médica en la que se introduce un dedo a través del esfínter anal para realizar palpación digital

[2] Análisis de sangre que busca la presencia del antígeno prostático específico (PSA). El PSA es una prueba inespecífica en la cual dicho antígeno puede verse elevado por diferentes razones y por ello, sus resultados deben ser interpretados con cautela.

Si los resultados de las pruebas de detección resultan anormales, se realizan más pruebas para confirmar el diagnóstico [5]:

- Análisis del PCA3. En este análisis se busca el gen *PCA3* en la orina, que resulta de gran utilidad en el diagnóstico, ya que este gen se encuentra exclusivamente en las células cancerosas de próstata, es decir, es específico del cáncer de próstata [5].
- Ecografía transrectal (TRUS). Procedimiento en el que se introduce una sonda en el recto que emite ondas sonoras de alta energía produciendo unos ecos cuando rebotan en los tejidos. Estos ecos permiten formar las imágenes computarizadas de la próstata [6]. Tiene varías funciones, pues además de servir para examinar la próstata cuando el paciente presenta un alto nivel de PSA, se utiliza también durante la biopsia para guiar las agujas a una zona determinada.
- Biopsia. Consiste en la extirpación de una pequeña cantidad de tejido para posteriormente examinarla bajo el microscopio y analizar posibles indicios de cáncer.

1.1.2 Biopsia e imagen histológica

1.1.2.1 Biopsia

La biopsia se suele realizar para poder formular un diagnóstico definitivo cuando el tacto rectal o el análisis de PSA proporcionan algún resultado sospechoso [7]. La biopsia conlleva ciertos riesgos como la dificultad para orinar, la hemorragia y la infección, debido a que la prueba se realiza vía transrectal. Se toman varias muestras de tejido de diversas zonas de la próstata para garantizar mayor fiabilidad de los resultados al disponer de una mayor cantidad de muestras. Se suelen extraer entre 12 y 14 piezas de tejido [5]. El médico usa la TRUS para identificar la próstata y los nódulos sospechosos y así, poder insertar la aguja de biopsia correctamente para la obtención de las regiones tisulares de interés.

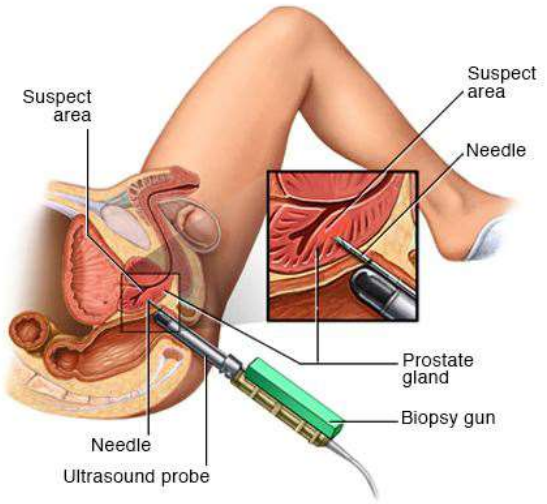


Figura 1.3: Biopsia de próstata. Fuente: MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH.

Existe una variante en la cual la biopsia se realiza con fusión de imágenes por resonancia magnética (MRI)[3]. Un software informático combina las imágenes obtenidas por el examen de MRI y la TRUS dando lugar a una imagen tridimensional que permite determinar con más precisión el área a examinar. Una vez realizada la biopsia, se obtendrán los cortes histológicos de las muestras de la próstata.

1.1.2.2 Imagen histológica

La histología es una rama de la biología que estudia la composición, la estructura y las características de los tejidos orgánicos de los seres vivos. Para poder llevar a cabo el estudio histológico es necesario un procesado del tejido extraído mediante la biopsia. La técnica histológica consiste en un conjunto de operaciones a las que se somete el material biológico para que sea posible su estudio bajo el microscopio. Esto permite a los especialistas en anatomía patológica observar las estructuras no visibles al ojo humano [9].

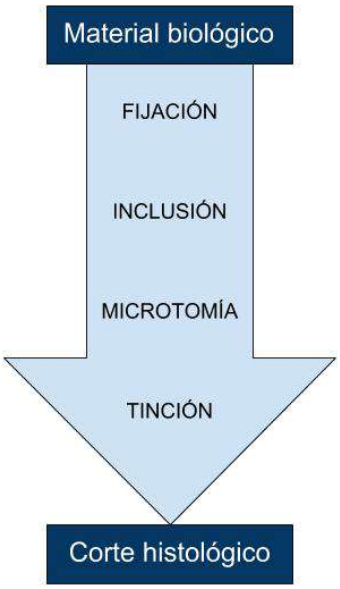


Figura 1.4: Procedimiento general para la preparación del corte histológico.

En el procedimiento general para la preparación del corte histológico, el primer paso es la fijación que consiste en colocar una sustancia fijadora a la muestra, normalmente líquida, para lograr conservar la forma original del tejido. A continuación, se lava el tejido con agua destilada para quitar el exceso de fijador. Luego se realiza la deshidratación, el aclaramiento y la infiltración: procedimientos que pueden hacerse manual o automáticamente. El siguiente paso es la inclusión, cuyo objetivo es proporcionar al tejido un soporte sólido para que sea posible realizar cortes muy finos. Hay que asegurarse de que la muestra se embebe bien en la parafina. Una vez realizada la inclusión, se procede a la microtomía: proceso por el cual se realizan cortes histológicos muy delgados. El tamaño de los cortes varía dependiendo del laboratorio o de las necesidades. Final-

[3]Método de diagnóstico por imagen que utiliza un imán de gran potencia y ondas de radiofrecuencia para producir imágenes de gran detalle de cualquier parte del cuerpo. [8]

mente, se debe aplicar una tinción al corte para poder observar la morfología tisular. El tipo de tinción se elige en función de las estructuras a diferenciar. La tinción más común utilizada para el estudio del cáncer de próstata es la hematoxilina y eosina (H&E). En este proyecto, se ha trabajado con imágenes teñidas con H&E que posteriormente fueron escaneadas para su estudio computacional, como se puede observar en la figura 1.5.

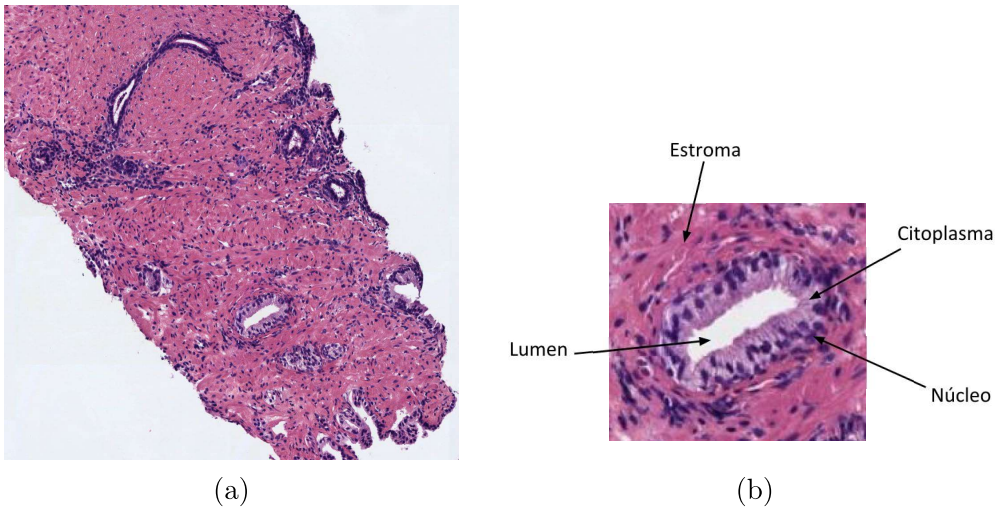


Figura 1.5: (a) Imagen histológica de una muestra de tejido prostático teñida con H&E. (b) Ejemplo de estructura donde se aprecian los diferentes componentes de una glándula histológica de la próstata.

Posteriormente, un patólogo[4] se encarga de analizar los cortes histológicos. En los análisis, el patólogo se encarga de buscar anomalías en el tejido mediante el uso de un microscopio. Se trata de un proceso manual muy laborioso que conlleva un tiempo importante y un alto nivel de discordancia entre diferentes patólogos. Uno de los objetivos de este proyecto es, además de mejorar en términos de coste-efectividad, disminuir el nivel de subjetividad a la hora de analizar las muestras.

Actualmente, la anatomía patológica se enfrenta a una creciente demanda de volumen de trabajo y a mayores cantidades de información, así como una búsqueda de mayor calidad y exactitud diagnóstica para proporcionar seguridad a los pacientes. Con el avance de la tecnología, las posibilidades en la mejora del diagnóstico aumentan considerablemente y los problemas actuales pueden ser solucionados por medio de la patología digital.

La patología digital incluye todos los aspectos relacionados con el uso de las TIC[5] en anatomía patológica. De esta forma, es posible adquirir las imágenes a partir de la preparación convencional y obtener un fichero de imágenes digitales que se puede visualizar en un monitor y analizar mediante herramientas software. Dichas herramientas permiten automatizar y objetivar el proceso para acelerar el diagnóstico, así como aumentar la precisión, la sensibilidad y la especificidad del mismo [10].

En el presente TFG se ha tratado de desarrollar un algoritmo capaz de ayudar a los patólogos en la exploración microscópica del tejido prostático a la hora de diferenciar entre muestras histológicas sanas y patológicas de grado 3.

[4]El patólogo es un médico especializado en el estudio de las alteraciones estructurales, bioquímicas y funcionales en células, tejidos y órganos.

[5]Tecnologías de la información y la comunicación.

1.1.3 Estado del arte

En la actualidad, los expertos basan sus análisis en el sistema de puntuación conocido como *Gleason Score*, el cual se utiliza para determinar la agresividad de un cáncer prostático. Este sistema puede ser empleado para elegir el tratamiento apropiado, a pesar de que el análisis histológico solo se realiza de algunas muestras pequeñas de la próstata.

En el presente TFG, se hace uso del sistema de Gleason mencionado anteriormente, que es una adaptación del sistema tradional. En él, se asignan dos grados a cada paciente, de manera que se utiliza una calificación primaria para describir el tipo de células predominante en la muestra y una calificación secundaria para el segundo tipo de células que presenta en mayor abundancia la región tisular. De esta forma, cuanto más alto sea el grado, más severo es y más probable será que crezca y se propague rápidamente. El sistema de clasificación Gleason va de 1 a 5 e indica la severidad del posible cáncer de próstata [11]. A continuación, se muestra una tabla en la que se describen con más detalle las características histológicas que definen cada grado y las diferencias entre ambos.

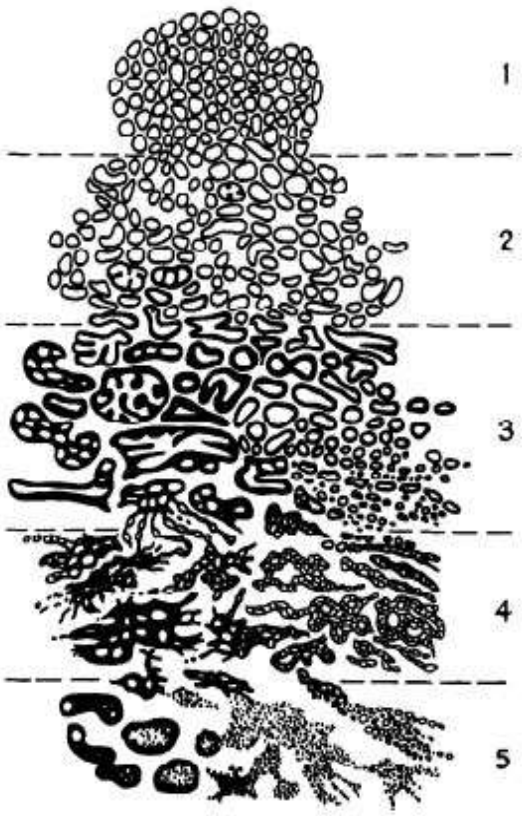


Figura 1.6: Escala de patrones de Gleason.

1	Glándulas pequeñas y uniformes. Sólo glándulas discretas bien formadas.
2	Más espacio (estroma) entre glándulas. Predominantemente glándulas bien formadas pero se empiezan a apreciar glándulas pobremente formadas, fusionadas o cribiformes.
3	Infiltración de células fuera de los márgenes que empiezan a invadir tejido circundante. Predominantemente glándulas pobremente formadas, fusionadas o cribiformes con un pequeño componente de glándulas bien formadas.
4	Masas irregulares de células neoplásicas con pocas glándulas. Solo glándulas pobremente formadas, fusionadas o cribiformes.
5	Falta de glándulas ocasionales y láminas de células. Carece de glándulas formadas. No hay estructura propia del tejido.

Tabla 1.1: Principales características histológicas de cada grado de la escala de patrones Gleason [11].

Una vez conocidos los diferentes grados, se puede observar que los grados 1 y 2 son tejidos con patrones similares al tejido sano cuya repercusión clínica es mínima. Por tanto, en la práctica el estudio del cáncer comienza en el grado 3. Por otra parte, el grado 5 es el más avanzado, pero su

distinción es clara y el diagnóstico es más sencillo. En cuanto a los grados 3 y 4, su diagnóstico es el más complicado y en la literatura encontramos una gran cantidad de estudios referidos a estos, como los propuestos en [12] y [13]. El grado 3 sigue un patrón que se asimila al tejido sano pero tiene indicios de rotura de glándulas e infiltración celular, mientras que el grado 4 ya ha perdido prácticamente la estructura del tejido sano. En base a la literatura, se ha obtenido una serie de características que definen a grandes rasgos tanto las glándulas en su conjunto como los lúmenes para estos grados más conflictivos. Dichas características se han tenido en cuenta a la hora de desarrollar el algoritmo de este proyecto.

	Benigno	**Grado 3**	**Grado 4**
Glándulas	Grandes, densas y separadas	Más pequeñas y más circulares	Fusión de glándulas mal definidas
Lúmenes	Grandes con formas variables (circulares, ovales, ramificados) y los límites anchos (muchas capas de células alrededor)	Pequeños con formas circulares y los límites estrechos (una sola capa de células alrededor)	Sin capas ordenadas de células alrededor

Tabla 1.2: Características histológicas del tejido prostático.

En base a lo expuesto anteriormente, el presente trabajo consiste principalmente en el desarrollo de un algoritmo cuyo objetivo es distinguir con la mayor precisión posible entre glándulas sanas y patológicas de grado 3. Para ello, se aborda el análisis hacia la glándula y sus partes. Como se puede ver en la figura 1.5, hay cuatro estructuras diferentes en la glándula; a partir de las cuales se obtienen diferentes características para definir su forma, su textura y su color. El estroma, es el área más rosada, correspondiente al tejido conjuntivo extracelular que sirve de sostén. El lúmen o luz es la región más clara, de color blanca y en ocasiones vemos mucina[6] en su interior. El citoplasma es la zona de tono púrpura claro; y finalmente, los núcleos son la zona de color más oscuro.

Una vez se han analizado los componentes de las glándulas y las diferencias que presentan, según el grado del sistema de clasificación de Gleason, es posible llevar a cabo una extracción de características que permita la distinción automática de dichos grados. A continuación, se presenta la tabla 1.3 a modo de resumen, en la que se describen algunos ejemplos de las características que se explican en la literatura médica para distinguir los diferentes tipos de tejidos.

[6]Familia de proteínas de alto peso molecular y altamente glicosiladas que son producidas por células de tejidos epiteliales [14]. La mucina posee un color azul tras realizar la tinción y se puede encontrar en algunos lúmenes de glandulas cancerosas.

Autores	**Tipos de tejido**	**Características**
Nguyen, Sabata y Jain [13]	Benigno, grado 3 y grado 4	▪ Del lumen: • Estadísticos de área, perímetro y circularidad • Porcentaje de lumen en la glándula • Número de lúmenes en la glándula ▪ De los núcleos: • Densidad de núcleos • Porcentaje de núcleos en la glándula ▪ Media y varianza de la distancia entre el centroide del lumen y el borde
Nguyen, Sarkar y Jain [15]	Artefactos, benigno y cancerígeno	▪ Del lumen: • Área • Solidez • Circularidad ▪ De los núcleos: • Media y desviación típica del vecindario • Media y desviación típica del vecindario que contiene núcleos ▪ Del citoplasma: • Media y desviación típica del vecindario que contiene núcleos ▪ Media y desviación típica de la distancia entre los píxeles del lumen y los núcleo de alrededor
Jian Ren [12]	Grado 3 y grado 4	▪ Bag-of-Words en características de SIFT + k-medias ▪ Histogram Of Gradients (HOG) ▪ Media y desviación estándar de los histogramas de intensidades para cada canal (RGB)
Kwak y Hewitt [16]	Benigno y cancerígeno	▪ Del lumen: • Área, extensión, compacidad y redondez • Radio del perímetro epitelial • Distancia al lumen más cercano • Número de lúmenes ▪ Distancia del lumen al núcleo más cercano ▪ Distancia del núcleo al lumen más cercano

Tabla 1.3: Algunos ejemplos de tipos de características obtenidas para la clasificación de cáncer de próstata a partir de imagen histológica.

1.2 Objetivos

El fin último del presente trabajo es la identificación automática de cáncer de próstata de grado 3 a partir de glándulas segmentadas de imágenes histológicas teñidas con hematoxilina y eosina, con la finalidad de ayudar a los profesionales en anatomía patológica a formular un diagnóstico preciso y en el menor tiempo posible. Dicho proceso se traduce en un problema de clasificación de dos clases mediante la extracción avanzada de características.

En este proyecto se plantean varios objetivos secundarios que se detallan brevemente a continuación:

1. Realizar una revisión de la literatura científica para obtener información sobre el estado del arte, a fin de conocer los procedimientos, técnicas de extracción de características y métodos de clasificación más empleados.
2. Crear una base de datos de glándulas de imágenes histológicas etiquetadas según sean sanas o patológicas de grado 3.
3. Preparación de la base de datos mediante un procesado de las imágenes y la obtención de diferentes espacios de color.
4. Dividir las muestras en dos grupos. Por una parte, el grupo de *train* con el 80 % de las glándulas que será empleado en la selección de características y en el entrenamiento del clasificador. Y, por otra parte, el grupo de *test* con el 20 % restante que será usado en la etapa de predicción.
5. Utilizar diferentes tipos de descriptores y estudiar las mejores características de cada uno de ellos. Extraer parámetros de descriptores de forma, de textura y de color para diferenciar entre glándulas sanas y patológicas de grado 3.
6. Analizar estadísticamente las características anteriormente extraídas para seleccionar aquellas que ofrezcan una información más relevante en términos de independencia con la clase y entre pares de variables.
7. Analizar el rendimiento de diferentes clasificadores para escoger el que mejor resultados proporcione mediante un proceso de validación.
8. Predecir el comportamiento que tendrá el clasificador ante muestras nuevas mediante el conjunto de datos de *test* y analizar los resultados de la clasificación.
9. Extraer conclusiones mediante el análisis de los resultados obtenidos, comprobar si se ha cumplido el objetivo principal e identificar los problemas surgidos a lo largo del proyecto. Y en base a esto, proponer posibles mejoras y cambios a realizar para investigaciones futuras.

1.3 Guía de la memoria

El capítulo 2 consta de dos apartados bien diferenciados. En la primera sección, se exponen los materiales que han sido necesarios para lograr el cumplimiento del objetivo final del trabajo. Y, en la siguiente sección, se explica detalladamente la metodología implementada para el desarrollo de los algoritmos relacionados con los procesos de extracción y selección de características, así como de clasificación supervisada. Por último, se presentan ejemplos con imágenes glandulares y se listan las funciones y comandos utilizados.

En el capítulo 3, se describen y se discuten los resultados que se han obtenido, tanto los de validación del mejor clasificador como los de la predicción con nuevas imágenes.

Finalmente, en el capítulo 4, se recopilan las conclusiones más importantes extraídas del estudio realizado y se detallan las posibles líneas futuras de investigación.

Capítulo 2

Material y método

2.1 Material

2.1.1 Base de datos

La base de datos usada para el desarrollo de este proyecto es una base de datos privada que se ha construido a partir de imágenes histológicas de biopsias de próstata procedentes del departamento de anatomía patológica del Hospital Clínico Universitario de Valencia.

En este trabajo, se utilizan imágenes de 25 pacientes diferentes, de los cuales 16 aportan imágenes benignas y 9, patológicas. Además, de cada uno de los pacientes se obtuvieron varias imágenes, dando un total de 251 imágenes de tejido sano y 71 imágenes de tejido patológico, todas ellas del mismo tamaño, 1024×1024.

Las muestras biopsiadas se escanearon para su análisis digital mediante un miscroscopio disponible en el Hospital Clínico de Valencia que permite capturar las imágenes con unas dimensiones de $[50000, 90000] \times [90000, 190000]$ píxeles y un tamaño desde 300 hasta 1.500MB (1,5GB). Debido a sus enormes dimensiones, las imágenes fueron remuestreadas de 40x a 10x y fueron divididas en diferentes *patches*[1] más pequeños a fin de poder visualizarlas en el software de MATLAB®. De esta forma, finalmente se trabaja con imágenes de $1024 \times 1024 \times 3$ píxeles.

Por una parte, para la preparación de las imágenes con cáncer de grado 3 se requiere la participación de médicos especializados en anatomía patológica que anoten las áreas relevantes que poseen información sobre el cáncer de grado 3, es decir, las zonas donde se encuentra la lesión. Para ello, se le facilitó la aplicación *MicroDraw* en la que se podían marcar zonas de interés en la muestra, esta aplicación es una plataforma especializada en la visualización y etiquetado de imágenes [17]. Dichas anotaciones, se convirtieron en máscaras en las cuales las zonas patológicas están a una intensidad de 1 mientras que el resto a 0. Estas máscaras son de gran utilidad para eliminar las zonas que no aportan información y son innecesarias para el presente trabajo.

[1] Fragmentos de una imagen que se obtienen al descomponerla en varias sub-imágenes con el objetivo de estudiar características de interés en regiones ampliadas.

Una vez los patólogos llevan a cabo las anotaciones pertinentes, se genera una base de datos privada donde se almacenan las imágenes etiquetadas según sean sanas o enfermas. Partiendo de esta, se hace uso de la librería *OpenSlide* que proporciona una interfaz sencilla que permite leer las imágenes *Whole Slide*. Estas se caracterizan por poseer alta resolución y ser considerablemente grandes, lo cual conlleva un problema a la hora de cargarlas con librerías estándares.

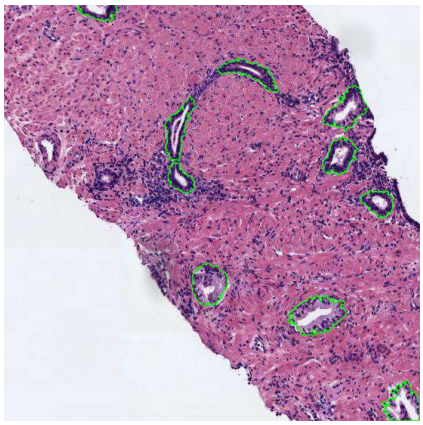


Figura 2.1: Imagen histológica de 1024x1024 píxeles donde se pueden apreciar las glándulas segmentadas.

Por otra parte, tanto para las imágenes benignas como para las patológicas, se realiza una segmentación manual en la que se marcan todas las glándulas de cada imagen mediante líneas de color verde, como se observa en la Figura 2.1.El objetivo en este punto es trabajar con las regiones de interés, previamente detectadas, para extraer características únicamente de aquellas zonas capaces de aportar información relevante al estudio.

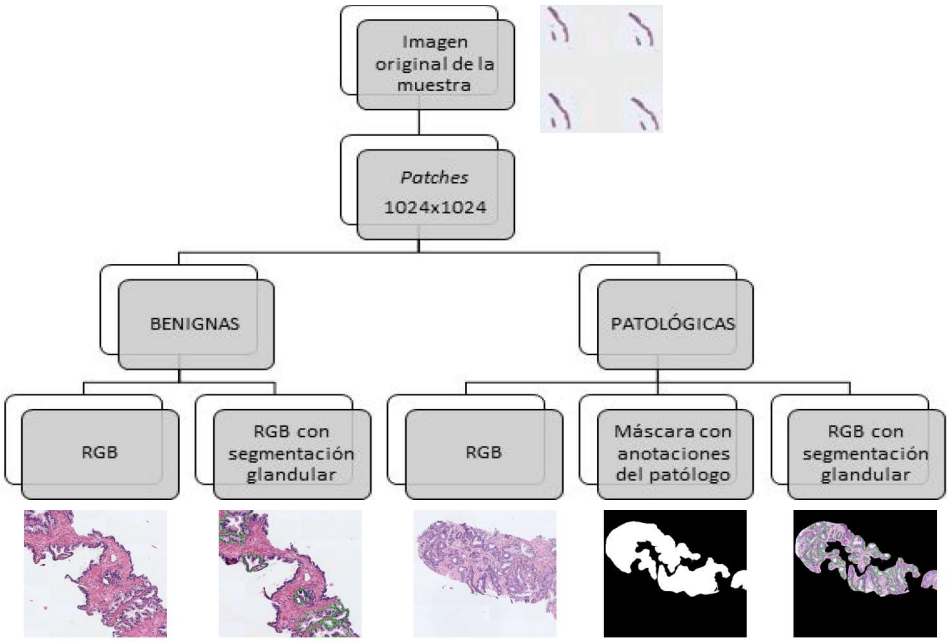


Figura 2.2: Esquema del proceso hasta la obtención de la base de datos del estudio.

Partiendo de las imágenes originales, se ha agrupado en 2 y 3 carpetas la información correspondiente a las muestras sanas y patológicas, respectivamente, tal y como se muestra en el esquema

de la Figura 2.2. A partir de estas, se realizará el preproceso explicado en la sección 2.2.1 y se obtendrán 913 glándulas benignas y 1026 patológicas que constituirán la base del resto del estudio.

2.1.2 Software y hardware

2.1.2.1 Software

Para la implementación de los distintos algoritmos empleados en este TFG y para el cálculo de todos los resultados se ha utilizado el programa MATLAB® v.R2016b, de The MathWorks, Inc. (Natick, Massachusetts, Estados Unidos). Este software combina un entorno de escritorio perfeccionado para el análisis iterativo y los procesos de diseño con un lenguaje de programación que expresa las matemáticas de matrices y vectores directamente. MATLAB® resulta de gran utilidad para numerosas aplicaciones como analítica de datos, comunicaciones inalámbricas, aprendizaje profundo, procesamiento de señales, visión artificial, finanzas, robótica y sistemas de control [18].

En este TFG se utilizará una de las apps que incluye la plataforma de MATLAB® llamada *Classification Learner*. Esta se caracteriza principalmente por su capacidad para entrenar modelos para clasificación supervisada de datos.

2.1.2.2 Hardware

Es necesario conocer las características del hardware en el que se ha llevado a cabo el proyecto, ya que esto repercutirá tanto en el funcionamiento como en la velocidad y rendimiento que se podrá alcanzar al ejecutar los distintos algoritmos.

Este proyecto ha sido desarrollado y ejecutado en el ordenador HP Pavilion TS 14 Notebook PC, cuyas especificaciones son las siguientes:

Sistema operativo:	Windows 10 Home 64 bits
Microprocesador:	Intel(R) Core(TM) i5-4200U CPU
Velocidad:	1.60GHz
Memoria del sistema:	8GB
- Ranura de memoria 1:	4GB Hynix 1600MHz
- Ranura de memoria 2:	4GB Hynix 1600MHz
Dispositivo gráfico 1:	Radeon (TM) HD 8670M
Dispositivo gráfico 2:	Intel(R) HD Graphics Family
Memoria de gráficos:	1.792 MB

Tabla 2.1: Especificaciones del hardware empleado.

2.2 Método

En el presente TFG, se realizará la extracción de características de unidades de glándula correspondientes a las muestras histológicas sanas y patológicas, bajo la hipótesis que sostiene que el estudio de patrones característicos en dichas regiones de interés permite detectar el caracter benigno o maligno del tejido prostático del paciente.

Para ello, en primer lugar se realizará una preparación de la base de datos de la que se ha hablado anteriormente para tener cada glándula de las imágenes anteriores encuadrada en una imagen individual. Después, se extraerán características de forma, de textura y de color. Para cada uno de estos tipos de características se realizará un preprocesado de la base de datos para la correcta preparación de las imágenes con las que posteriormente se va a trabajar.

Una vez extraídas todas las variables de todas las glándulas de cada imagen, se realizará una partición aleatoria para dividir las muestras en dos conjuntos de datos: uno de entrenamiento y otro de test. De esta forma, se garantiza la robustez del modelo al no testear con las mismas imágenes con las que se habrá entrenado.

Posteriormente, se llevará a cabo un proceso de selección de características mediante el conjunto de entrenamiento con la finalidad de que únicamente las variables que son de interés sean usadas como entrada del clasificador. La clasificación será supervisada, por tanto, la finalidad de este paso es que el clasificador "aprenda" qué patrones van asociados a qué clases.

A continuación, se puede observar un diagrama de flujo en el que se muestran encuadradas las funciones desarrolladas y redondeadas las principales tablas de datos.

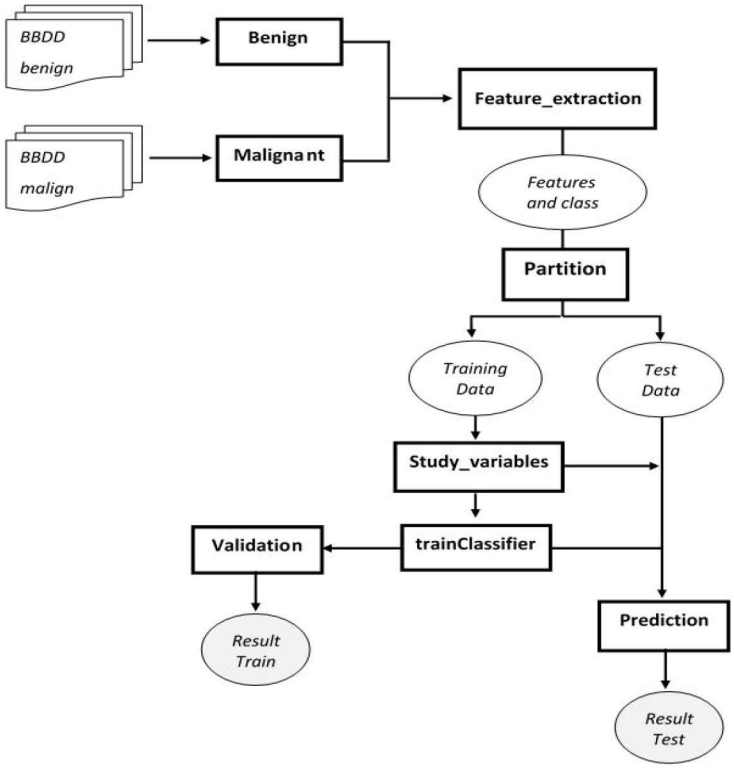


Figura 2.3: Diagrama resumen del método empleado.

Lista de funciones para el método y los resultados

Seguidamente, se listan todas las funciones que se han programado a lo largo del presente TFG con sus respectivos *inputs* y *outputs* y que han sido implementadas en el método propuesto que se expone en el diagrama de la Figura 2.3. También, se exponen las funciones desarrolladas para obtener los resultados tanto de la validación como de la predicción.

A lo largo de esta memoria se irán explicando, cuando sea necesario, las diferentes variables de entrada y de salida de cada función con la finalidad de dar a conocer por qué se necesitan, qué hacen y qué se consigue con ellas.

1. $[FEAT\_SELECTION\_m] = \textbf{benign}(carpetaRGB, carpetaPatologa, carpetaSegGlands)$
2. $[FEAT\_SELECTION\_b] = \textbf{Malignant}(carpetaRGB, carpetaSegGlands)$
3. $[caract] = \textbf{Feature\_Extraction}(Gland, GlandBW, GlandsinMask, prop, bb)$
4. $[GlandCitoplasma, GlandLumen, GlandNucleo, GlandEstroma] = \textbf{clustering}(Gland)$
5. $[HGlandsinMask, EGlandsinMask,] = \textbf{colour\_deconvolution}(GlandsinMask,' H\&E')$
6. $[FEAT\_SELECTION] = \textbf{feat}(FEAT\_SELECTION\_b, FEAT\_SELECTION\_m)$
7. $[training\_data, test\_data] = \textbf{Partition}(FEAT\_SELECTION)$
8. $[training\_data\_selected, names\_selected, mu, sigma] =$ $\textbf{Study\_variables}(training\_data, Allnames)$
9. $[trainedClassifier, validationAccuracy, validationPredictions, validationScores,$ $compactSVM] = \textbf{trainClassifierSVMQ}(training\_data)$
10. $[test\_data\_selected] = \textbf{Select\_features\_test}(test\_data, Allnames, names\_selected,$ $mu, sigma)$
11. $[Result\_train] = \textbf{Validation}(training\_data\_selected)$
12. $[Result\_test] = \textbf{Prediction}(test\_data\_selected)$

2.2.1 Preparación de la base de datos de imágenes

Como ya se ha mencionado anteriormente, para llevar a cabo el objetivo final, se parte de diferentes tipos de imágenes que se agrupan de la siguiente forma:

- Benignas:
  - Imágenes histológicas RGB (del inglés *Red*, *Green*, *Blue* - rojo, verde, azul).
  - Imágenes histológicas segmentadas mediante una línea de color verde que rodea las glándulas.
- Malignas:
  - Imágenes histológicas RGB.
  - Imágenes histológicas segmentadas mediante una línea de color verde que rodea las glándulas.
  - Máscaras en blanco y negro donde el patólogo ha señalado las zonas malignas.

El objetivo será obtener por una parte, las glándulas benignas y por otra, las patológicas. Para ello, se empezará enmascarando las imágenes RGB malignas a fin de obtener solo las zonas que el patólogo ha anotado, ya que es en dichas zonas donde es posible encontrar características asociadas al cáncer de grado 3. Se tendrá que multiplicar ambas imágenes y repetir el proceso para cada canal de color de la imagen RGB. Finalmente, se obtendrá la última imagen de la Figura 2.4.

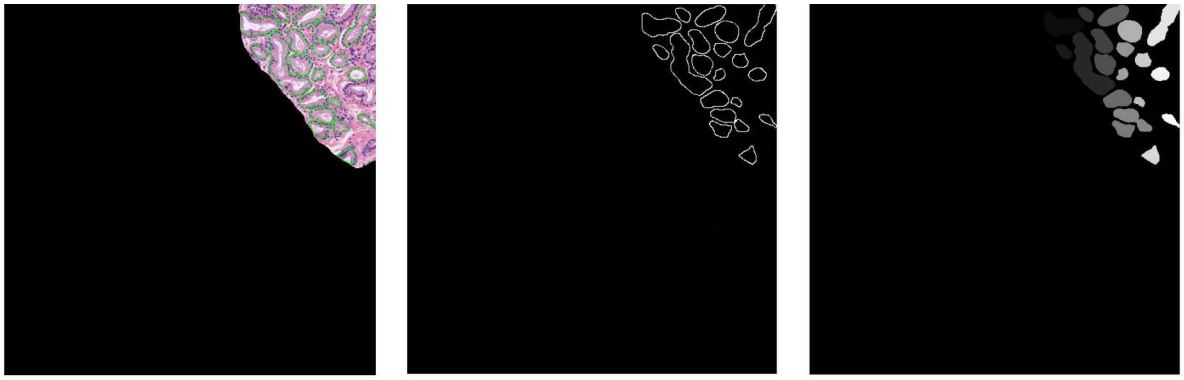
Figura 2.4: Enmascaramiento de la imagen RGB con las anotaciones del patólogo.

De esta forma, se tendrán las imágenes RGB benignas y las RGB malignas, así como sus respectivas imágenes con las glándulas señaladas mediante una línea verde. Con estos dos grupos de imágenes, se pretende obtener finalmente las imágenes individuales correspondientes a:

- Glándula
- Glándula enmascarada
- Máscara glándula

Atendiendo a la segmentación manual (Figura 2.5.a), es posible obtener una máscara binaria relativa a las líneas de segmentación destacadas en color verde (Figura 2.5.b). Para mejorar la

obtención de las máscaras, se aplicará un umbral mediante el método de Otsu[2]. A continuación, se rellenarán los huecos y se obtendrán las máscaras, en ellas se etiqueta cada glándula para poder ir obteniéndolas una a una como se muestra en la Figura 2.5.c.

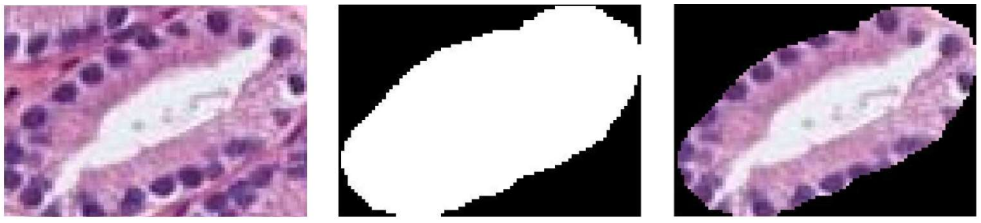


(a) (b) (c)

Figura 2.5: (a) Imagen original de la segmentación glandular manual, (b) bordes de las glándulas segmentadas manualmente y (c) máscara de etiquetas.

El siguiente paso será seleccionar cada glándula dentro de la imagen y obtener la *bounding box*[3] de cada una de ellas.

Una vez adquiridas las diferentes coordenadas de la *bounding box*, se podrá obtener de manera individual cada glándula de la imagen deseada. Las imágenes que se emplearán son la RGB (Figura 2.6.a) y la imagen de etiquetas binarizada (Figura 2.6.b), que permiten obtener la glándula de forma aislada (Figura 2.6.c). De esta manera, se lograrán las tres imágenes de la glándula que son necesarias para posteriormente extraer las distintas características de forma, textura y color.



(a) (b) (c)

Figura 2.6: (a) Glándula sin enmascarar, (b) máscara de la glándula y (c) glándula enmascarada.

[2]Se trata de un método no paramétrico que selecciona el umbral óptimo maximizando la varianza entre clases mediante una búsqueda exhaustiva.

[3]Es el rectángulo más pequeño que contiene la región seleccionada.

2.2.2 Preprocesado de las imágenes

En este apartado, se prepararán las imágenes obtenidas anteriormente para la extracción de parámetros. Se utilizarán unas imágenes u otras en función del tipo de características que se vayan a extraer.

2.2.2.1 Clustering

Para extraer las características relacionadas con los descriptores de forma, se necesitarán tener máscaras de cada una de las partes de la glándula: lumen, citoplasma, núcleo y estroma (véase la Figura 1.5). Para conseguir separar las diferentes regiones se ha desarrollado la siguiente función:

$$[\text{GlandCitoplasma, GlandLumen, GlandNucleo, GlandEstroma}] = \textbf{clustering}(\text{Gland})$$

El *input* de la función será una determinada glándula enmascarada y los *outputs*, las imágenes correspondientes a las diferentes partes de la imagen de la glándula.

La función hace uso de una técnica conocida como *k-means clustering* que consiste en un método de agrupamiento destinado a clasificar cada píxel de la imagen en uno de los k grupos definidos inicialmente, en función del nivel de intensidad de los píxeles.

Dicho método es un algoritmo iterativo de partición de datos que asigna n observaciones a uno de los k grupos definidos por los centroides[4], donde k es un valor que se elige antes del comienzo del algoritmo. El funcionamiento interno de dicho algoritmo procede de la siguiente manera [19]:

1. Elige k centros (o centroides) de *cluster* iniciales.
2. Calcula las distancias de todas las observaciones a cada centroide.
3. Asigna cada obsevación al *cluster* cuyo centroide es el más cercano en términos de distancia euclídea.
4. Calcula el promedio de las observaciones en cada grupo con la finalidad de obtener k nuevas ubicaciones de centroides.
5. Repite desde el paso 2 hasta que las reasignaciones no cambien o hasta que se alcance un número máximo de iteraciones.

El *k-means clustering* pondrá al mismo nivel de intensidad los píxeles que presentaban una intensidad parecida en la imagen original y devolverá una imagen de etiquetas caracterizada por presentar k niveles diferentes de grises, correspondientes a los k grupos definidos (véase la Figura 2.7). En el presente TFG, se distinguirán 4 clases y se establecerá el número de iteraciones en 5. De esta forma, es posible obtener las máscaras asociadas a cada uno de los cuatro elementos de interés presentes en las regiones glandulares de las imágenes histológicas de la próstata, tal como se muestra en la Figura 2.7.

[4] Aquí, hace referencia al centro de un *cluster* (o grupo).

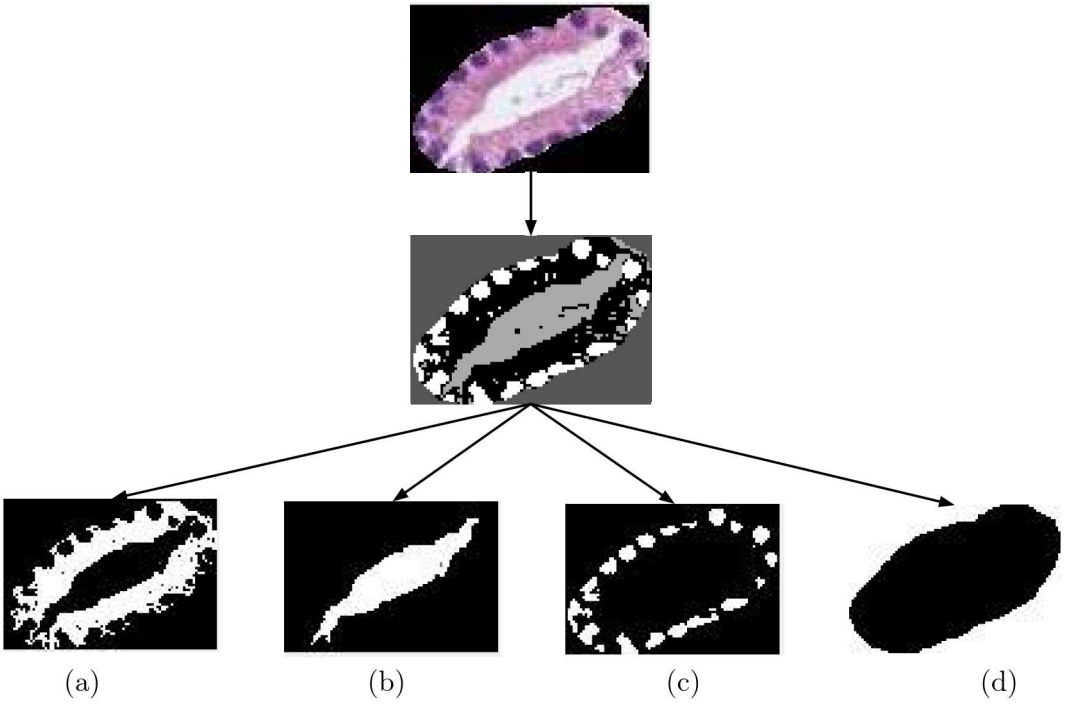


Figura 2.7: *Clustering* de una glándula maligna de próstata donde la primera imagen es la glándula original, la segunda es la imagen de etiquetas y las de abajo son: (a) citoplasma, (b) lumen, (c) núcleos y (d) estroma.

2.2.2.2 Deconvolución de color

La tinción H&E mencionada en la sección 1.1.2, es la usada comúnmente para medicina diagnóstica y se usa en la técnica histológica de muestras de cáncer de próstata. La hematoxilina, que por ser básica, tiñe en tonos azul y púrpura las estructuras ácidas (basófilas), como por ejemplo, los núcleos. Por el contrario, la eosina, debido a su naturaleza ácida, tiñe componentes básicos (acidófilos) en tonos rosa, como por ejemplo, el estroma [20].

Gracias a estas propiedades de la tinción, mediante la deconvolución de color se podrán separar las contribuciones de cada tinción en la imagen final. Por este motivo, se usará en las imágenes que se empleen en la extracción de características de textura.

De la misma manera que en el estudio llevado cabo en [21], el parámetro que se usará para separar las contribuciones de cada tinción en la imagen es la Densidad Óptica (OD) para cada canal; parámetro que se expresa mediante la siguiente ecuación:

$$OD_C = -\log\left(\frac{I_C}{I_{0,C}}\right) = A * c_C \tag{2.1}$$

Como se puede apreciar, la OD es lineal con respecto a la concentración de tinción absorbida por el material. Cada tinción pura tiene una OD asociada a cada canal RGB. Como existe una mezcla de dos tinciones, se puede caracterizar la OD mediante la siguiente matriz:

$$\begin{array}{c} \\ Hematoxilina \\ Eosina \end{array} \begin{array}{c} \begin{array}{ccc} R & G & B \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \end{pmatrix} \end{array}$$

Para obtener los diferentes valores de las contribuciones individuales de cada tinción, se debe realizar una transformación ortonormal previa de la información RGB. Además, se ha de normalizar para conseguir el peso correcto del factor de absorción para cada tinción. Por consiguiente, se divide cada vector de la OD por la longitud total y se obtiene la matriz normalizada (M).

Si C es el vector 2x1 para las cantidades de las dos tinciones en un píxel en particular, el vector de niveles OD detectado en ese píxel es:

$$y = CM \tag{2.2}$$

A partir de dicha ecuación se deduce:

$$C = M^{-1}[y] \tag{2.3}$$

En base a lo expuesto, se revela que la multiplicación de la imagen OD con la inversa de la matriz normalizada, conocida como la matriz de deconvolución de color (D), constituirá una representación ortogonal de las tinciones que forman la imagen.

$$C = D[y] \tag{2.4}$$

La diagonal de D contiene valores superiores a 1 y el resto de la matriz son valores negativos. Si las tinciones fueran puras, la matriz D sería una matriz unitaria. Los valores de los niveles de OD correctos para cada tinción se forman de la siguiente manera:

- Se obtiene la contribución de la hematoxilina restando una porción del canal verde y del azul al canal rojo.
- Se obtiene la contribución de la eosina restando una porción del canal rojo y del azul al canal verde.

La deconvolución de color pretende resaltar las diferencias existentes entre los diferentes elementos de interés de las regiones glandulares, con la finalidad de obtener características relevantes al aplicar los descriptores de textura. Véase ejemplo en la Figura 2.8.

(a) (b) (c)

Figura 2.8: Deconvolución de una glándula prostática maligna. (a) imagen original (b) hematoxilina y (c) eosina.

2.2.3 Extracción de características

La extracción de características consiste en la cuantificación de la información contenida en las imágenes y es uno de los aspectos cruciales de este campo, ya que permitirá diferenciar estadios benignos y patológicos [22]. Por ello, este paso es primordial para la clasificación, ya que dicha información permitirá al clasificador "aprender" analizando qué valores de las características se asocian con una clase u otra.

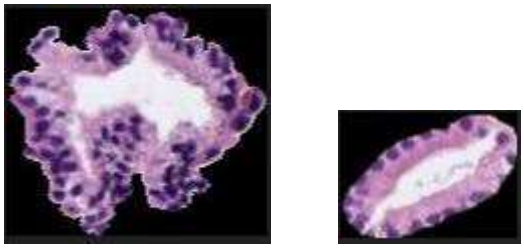
El contenido de cada imagen está codificado en el valor de cada una de las unidades mínimas, los píxeles. Los píxeles guardan la información contenida en la imagen y, por ello, los descriptores de imagen deberán guardar algún tipo de relación o actuar sobre los mismos. Idealmente, un descriptor de imagen debería gozar de ciertas propiedades, algunas son [23]:

- Simplicidad. Debe representar las características de la imagen de forma clara y sencilla con el fin de proporcionar una interpretación fácil.
- Diferenciabilidad. Debe poseer una gran discriminabilidad respecto de otras imágenes. Además, debe poseer información que permita relacionar imágenes parecidas.
- Reproducibilidad. Debe ser independiente del tiempo en el que se genere.
- Eficiencia. Debe consumir una cantidad de recursos que sea asequible, en términos de coste computacional.

Para obtener las características que permitan el aprendizaje del clasificador se emplearán descriptores de bajo nivel, llamados descriptores visuales. Concretamente se usarán descriptores que aporten información acerca de la forma, la textura y el color que presenten las estructuras de interés de las imágenes tisulares.

De cada uno de estos descriptores se obtendrán diferentes características que se irán almacenando en un vector. El vector de cada glándula será el *output* de la función **Feature_Extraction** que se puede observar en el diagrama de la Figura 2.3.

Como se puede apreciar en la Figura 2.9, visualmente existen disimilitudes muy aparentes entre una glándula benigna y una patológica. Dichas diferencias se encuentran resumidas en la tabla 1.2.



(a) (b)

Figura 2.9: (a) Glándula benigna y (b) glándula patológica de grado 3.

2.2.3.1 Descriptores de forma

Los descriptores de forma definen las regiones, contornos y formas de imágenes. En este TFG se hará uso de diferentes tipos de los descriptores mencionados.

Por un lado, se utilizarán los descriptores geométricos, que permiten estudiar la forma geométrica de los objetos y los contornos de las regiones; por ejemplo, el área o el perímetro. Existen algunos problemas debidos a la segmentación de las unidades glandulares que pueden afectar a la calidad de los resultados, como los contornos mal definidos, los salientes incorrectos o los agujeros inexistentes. Por otro lado, se emplearán los descriptores topológicos, que estudian las configuraciones geométricas con propiedades específicas como la invarianza bajo ciertas transformaciones y los cambios de escala; por ejemplo, la compacidad o la suavidad .

En este apartado se obtendrán diferentes características relacionadas con cada una de las glándulas y sus componentes correspondientes. Para ello, se han de usar las máscaras obtenidas tras el *clustering*, como las que se muestran en la Figura 2.7.

Seguidamente, se expone la tabla 2.2 donde se resumen los valores que se han extraído de la glándula y de cada una de las regiones de la misma.

Glándula	Área, Perímetro, Solidez, Redondez, Suavidad, Excentricidad y Compacidad
Lumen	Área, Perímetro, Solidez, Excentricidad, Compacidad, Relación entre el lumen y la glándula y Distancia del centroide al borde del lumen
Núcleos	Área, Relación entre los núcleos y la glándula y Densidad de núcleos
Citoplasma	Área

Tabla 2.2: Descriptores de forma que se emplean.

A continuación, se detallarán los diferentes parámetros obtenidos y las fórmulas con las que son computados.

- **Área**. Número de píxeles que tienen una etiqueta. En este caso, se cuentan los píxeles con valor 1. Este descriptor aporta una idea del tamaño.

  Como se puede observar en la Figura 2.9, generalmente, una glándula benigna posee un tamaño mayor que una maligna.

- **Perímetro**. Número de píxeles del contorno del objeto. Un píxel del contorno es aquel que pertenece al objeto y forma parte del límite del mismo.

  Por el mismo motivo que el descriptor anterior, en la Figura 2.9 se puede apreciar que el perímetro de la glándula sana será mayor que el de la patológica.

- **Solidez**. Se computa como:

$$Solidity = \frac{Area}{ConvexArea} \tag{2.5}$$

donde 'Area' es el parámetro obtenido anteriormente y 'ConvexArea' es el número de píxeles de la 'ConvexImage'. La 'ConvexImage' es una imagen binaria que especifica la 'ConvexHull', que es el polígono convexo más pequeño que puede contener a la región.

- **Redondez**. Se computa como:

$$Roundness = \frac{\sqrt{\frac{Area}{\Pi}} * Perimetro}{Area} \tag{2.6}$$

- **Suavidad**. Define el borde de la región. Se computa como:

$$Smoothness = \frac{Perimetro}{PerimetroBB} \tag{2.7}$$

donde 'Perimetro' es el valor obtenido anteriormente y 'PerimetroBB' es el perímetro de la *BoundingBox*, es decir, el perímetro del rectángulo más pequeño que contiene el objeto.

Como se puede observar en la Figura 2.9 el borde de la glándula benigna es abrupto mientras que el de la de grado 3 es más suave.

- **Excentricidad**. Se define como un parámetro que determina el grado de desviación de una sección con respecto a una circunferencia.

Se computa como la relación de la distancia entre los focos de la elipse y la longitud de su eje principal. Nos devolverá un valor entre 0 y 1. Siendo 0 el valor para un círculo y 1 el de un segmento de línea.

- **Compacidad**. Se trata de un parámetro que no depende del tamaño de la región y viene dado por:

$$Compactness = \frac{Perimetro}{\sqrt{Area}} \tag{2.8}$$

donde 'Perimetro' y 'Area' son los valores obtenidos anteriormente.

- **Relación entre el área del lumen y de la glándula**. Se define como la proporción de lumen que hay en la glándula. Se computa como:

$$RelationLG = \frac{AreaLumen}{AreaGlandula} \tag{2.9}$$

donde ambas áreas han sido extraídas antes.

- **Distancia del centroide al borde del lumen**. Se calculará la distancia euclídea del centroide[5] del lumen a cada píxel del borde del lumen. Para ello, se necesitará obtener, en primer lugar, los píxeles del borde del lumen y se hará de la siguiente manera:

  1. Erosión de la imagen del lumen. Se creará un elemento estructurante y se realizará la erosión de la imagen. La erosión es el valor mínimo del conjunto de píxeles bajo el elemento estructurante.
  2. Se calculará la diferencia entre la imagen del lumen inicial y la imagen erosionada.

[5]Hace referencia al centro de masas de la región.

Una vez obtenidos los píxeles del borde, se calcula la distancia euclídea de la siguiente forma:

$$DistanciaEuclidea_i = \sqrt{(c_x - x_i)^2 + (c_y - y_i)^2} \tag{2.10}$$

donde c_x y c_y son las coordenadas del centroide y x e y son las coordenadas del píxel i.

Se calculará la distancia a cada uno de los píxeles del borde de la estructura y se obtendrá la media y la varianza de todas ellas.

- **Relación entre el área de los núcleos y de la glándula**. Se define como la proporción de núcleos que hay en la glándula. Se computa como:

$$RelationNG = \frac{AreaNucleos}{AreaGlandula} \tag{2.11}$$

donde ambas áreas han sido extraídas antes.

- **Densidad de núcleos**. Se define como la proporción de núcleos que forman parte del citoplasma. Lo computamos como:

$$DensityNucleos = \frac{AreaNucleos}{AreaNucleos + AreaCitoplasma} \tag{2.12}$$

donde ambas áreas han sido obtenidas anteriormente.

Como se observa en la Figura 2.9, es una diferencia aparente que la densidad de núcleos es mayor en las benignas que en las malignas. La glándula benigna presenta varias capas de núcleos, mientras que la patológica de grado 3, solo una.

2.2.3.2 Descriptores de textura

Los descriptores de textura definen la disposición espacial de las intensidad de color en una imagen o región seleccionada. Una región de una imagen tiene una textura constante si el conjunto de estadísticos locales y demás propiedades locales son constantes o varían lentamente [24]. Si se cuantifica la información textural contenida en la imagen se estará en disposión de poder detectar cambios en ella. Existe una gran cantidad de descriptores: los estadísticos, la matriz de co-ocurrencias, el *Histogram Of Gradients* (HOG), los *Local Binary Patterns* (LBP), la granulometría o el *Scale-invariant feature transform* (SIFT), entre otros.

En el presente TFG, se tratarán dos descriptores de textura: la matriz de co-ocurrencias y los LBP. Para la implementación de estos descriptores, se usarán las imágenes de las glándulas sin enmascarar (véase la Figura 2.6.a), ya que las enmascaradas provocarían cambios muy bruscos entre el fondo negro y el borde de la glándula y no se definiría realmente la textura de la imagen.

A continuación, se explica en profundidad cada uno de los tipos de descriptores de textura implementados en el presente TFG.

Matriz de co-ocurrencias

La matriz de co-ocurrenias consiste en una estructura 2D, cuya finalidad es capturar las dependencias espaciales de los valores de gris que contribuyen a la percepción de las texturas presentes en una imagen.

La matriz de co-ocurrencias es una matriz de frecuencias en la que un píxel con un nivel de gris i matiene una relación espacial específica con otro píxel con un nivel de gris j [22]. La matriz de co-ocurrencias permite tener en cuenta parejas de píxeles separadas una distancia d en un determinado ángulo. Por tanto, dicha matriz *P(i,j)* se define especificando una dirección de desplazamiento *d=d(i,j)* y contando todos los pares de píxeles separados por d que tienen los valores de gris i y j.

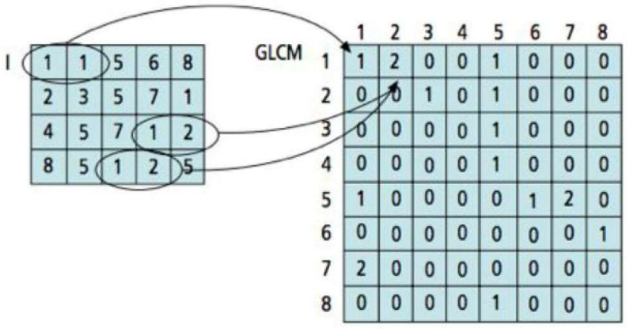


Figura 2.10: Ejemplo sencillo del proceso de construcción de una matriz de co-ocurrencias de 8 niveles de gris con una distancia de separación de 1 píxel y un ángulo de 0º

La hipótesis de partida es que cuando la matriz contiene valores elevados en la diagonal, se está describiendo una textura más suave debido a que los pares de píxeles tendrán valores más próximos, como por ejemplo la imagen de la Figura 2.11.c; mientras que, cuando los valores de la diagonal son bajos se revela una macrotextura con cambios apreciables donde los pares de píxeles presentarán valores de gris muy diferentes, como por ejemplo la imagen de la Figura 2.11.a. Se pueden observar las matrices de co-ocurrencia asociadas a los ejemplos nombrados anteriormente en la Figura 2.11.d y 2.11.b respectivamente.

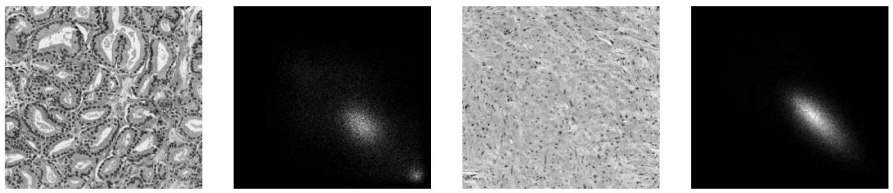


Figura 2.11: Ejemplo de matrices de coocurrencia: (b) es la matriz de (a) y (d) de (c).

En el presente TFG se obtendrá la matriz de co-ocurrencias para la imagen de grises del canal cian y para las imágenes obtenidas de la hematoxilina y la eosina al aplicar la deconvolución de color explicada en el apartado 2.2.2. Además, para cada uno de los tres canales se obtendrá una matriz de co-ocurrencias para un ángulo de 0º y para uno de 45º. Por tanto, se calcularán 6

matrices diferentes de las que se extraerán parámetros intrínsecos que la definen y que serán los utilizados en el entrenamiento del clasificador.

(a) (b) (c)

Figura 2.12: Imágenes de las que se obtendrá la GLMC: (a) imagen original (b) hematoxilina y (c) eosina.

El hecho de que se utilicen los 3 canales que se observan en la Figura 2.12 es porque cada uno de ellos muestra una estructura más realzada que el resto y no se tiene certeza de cuál o cuáles de ellos podrán proporcionar mejores resultados. Por ello, para detectar patrones de textura de forma más precisa será necesario calcular combinaciones de matrices con diferentes canales y direcciones.

La forma de especificar la relación de posición a nivel matricial entre un par de píxeles cuya correlación se quiere estudiar será mediante un vector 1x2 llamado 'Offset' donde el primer valor es el número de filas entre el píxel de interés y su vecino, y el segundo valor es el número de columnas. El desplazamiento se expresa como un ángulo y cada ángulo corresponde a un vector diferente como se observa en la tabla 2.3 donde D es la distancia entre píxeles. La Figura 2.13 ejemplifica como funciona el 'Offset' para una D igual a 1.

Ángulo	Offset
0	[0 D]
45	[-D D]
90	[-D 0]
135	[-D -D]

Tabla 2.3: Ángulo y su respectivo 'Offset'.

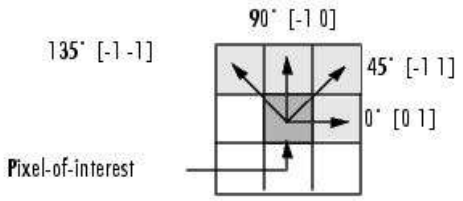


Figura 2.13: Ejemplo de los diferentes 'Offset' [25].

En el presente TFG, se utilizará la matriz de 0º, [0 D], y la matriz de 45º, [-D D], con una D igual a 2. A continuación, se explicarán los diferentes parámetros que se han obtenido de la matriz de coocurrencia (GLCM, del inglés - *Gray Level Coocurrence Matrix*):

- **Contraste**: medida de la cantidad de variaciones locales en los niveles de gris de una imagen completa. El contraste será 0 para una imagen con niveles de gris constantes.
- **Correlación**: medida de la dependencia de un píxel con respecto a su vecino. Mide las variaciones locales de los niveles de gris. Va del rango -1 a 1, siendo estos, valores propios de una imagen perfectamente negativa o positivamente correlacionada.
- **Energía**: medida del orden de la imagen. El rango va de 0 a 1.
- **Homogeneidad**: valor que mide la proximidad de la distribución de elementos en la GLCM respecto a su diagonal GLCM. Mide la homogeneidad local, asignando valores más altos a

las diferencias más pequeñas. El rango va de 0 a 1, siendo 1 para un GLCM diagonal, es decir, una imagen totalmente homogénea.

- **Media**: medida de tendencia central que se calcula mediante la suma de todos los valores y dividiendo entre la cantidad de los mismos.
- **Desviación estándar**: valor numérico que indica la dispersión del conjunto de datos de la matriz.
- **Entropía**: medida estadística de la aleatoriedad que se puede utilizar para caracterizar la textura de una imagen de grises. Dicho de otra manera, es una medida del desorden de la imagen; a mayor desorden, mayor entropía.

Local Binary Patterns (LPB)

El LBP es un descriptor de textura simple pero eficiente que etiqueta cada píxel analizando su vecindario. En general, estudia si el nivel de gris de cada píxel supera un umbral y codifica dicha comparación mediante un número binario. Los LBP presentan un gran poder discriminativo y gracias a su sencillez, su coste computacional es considerablemente bajo. Además, la propiedad más importante que posee es la robustez que ofrece ante las variaciones en las intensidades de los niveles de gris debido a, por ejemplo, las diferencias de iluminación. Por todo ello, en estos últimos años se ha convertido en uno de los operadores más empleados en el análisis de texturas para diferentes aplicaciones [26].

El operador de texturas propuesto permite detectar patrones binarios locales uniformes en vecindarios circulares con cualquier cuantificación del espacio y con cualquier resolución espacial. Se define el caso general basado en un vecindario circular simétrico de P miembros y radio R, se pueden observar algunos ejemplos de diferentes vecindarios en la Figura 2.14. De esta forma, P controlará la cuantificación del espacio angular y R determinará la resolución espacial del operador [26].

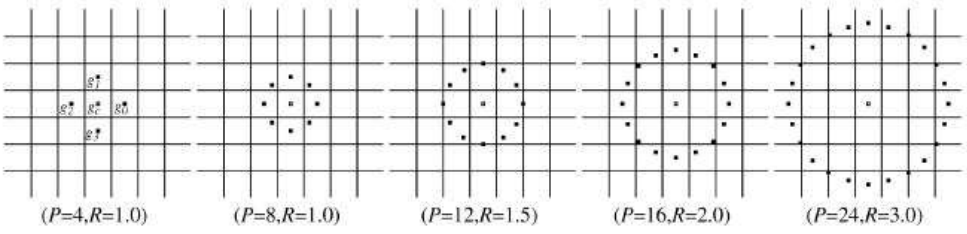


Figura 2.14: Vecindarios circularmente simétricos con diferente (P,R) [27].

El histograma discreto de ocurrencia de los patrones uniformes calculados sobre una imagen o región de la misma constituye una característica con un gran potencial discriminatorio, ya que se combinan enfoques estructurales y estadísticos, tal que el patrón binario local detecta microestructuras.

Se considera que la textura de una imagen está definida bidimensionalmente mediante dos propiedades ortogonales: la estructura espacial (patrón) y el contraste ("cantidad" de textura local en la imagen). Cabe destacar que donde el patrón espacial se ve afectado por la rotación, el contraste no, y donde el contraste se ve afectado por la escala de grises, el patrón espacial no.

En consecuencia, si se restringe al análisis de textura invariante a la escala de grises, el contraste no interesará porque depende de esta [27].

El operador $LBP^{riu2}_{P,R}$ es una excelente medida de la estructura espacial de la textura local de la imagen pero, por definición, descarta la otra propiedad, el contraste. Si solo se requiere análisis textural invariante ante la rotación y no se requiere invarianza de la escala de grises, el $LBP^{riu2}_{P,R}$ puede ser mejorado mediante la combinación con una medida de la varianza que es invariante a la rotación pero que caracteriza el contraste de la imagen, $VAR_{P,R}$. En este proyecto se obtendrán los histogramas de ambos operadores.

Para conseguir la derivación del operador de textura hacia la invarianza rotacional y de escala de grises se define la textura T en un vecindario monocromático como la distribución de los niveles de gris de los píxeles de la imagen:

$$T = t(g_c, g_0, ..., g_{p-1}), \tag{2.13}$$

donde g_0 es el valor de gris del píxel central del vecindario local y g_p para p = 0,...,P-1 corresponde a los valores de gris de píxeles espaciados en un círculo de radio R. Los valores de gris que no caigan exactamente en el centro de los píxeles serán estimados por interpolación.

El primer paso para llegar a la invarianza en la escala de grises es restar, sin perder información, el valor de gris del píxel central a los valores de grises del vecindario circular (véase la Figura 2.15):

$$T = t(g_0 - g_c, ..., g_{p-1} - g_c). \tag{2.14}$$

Este es un operador de textura altamente discriminativo, ya que registra las ocurrencias de varios patrones del vecindario de cada píxel en un histograma. Por ejemplo, para las regiones constantes, las diferencias son cero en todas las direcciones. Los signos de estas diferencias no se ven afectados por los cambios en la luminancia media, por lo que la distribución de la diferencia conjunta será invariable contra los cambios de la escala de grises. Entonces, esto se conseguirá al considerar solo los signos de las diferencias en lugar de los valores exactos:

$$T = t(s(g_0 - g_c), s(g_1 - g_c), ..., s(g_{p-1} - g_c)) \tag{2.15}$$

donde

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{2.16}$$

Asignando un factor binomial 2^p a cada signo de cada diferencia se consigue el operador $LBP_{P,R}$ que se caracteriza por presentar una invariancia frente a cualquier transformación de la escala de grises. Matemáticamente se expresa mediante la siguiente ecuación:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \tag{2.17}$$

donde P es el número de píxeles vecinos alrededor del píxel central que regula la cuantificación del espacio angular, R el radio de la circunferencia que determina la resolución espacial y g_c y g_p son los valores de gris del píxel central y de cada uno de los p píxeles considerados, respectivamente.

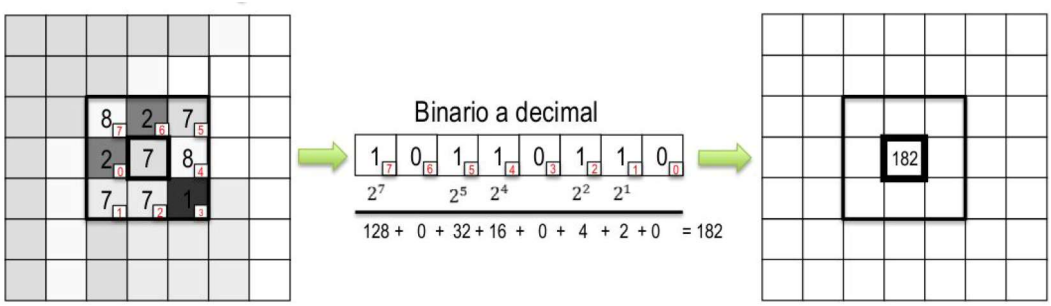


Figura 2.15: Ejemplo de los pasos para obtener el valor del LBP para un píxel concreto [28]

Por tanto, el vecindario local es umbralizado con respecto al valor de gris del píxel central y estos valores son transformados al sistema binario según la ecuación 2.17.

Cuando la imagen sufre una rotación, los valores de gris g_p se moverán a través del perímetro del círculo, es decir, alrededor del píxel central. Dado que g_0 siempre es el elemento a la derecha del píxel central, la rotación proporciona como resultado un valor final diferente, salvo que sean todo 0's o 1's.

Para eliminar este efecto y conseguir invarianza a la rotación, se asigna un único valor a cada patrón binario local, se define

$$LBP^{ri}_{P,R} = \text{mín}\{ROR(LBP_{P,R}, i)|i = 0, 1, ..., P-1\} \tag{2.18}$$

donde *ROR(x, i)* es un desplazamiento hacia la derecha de i píxeles para el patrón x.

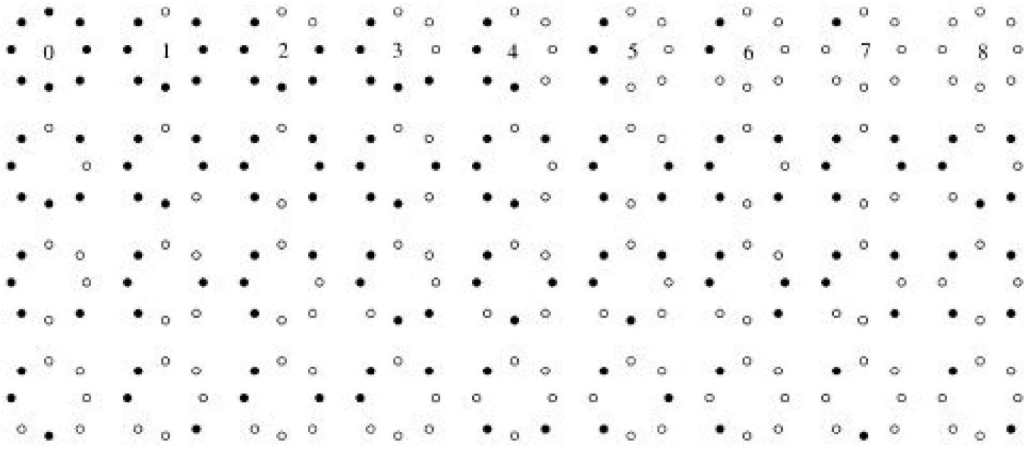


Figura 2.16: Los 36 patrones binarios únicos invariantes a rotación que pueden ocurrir en un vecindario simétrico circular de P=8. Los círculos blancos corresponden a valores de bit 1 y los negros, a 0. La primera fila contiene los 9 patrones uniformes [27].

$LBP^{ri}_{P,R}$ cuantifica las estadísticas de ocurrencia de los patrones invariantes de rotación correspondientes a ciertas microtexturas de la imagen, por tanto, los patrones se pueden considerar detectores de características.

Con el fin de reducir el número de patrones posibles a los realmente discriminantes, se definen los patrones "uniformes", estos tienen en común una estructura circular uniforme que presenta muy pocas transiciones espaciales. Para definir estos patrones se introduce la medida de uniformidad U que corresponde al número de transiciones espaciales (cambios de bit 1/0) en el patrón. En [27] se demostró que los patrones con un número de transiciones espaciales menor o igual que 2 tenían una capacidad de discrimación superior, por lo que dichos patrones fueron definidos como "LBP uniformes" dando lugar a la siguiente expresión matemática:

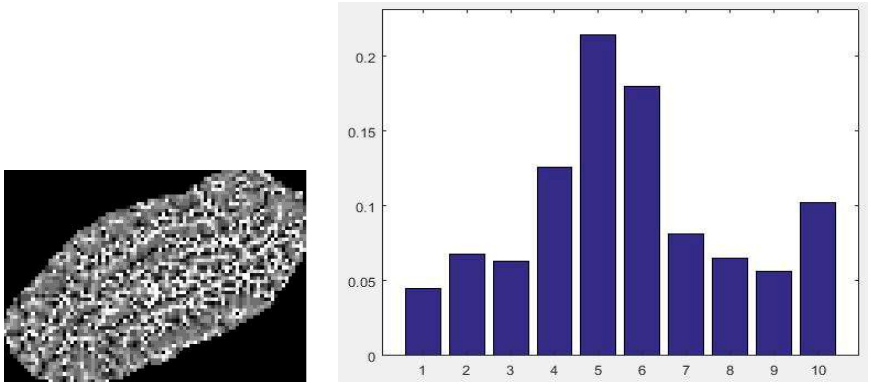
$$LPB^{riu2}_{P,R} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c), & \text{si } U(LBP_{P,R)} \leq 2 \\ P+1, & \text{el resto ,} \end{cases} \tag{2.19}$$

donde

$$U(LBP_{P,R)} = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=0}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)|. \tag{2.20}$$

De esta manera, existen, exactamente, P+1 patrones binarios "uniformes" posibles en un vecindario circular simétrico de P píxeles, siendo cada uno de ellos etiquetados en el rango de 0 a P. En la Figura 2.16 los casos de la primera fila corresponden a los P+1 patrones uniformes, ya que el número de transiciones espaciales es menor que 2, mientras que los 27 restantes son etiquetados como P+1 dentro del grupo de patrones no uniformes. Por tanto, el operador $LBP^{riu2}_{P,R}$ tiene P+2 valores de salida distintos: de 0 a 8 para los patrones uniformes y la etiqueta P+1 para los no uniformes.

Finalmente, para describir el análisis final de la textura, se usará el histograma de 10 barras de las etiquetas de los patrones acumulados sobre una muestra (véase la Figura 2.17).



(a) (b)

Figura 2.17: (a) Representación de la textura obtenida mediante $LBP^{riu2}_{P,R}$ de una glándula prostática maligna y (b) su correspondiente histograma.

La razón por la que el histograma de patrones "uniformes" proporciona la mejor discriminación es debido a que reduce las diferencias en sus propiedades estadísticas. La proporción de patrones

"no uniformes" de todos los acumulados en un histograma es tan pequeño que no se puede estimar de forma confiable.

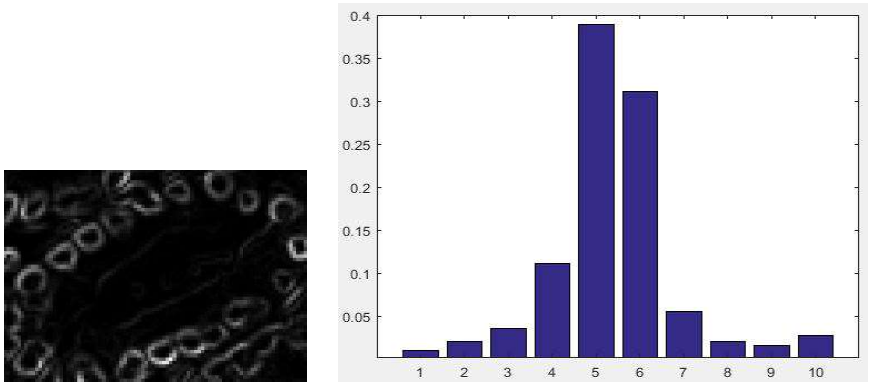
Como ya se ha mencionado anteriormente, el $LBP^{riu2}_{P,R}$ es un operador invariante ante la escala de grises. Es una medida excelente del patrón espacial pero, por definición, descarta el contraste, el cual sí depende de la escala de grises. Como en el presente proyecto se pretende incorporar información relacionada con los cambios de contraste de la textura local, se añadirá una medida de la varianza local invariante ante la rotación que se expresa mediante el operador $VAR_{P,R}$ (*Rotational Invariant Local Variance*):

$$VAR_{P,R} = \frac{1}{P}\sum_{p=0}^{P-1}(g_p - \mu)^2, \tag{2.21}$$

donde

$$\mu = \frac{1}{P}\sum_{p=0}^{P-1} g_p. \tag{2.22}$$

De la misma manera, se pretende obtener las métricas relacionadas con el histograma normalizado que proporciona la combinación de las imágenes $LBP^{riu2}_{P,R}$ y $VAR_{P,R}$ (véase la Figura 2.18).



(a) (b)

Figura 2.18: (a) Representación de la textura obtenida mediante $VAR_{P,R}$ de una glándula prostática maligna y (b) su correspondiente histograma.

$VAR_{P,R}$ y $LBP^{riu2}_{P,R}$ son complementarias y se espera que su distribución conjunta sea una medida poderosa que proporcione resultados interesantes para la discriminación entre glándulas sanas y patológicas de grado 3.

2.2.3.3 Descriptores de color

El espacio de color se define como un modelo de representación del color con respecto a los valores de intensidad. La dimensionalidad puede estar comprendida de una a cuatro dimensiones, siendo el espacio de color más representativo el RGB. Por consiguiente, en este TFG se hará uso de los tres canales asociados al color rojo, al verde y al azul para extraer características de interés.

Además, se implementarán descriptores globales, los cuales resumen el contenido de la imagen en un vector de características. Tienen la ventaja de encapsular una gran cantidad de información en una pequeña cantidad de datos que describen la imagen. Este tipo de descriptores se encuentra ampliamente en la literatura, como se expone en la tabla 1.3, debido a las ventajas que presenta como, por ejemplo, su bajo coste computacional.

En primer lugar, se obtendrá la media y la desviación típica de las intensidades de color de cada canal y se almacenarán los valores obtenidos para utilizarlos como características de la imagen.

A continuación, se obtendrá el histograma de color, el cual representa la frecuencia de aparición de cada una de las intensidades presentes en la imagen, se contabilizan los píxeles que comparten un valor o unos valores de intensidad de color. Está compuesto por diferentes rangos que representan un valor o conjunto de valores de intensidad. Previamente a la contabilidad de los píxeles, existe una etapa de cuantificación de los intervalos con los que representaremos el histograma, este proceso consiste en agrupar intensidades de color cuyos valores se encuentran próximos entre sí [23]. Mediante este etapa, se reduce la información considerablemente y consecuentemente, el tiempo de cálculo también.

En el presente TFG, se calculará un histograma para cada uno de los canales de color RGB y se reducirá de 256 niveles de gris a 5 conjuntos diferentes de intensidades. Por tanto, se agruparán una gran cantidad de valores por cada intervalo, lo cual conllevará una reducción del poder discriminatorio del descriptor. No obstante, es necesario proceder de esta manera, ya que existe un compromiso entre la capacidad discriminatoria y el coste computacional.

Finalmente, de los descriptores de color se obtendrán 21 características diferentes de cada glándula.

Una vez se han obtenido todas las características anteriormente mencionadas, se generará un vector con los diferentes valores para cada parámetro. Dicho vector se añadirá a la matriz *FEAT_SELECTION* donde cada fila corresponderá a una glándula diferente y cada columna a una característica. Además, se utilizará una columna de la matriz para definir la etiqueta correspondiente al *goundtruth* (GT): 1 para sanas y 2 para patológicas. Finalizado este proceso, se obtendrá una tabla de características que poseerá un tamaño de 1939x107.

2.2.4 Partición de los datos

El paso que se explica en la presente sección es de mera importancia pues permitirá diseñar un modelo robusto y proporcionar resultados de manera fiable.

Cabe destacar que el número total de glándulas estudiadas se encuentra balanceado siendo 1024 malignas y, 913 benignas. Una vez obtenidas las características de las 1939 glándulas, se utilizarán el 80 % de ellas para la selección de características que serán utilizadas como entrada durante la etapa de entrenamiento del clasificador y se llamarán *'training_data'*; y el 20 % restante

de las glándulas será utilizado para analizar la bondad del clasificador mediante una etapa de predicción, se llamarán *'test_ data'*.

El *script* desarrollado procederá de forma que el conjunto de muestras *'test_ data'* serán seleccionadas aleatoriamente y permitirán medir la calidad del algoritmo. Tras la partición de los datos se obtendrá la matriz *'training_ data'* que contiene 1552 muestras y la matriz *'test_ data'* con 387 muestras.

2.2.5 Selección de características

2.2.5.1 Método de selección de características

Una vez terminado el proceso de extracción de características, será necesario realizar un análisis estadístico que permita escoger cuál es el espacio de características más discriminativo y de esta forma, eliminar aquellas características que no aportan información relevante o son redundantes e incluso pueden empeorar el rendimiento de la clasificación. El objetivo será seleccionar el subconjunto de elementos que minimiza una función de separabilidad o que mejor clasificación proporciona.

El análisis estadístico o proceso de selección de características se realizará solo sobre la matriz *'training_ data'*, ya que está sera la empleada en el entrenamiento del clasificador. Como ya se ha mencionado anteriormente, está matriz consta de 1552 filas correspondientes a glándulas benignas y patológicas, y 107 columnas que comprenden 102 características y otros datos de interés, como:

- Número de la imagen en el que se encuentra la glándula bajo estudio.
- Centroides x e y de la glándula.
- Variable respuesta o *groundtruth*. Vector que posee la clase correspondiente a cada observación. Resulta muy importante tanto para el análisis estadístico como para la etapa de la clasificación supervisada, en la cual, el clasificador 'aprenderá' mediante las etiquetas de los datos.

Una vez se han separado las diferentes matrices y vectores de interés, se procederá al análisis estadístico siguiendo los siguientes pasos:

1. **Normalización** o estandarización de las características mediante el cálculo del *zscore*. La expresión es la siguiente:

$$z = \frac{(x - \overline{X})}{S}, \tag{2.23}$$

donde x es la característica original, z, la normalizada y $\overline{X}$ y S son la media y la desviación típica de todos los valores de dicha característica, respectivamente.

Tanto el valor de la media como el de la desviación estándar utilizados para centrar y escalar los datos serán empleados posteriormente para normalizar valores nuevos con respecto al mismo grupo de datos.

2. **Selección de parámetros**:
   - Nivel de confianza con el que se acepta o se rechaza la H0. Concretamente, en este TFG se define el nivel de confianza como *trust = 0.9999*
   - Nivel de significancia. Se define como α = *1-trust*. De esta forma, si el p-valor calculado para una hipótesis nula es menor que el nivel de significancia establecido, se rechaza la hipótesis nula. De lo contrario, no habría evidencia para rechazar la hipótesis nula y por tanto, se asumiría la hipótesis alternativa.
   - Umbral para eliminar variables que están correladas. Que definido mediante *th_ cor=0.9*
3. **Capacidad discriminatoria**. En este apartado y antes de empezar el estudio, se separarán las variables con respuesta negativa (benignas) de las de respuesta positiva (malignas).
4. **Media y desviación estándar**. Con el objetivo de analizar visualmente la capacidad discriminatoria de las variables, se representan la media y la desviación estándar de cada una de ellas diferenciando en función del tipo de glándula. Es decir, para cada variable, se representan en rojo los valores de media y desviación estándar para una glándula patológica, y en azul, para una glándula sana.

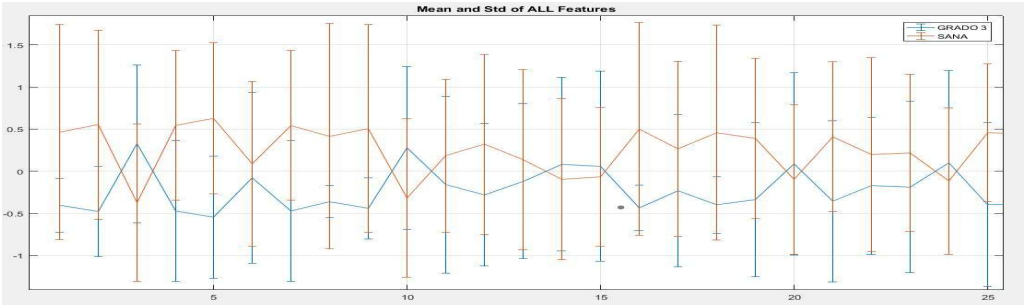


Figura 2.19: Representación de la media y la desviación típica de las variables 1-25 para glándulas benignas y patológicas.

Se puede observar en la Figura 2.19 que las medias de las características son diferentes en función de su malignidad, es decir, existen parámetros cuyas medias se encuentran más separadas que otras. Estas serán las variables que permitirán discernir con más precisión el carácter maligno de una glándula.

5. ***Box and whisker***. Mediante este diagrama se podrá ver una representación visual de la capacidad discriminatoria de las variables, permite observar el solapamiento de los datos, es decir, para una misma característica se enfrentan los valores que toma en función de la clase. Además, describe varias propiedades al mismo tiempo, como la dispersión y la simetría. Para su realización se representarán los tres cuartiles y los valores mínimo y máximo de los datos sobre un rectángulo.

   Como se puede observar en la Figura 2.20 existen variables que visualmente separan a la perfección ambos grupos, estas poseen medianas diferentes y presentan un porcentaje de solapamiento considerablemente bajo en función de la etiqueta del *groundtruth*.

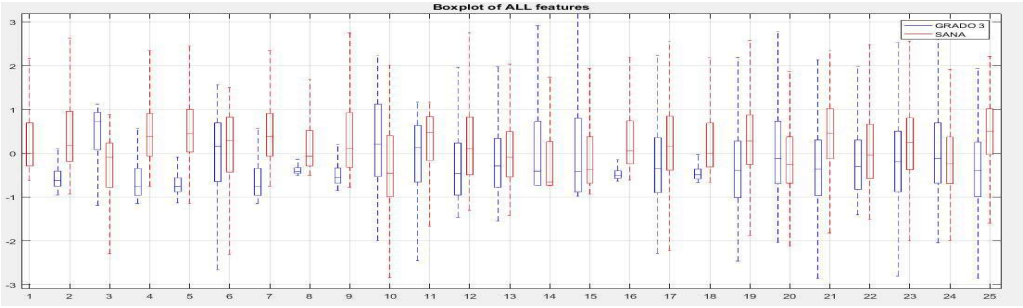


Figura 2.20: *Box and whisker* de las variables 1-25 para glándulas benignas y patológicas.

6. **Estudio de la normalidad de las características**. Resulta un paso de vital importancia ya que el objetivo del mismo será estudiar la distribución de las variables, ya que en función de ello se procede de una manera u otra a la hora de analizar su capacidad discriminatoria. Para ello,sse realizará el test de *Kolmogorov-Smirnov* (KS test).

   El KS test es una prueba no paramétrica de la hipótesis nula (H0) de una población, siendo la H0 la normalidad de las variables. Se trata de comparar la distribución muestral con la resultante de dar por cierta la hipotética distribución de la población, es decir, compararla con una distribución de probabilidad de referencia que en este caso será la distribución normal. Esta prueba analizará cada columna de la matriz de características, *features*, y resultará en un valor escalar en el rango [0 1] al que se le nombrará como *p-value*.

   El *p-value* indica la probabilidad de la variable de ser posible bajo la H0. Valores pequeños del *p-value* indican dudas sobre la validez de la H0. Este parámetro será comparado con el alpha del apartado de selección de parámetros de forma que:

   - Si p_norm <= alpha se rechazará la H0 y se considerará que la variable no es normal.
   - Si p_norm >alpha se aceptará la H0 y se considerará que la variable sigue una distribución normal.

   Se puede observar en la Figura 2.21 la representación de la distribución de dos variables y una línea roja superpuesta correspondiente a la Campana de Gauss que marca la tendencia que seguiría una distribución normal; como se puede ver a priori, la segunda de ellas se asimila mucho a la distribución normal mientras que la otra difiere considerablemente.

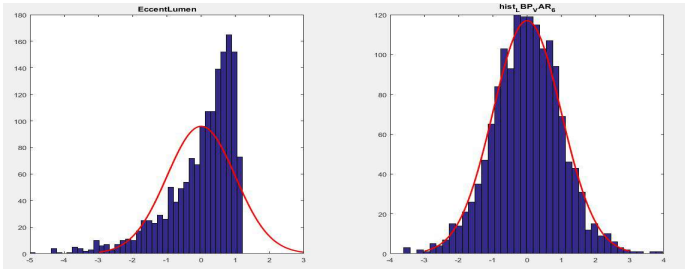


Figura 2.21: Representación de la distribución de dos variables ejemplo.

7. **Capacidad discriminatoria de las variables según su normalidad**. En este apartado se estudiará si las variables resultan de utilidad para separar las clases (glándulas benignas y patológicas). Se realizarán dos test estadísticos diferentes dependiendo de los resultados del apartado anterior, de forma que se usará el test *t de Student* para las variables que sigan una distribución normal y el test *Wilcoxon rank sum* para las que no.

   Cada test planteará una hipótesis nula diferente y del resultado de cada uno se obtendrá un *p-value* que se comparará con el alpha escogido. De este modo, si es menor, se rechazará la H0 con un porcentaje de confianza del 99.99 %, y si es mayor, se aceptará la H0 y se eliminará dicha variable del espacio de características.

   - *T de Student*. Este test será el utilizado cuando se demuestre que una variable sigue una distribución normal. En este, se compararán las medias de cada variable para ambas clases. La H0 sostiene que los datos de un vector y de otro provienen de muestras aleatorias independientes de distribuciones normales con medias iguales y la hipótesis alternativa de que provienen de poblaciones con medias desiguales.
   - *Wilcoxon rank sum*. Este test será utilizado cuando se demuestre que una variable no sigue una distribución normal. En este, se compararán las medianas de cada variable para ambas clases. La H0 defiende que los datos para una clase y para la otra son muestras de distribuciones con medianas iguales.

   Una vez realizado el estudio de la capacidad discriminatoria de las características, el algoritmo habrá eliminado las variables pertinentes y el espacio de características se habrá reducido.

8. **Cálculo de la correlación 2 a 2 de las variables**. Se obtendrá la matriz de coeficientes de correlación y la matriz de *p-values* para probar la hipótesis nula de que no hay relación entre las características obtenidas. Se comparará la tabla de características donde las filas representan observaciones y las columnas, variables, y como resultado se obtendrá la matriz de coeficientes de correlación (R) y la matriz de *p-values* (p_cor) que contiene los valores de p en el rango [0 1] para cada par de variables, tal que valores cercanos a 0 corresponden a una correlación significativa.

   Una vez se han obtenido los valores nombrados anteriormente, se obtendrá la Figura 2.22 donde se puede analizar visualmente cuan correlacionado está cada par de variables. Así, se procederá a comparar los coeficientes de correlación con el th_cor escogido y el *p-value* asociado con el nivel de significancia elegido. De forma que, si $R > th\_cor \cap p\_cor < \alpha$ se considerará que la variable está muy correlada y se eliminará de nuestro espacio de características.

De esta forma, es posible eliminar la información redundante que presentan ciertas características, es decir, es posible descartar del estudio aquellas variables que no aportan información novedosa. Cabe destacar que el proceso explicado en este apartado es fundamental para facilitar la clasificación y aumentar el rendimiento del algoritmo; por ello, es muy importante escoger correctamente los parámetros del paso 2.

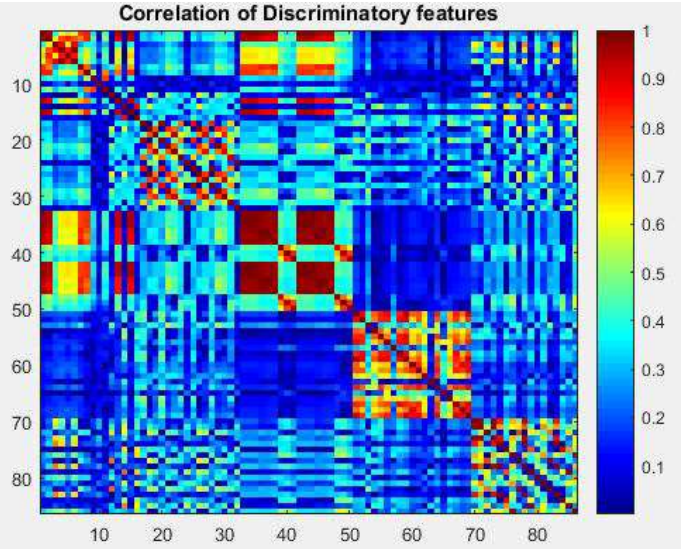


Figura 2.22: Representación de la matriz de coeficientes de correlación de las características.

2.2.5.2 Características seleccionadas

Como ya se ha comentado, el proceso anteriormente descrito permitirá descartar del conjunto total de características extraídas aquellas cuyo poder discriminatorio entre clases sea bajo. Este proceso constituye principalmente dos etapas en las cuales se eliminan características: una en la que se estudia la capacidad discriminatoria de las variables en función de su media o mediana y otra en la que se analiza la correlación entre pares de características para evitar que haya información redundante. De esta forma, tras realizar el análisis estadístico, el espacio de características se habrá reducido de 102 características a 49.

En primer lugar, tras realizar el estudio de la normalidad se obtendrá que 65 características no siguen distribuciones normales, mientras que 37, sí. Por consiguiente, se comparará la mediana de las primeras y la media de las segundas para determinar la capacidad discriminatoria de las variables. Y tras esta etapa, el espacio de característica se verá reducido de 102 a 86 características.

En segundo lugar, tras el cálculo de la correlación dos a dos de características, se eliminarán las características que presenten un porcentaje de información correlacionada superior al 90 % y el espacio pasará de 86 a las 49 características finales. Así, las variables seleccionadas se pueden observar en la Tabla 2.4.

De las 49 características finales seleccionadas se puede observar en la Figura 2.23 la distribución que sigue cada una, mostrando en rojo la Campana de Gauss característica de una distribución normal.

Tipo de descriptor	Variables
Forma	Densidad de núcleos
	Área, Solidez, Redondez y Suavidad de la glándula
	Área, Perímetro, Solidez, Excentricidad y Compacidad del lumen
	Relación entre el lumen y la glándula
	Relación entre los núcleos y la glándula
Textura	Correlación, Homogeneidad, Energía y Entropía de la GLCM de 0º del canal cian
	Correlación de la GLCM de 45º del canal cian
	Energía y Entropía de la GLCM de 0º del canal hematoxilina
	Correlación, Energía y Entropía de la GLCM de 0º del canal eosina
	Histograma LBP 1, 2, 3, 4, 5, 6, 7 y 8
	Histograma VAR 1, 2, 3, 4, 5, 6, 7, 8 y 9
Color	Media de las intensidades del canal rojo
	Desviación típica de las intensidades del canal verde
	Histograma del canal rojo 1, 2, 4 y 5
	Histograma del canal verde 2, 3 y 4
	Histograma del canal azul 2

Tabla 2.4: Características seleccionadas.

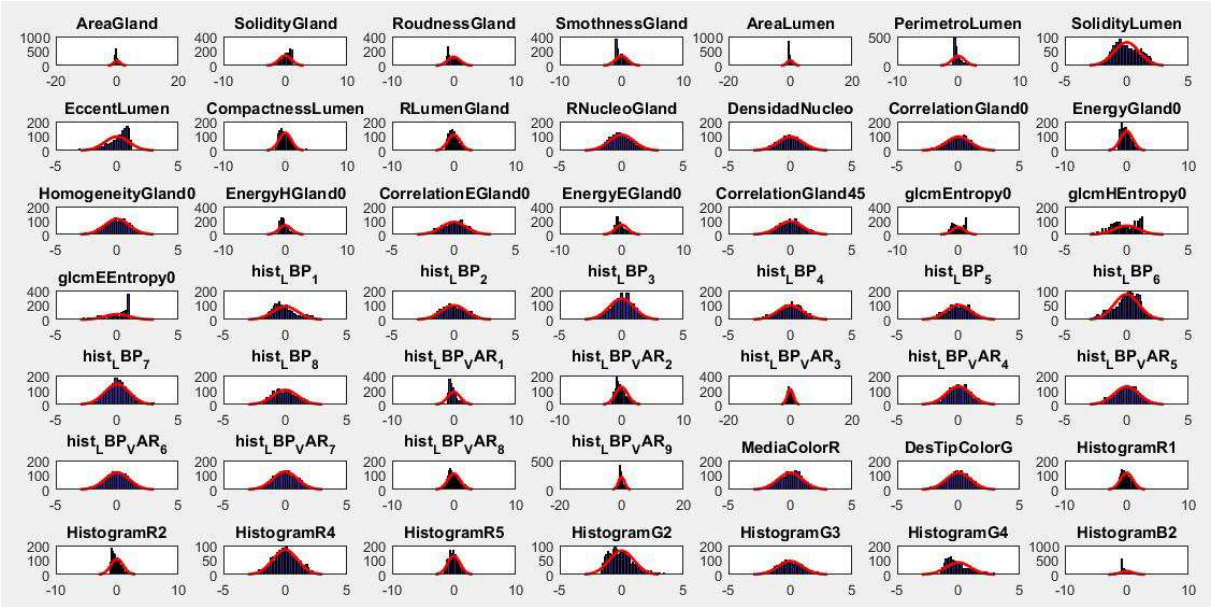


Figura 2.23: Normalidad del espacio de características final.

Una vez se ha diferenciado entre variables normales y no, se obtendrá como resultado que 34 no siguen una distribución normal y se compararán sus medianas y 15 si la siguen y se compararán sus medias. Se puede observar la media y la desviación estándar de las variables que siguen una distribución normal en la Figura 2.24 y la mediana y los cuartiles de las variables que no siguen una distribución normal en la Figura 2.25.

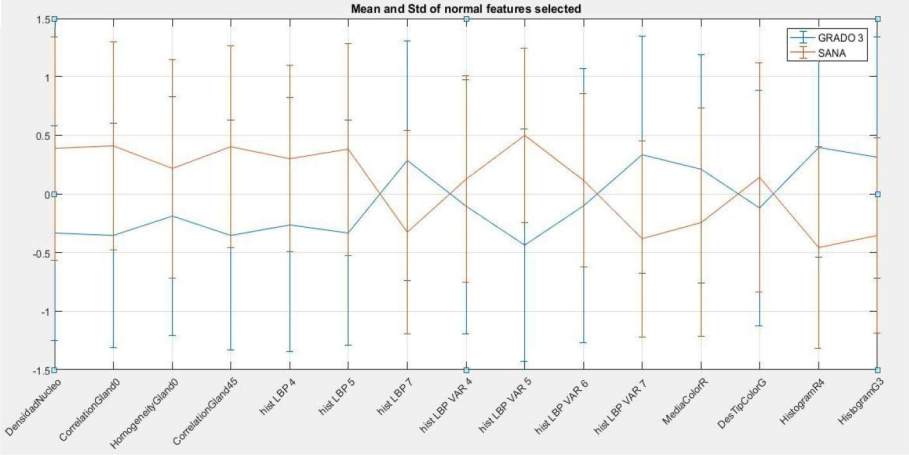


Figura 2.24: Representación de la media y la desviación estándar de las variables normales seleccionadas.

Como se puede observar en la Figura 2.24, las medias de las variables son diferentes en función de si se trata de glándulas sanas o de patológicas. Por ello, y tras haber realizado el test *T de Student*, se puede afirmar con un 99'99 % de confianza que estas variables poseen medias diferentes para cada clase. Por ejemplo y basándose en la gráfica, se puede apreciar que variables como la densidad de núcleos o la barra 4 del histograma del canal rojo diferencian bastante bien entre glándulas sanas y patológicas. Cabe destacar que la densidad de núcleos se calcula en este proyecto de forma novedosa y da aparentes resultados satisfactorios.

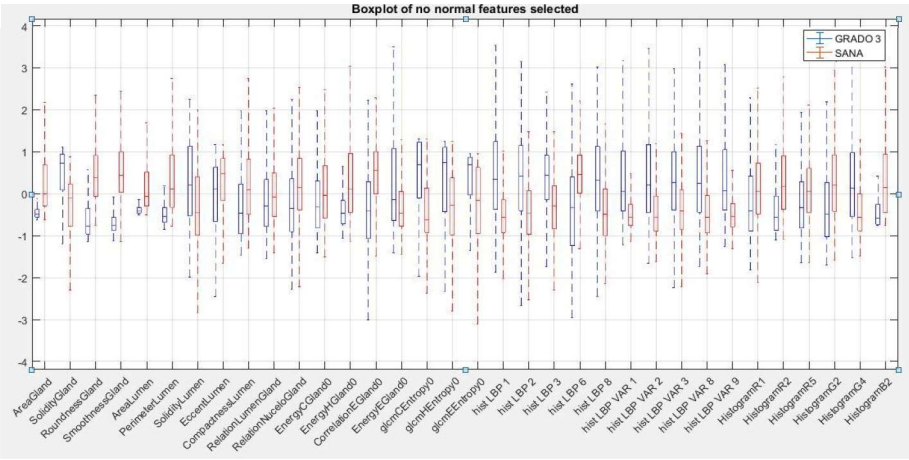


Figura 2.25: *Box and whisker* de las variables no normales seleccionadas.

El algoritmo selecciona más variables que no siguen una distribución normal. En la Figura 2.25, se demuestra visualmente que las variables seleccionadas son considerablemente dependientes de la clase, es decir, toman valores diferentes en función de si se trata de una glándula sana o patológica de grado 3. Dentro de estas, hay algunas que presentan mayor capacidad discriminatoria que otras y que contribuirán con mayor precisión a la clasificación. Algunos ejemplos son el área y la redondez de la glándula, así como algunas barras del histograma LBP.

Finalmente, se puede observar en la Figura 2.26 que, efectivamente, la correlación de las variables seleccionadas no supera el umbral propuesto.

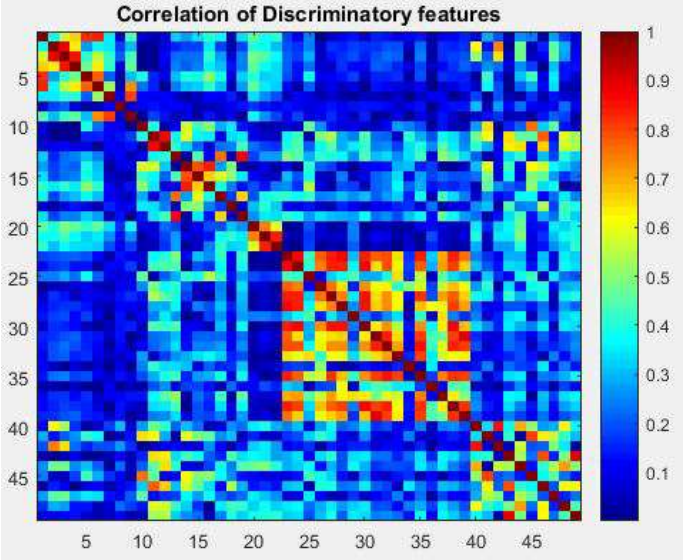


Figura 2.26: Correlación entre pares de variables seleccionadas.

2.2.6 Clasificación supervisada

Las técnicas de clasificación permiten agrupar muestras de acuerdo a criterios o métodos. El objetivo de la clasificación mediante el aprendizaje automático es la asignación de una observación a una de las diferentes clases [29]. La clasificación supervisada cuenta con información a priori, es decir, con muestras ya clasificadas que le servirán de ejemplo para construir el modelo general de clasificación.

2.2.6.1 Familias de clasificadores

Existe una gran variedad de familias de clasificadores dentro de los métodos de clasificación supervisada. En este TFG, se probarán varios modelos de clasificadores durante la etapa de entrenamiento, con el objetivo de analizar cuál de ellos presentan un porcentaje de precisión superior. A continuación, se explican brevemente los tipos de clasificadores implementados.

Árboles de decisión

Los árboles de decisión son uno de los posibles enfoques para la toma de decisiones por múltiples etapas. La idea básica es convertir una decisión compleja en una unión de varias decisiones más simples.

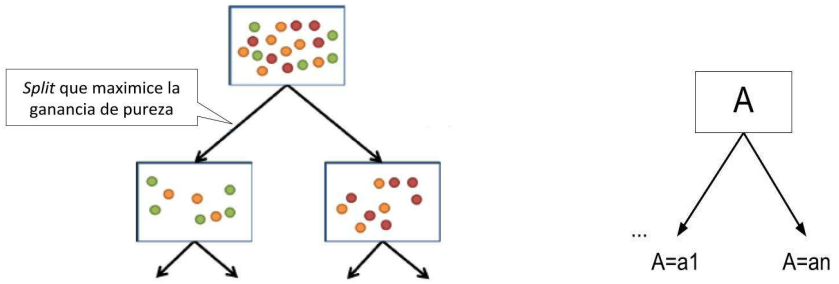
Los árboles de decisión constituyen uno de los métodos de aprendizaje inductivo[6] supervisado no paramétrico más utilizado, son estructuras resultantes de la partición recursiva del espacio de representación a partir de un conjunto de prototipos [28]. Los algoritmos para su construcción suelen trabajar de manera *top-down*[7], escogiendo en cada paso la variable que mejor divide el conjunto de elementos. Los árboles puede ser entendidos como un conjunto de reglas, donde cada

[6]Es un proceso que parte de la observación y el análisis de una característica con el fin de formular una regla que explique dicha característica.

[7]Estrategia de procesamiento de información que va desde la información global hasta la más detallada y específica.

nodo del árbol es una característica (o variable) y cada rama representa un posible valor de la misma.

Cada nodo tiene asociado un subconjunto de datos de entrenamiento e inicialmente, el nodo raíz tiene todo el conjunto de datos que tras la operación de expansión se particionará, como se observa en la Figura 2.27.a, según el número de nodos hijos de acuerdo a un *split* o condición de separación de los datos. Un *split* es una variable o atributo junto con una lista de condiciones sobre la misma, como por ejemplo el de la Figura 2.27.b. La parte compleja del algoritmo reside en encontrar el mejor *split* y, por ello, existen distintos métodos para evaluarlos buscando que los nodos hijos presenten la mayor pureza posible. Algunos de los criterios de selección son: el índice de Gini, la entropía (ganancia de información) o el test Chi-cuadrado [30].



(a) (b)

Figura 2.27: (a) Ejemplo de una partición de los datos en un árbol de decisión. (b) Ejemplo de un *split* de un árbol de decisión.

Existe una gran variedad de árboles de decisión que se clasifican en función de su complejidad. Para este proyecto se han estudiado varios árboles de decisión:

- *Simple Tree*. Número máximo de divisiones es igual a 4.
- *Medium Tree*. Número máximo de divisiones es igual a 20.
- *Complex Tree*. Número máximo de divisiones es igual a 100.

Todos ellos tienen propiedades en común: son rápidos a la hora de predecir, ocupan poco espacio en la memoria y son fáciles de interpretar. Además, todos emplean como criterio de selección de los *splits* el índice de Gini.

El índice de Gini se expresa de la siguiente manera:

$$Gini(t) = \sum_{k=1}^{j} p(k|t)(1 - p(k|t)) \tag{2.24}$$

donde p(k|t) es la probabilidad de que un caso asociado al nodo t sea de la clase j.

Los árboles de decisión presentan ventajas y limitaciones. Por una parte, emplean reglas de decisión fáciles de interpretar, son no paramétricos, requieren poca preparación de los datos, son capaces de manejar tanto datos numéricos como categóricos, utilizan un modelo de caja blanca[8],

[8]Quiere decir que si una situación dada es observable en un modelo entonces la condición se explica fácilmente.

son robustos frente a *outliers*[9] y pueden analizar grandes conjuntos de datos utilizando recursos estándares en un plazo de tiempo razonable [28]. En contra parte, tienden a sobreajustar, realizan divisiones perpendiculares en el espacio de características que no siempre son las mejores, no poseen los mejores aciertos y las pequeñas variaciones de los datos pueden cambiar mucho el árbol [31].

Análisis discriminante

Se trata de modelos generativos donde el clasificador se representará mediante un conjunto de funciones discriminantes. El espacio muestral donde se representan las instancias observadas quedará dividido en regiones de decisión separadas por las fronteras de decisión que están formadas por aquellos puntos para los que se cumple: $g_i{=}g_k$ [31].

Los modelos generativos calculan la probabilidad de que una muestra pertenezca a una clase o a otra mediante el teorema de Bayes, que se define de la siguiente manera:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)}, \text{donde } p(x) = \sum_{c' \in C} p(x|c')p(c') \tag{2.25}$$

Los resultados finales varían en función de los supuestos que hacemos sobre la probabilidad condicionada, *p(x/c)*. En este TFG se estudiarán:

- Análisis Discriminante Lineal (LDA). El LDA asume que los datos de cada clase siguen una distribución Gaussiana y que las matrices de covarianza son iguales para los datos de diferentes clases, por tanto, la frontera de clasificación será lineal. LDA falla cuando la información discriminante se encuentra en la varianza y no en la media. Sigue la regla de Bayes que dice que se debe elegir la clase que maximice la probabilidad a posteriori, dado el vector de características.
- Análisis Discriminante Cuadrático (QDA). El QDA sigue el mismo razonamiento, pero no asume que las matrices de covarianza tengan que ser idénticas, por ello, las fronteras de clasificación serán cuadráticas. QDA presenta una mayor capacidad de ajuste, pero esto conlleva más parámetros que ajustar y el consecuente problema de sobreajuste[10].

Regresión logística

Es el modelo discriminativo más empleado. Consiste en un clasificador binario que solo puede separar dos clases. Calcula la probabilidad de que una muestra x pertenezca a una clase o a otra.

$$g(x, w) = \log(\frac{p(y=1|x)}{p(y=0|x)}) \tag{2.26}$$

Es un modelo generativo, por este motivo, solo está interesado en la definición de la frontera. La frontera entre dos clases se definirá a partir de aquellos puntos en los que la probabilidad de pertenecer a una clase o a otra sea la misma, lo cual se expresa de la siguiente manera:

[9] Valores atípicos.

[10] Efecto producido debido a un entrenamiento excesivo del sistema que provoca que el algoritmo quede ajustado a unas características muy específicas de los datos del entrenamiento y no sea capaz de predecir el resultado de otras observaciones. También se le conoce como *overfitting*.

$$P(G = k|X = x) = P(G = l|X = x) \tag{2.27}$$

siendo k y l las diferentes clases y x los puntos que definen la frontera.

En la regresión logística se forzará a tener una frontera lineal y consecuentemente, la ecuación de partida será la denominada distribución logística:

$$P(y = 1|x) = \frac{\exp^{b_0+\sum_{i=1}^{n} b_i x_i}}{1 + \exp^{b_0+\sum_{i=1}^{n} b_i x_i}} \tag{2.28}$$

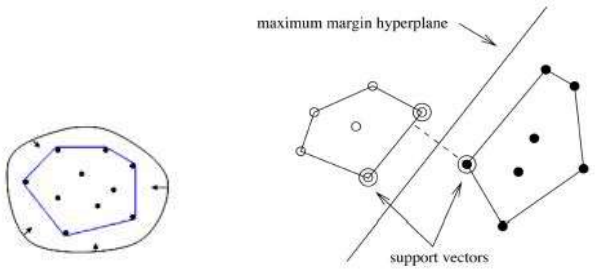
donde *P(y=1/x)* es la probabilidad de que la variable y tome el valor 1 en presencia de las covariables x siendo X un conjunto de n variables, b_0, la constante del modelo y b_i, los coeficientes de las covariables.

La distribución logística es similar a la distribución normal en su forma, pero tiene menos curtosis[11]. La regresión logística es la solución óptima que separa dos normales obtenidas por combinación lineal de las variables de entrada.

Máquina de vectores soporte (SVM)

Un modelo de clasificador basado en SVM permite definir un hiperplano óptimo en forma de superficie de decisión, de modo que el margen de separación entre las clases se amplía al máximo. Esta superficie se conoce como hiperplano de margen máximo, el cual proporciona la máxima separación entre dos clases linealmente separables, siguiendo los siguientes pasos [32]:

1. Calcular la envolvente convexa para cada clase, siendo la envolvente convexa el conjunto convexo mínimo que contiene a un conjunto de puntos y, además, la intersección de las envolventes de clases diferentes es un espacio vacío. Se puede observar un ejemplo en la Figura 2.28.a.
2. Determinar el segmento más corto que une ambas envolvente.
3. La perpendicular del segmento anterior en el punto medio es el hiperplano de margen máximo. Se puede observar un ejemplo en la Figura 2.28.b.



(a) (b)

Figura 2.28: (a) Ejemplo de envolvente convexa. (b) Ejemplo de hiperplano de margen máximo.

[11] Medida de cuán escarpada o achatada está una curva o distribución. A mayor curtosis, más pronunciada es la curva.

Los Vectores Soporte se refieren a un subconjunto de las observaciones de entrenamiento que es usado como soporte para la obtención de la ubicación óptima de la superficie de decisión, son las instancias de cada clase más próximas al hiperplano. El conjunto de vectores define de forma única del hiperplano óptimo.

Si las clases no son linealmente separables se debe realizar una transformación no lineal del espacio de entrada especificando el tipo y tamaño del *kernel* y, a continuación, se debe resolver el problema de optimización para ajustar el hiperplano ótpimo que permita clasificar las características transformadas.

La definición del *kernel* permite describir la forma del hiperplano, como se puede observar en el ejemplo de la Figura 2.29. De esta forma, se mejora la capacidad de generar límites de decisión no lineales utilizando métodos diseñados para clasificadores lineales; y se hace posible la aplicación de un clasificador a datos que no tienen una representación de espacio vectorial obvia [33].

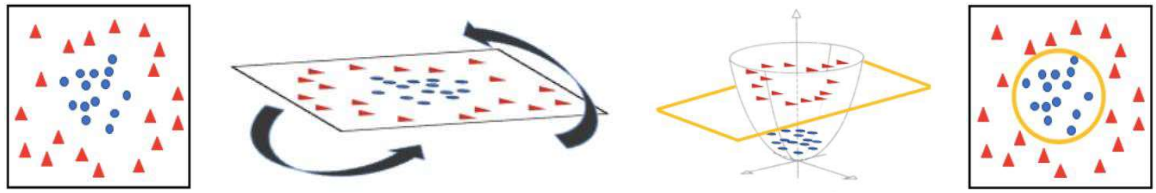


Figura 2.29: Ejemplo de clasificación no lineal utilizando una tranformación del espacio tridimensional. (a) Observaciones originales, (b) rotación del plano, (c) transformación al espacio tridimensional y (d) resultado final de la frontera de separación [34].

En este TFG se estudiarán SVM's con *kernels* lineales, polinómicos y gaussianos, y se obtendrán los resultados de precisión, para el conjunto de datos de entrenamiento, de todos ellos:

- *Linear SVM.* Presenta una interpretación sencilla. Lleva a cabo una separación lineal simple entre clases. Tiene poca flexibilidad.
- *Quadratic SVM.* Hace una separación mediante un polinomio de segundo orden entre clases. Presenta una flexibilidad media.
- *Cubic SVM.* Realiza una separación mediante un polinomio de grado 3 entre clases. Al igual que el SVM con kernel cuadrático, este también presenta una flexibilidad media.
- *Fine Gaussian SVM.* Realiza distinciones entre clases mediante un *kernel* gaussiano con la escala establecida en 1.8. Muy flexible.
- *Medium Gaussian SVM.* Realiza las distinciones entre clases mediante un *kernel* gaussiano con la escala establecida en 7. Flexibilidad media.
- *Coarse Gaussian SVM.* Realiza las entre clases mediante un *kernel* gaussiano con la escala establecida en 28. Flexibilidad baja.

Todos ellos permiten obtener predicciones de forma rápida si solo tienen que separar entre dos clases, reduciéndose la velocidad cuando hay 3 clases o más. De la misma manera, la capacidad de memoria se incrementa cuando se analizan tres o más clases.

K-vecinos más próximos (KNN)

Los modelos basados en memoria son aquellos que almacenan las observaciones para usarlas como prototipos[12]. Uno de los más utilizados es el KNN.

Los KNN son modelos no paramétricos basados en distancias. Estos modelos asumen que el espacio de muestras es métrico: $\{X,\ d\}$, donde X es el conjunto de puntos u observaciones y d es una métrica o distancia [31].

El aprendizaje basado en instancias consta de dos etapas principales [28]:

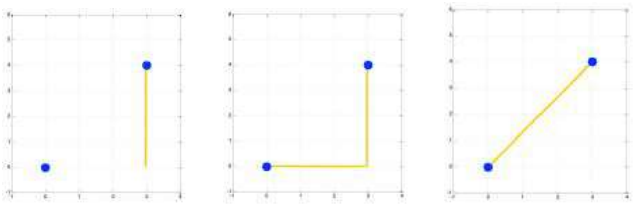
- **Aprendizaje**. Se almacenan todas las muestras de entrenamiento y su clase.
- **Clasificación**. Proceso de búsqueda de instancias más cercanas:
  1. Se representa cada instancia como un conjunto de características.
  2. Se calcula la distancia entre pares de instancias.
  3. Se selecciona la respuesta de la instancia más cercana (k=1) o se emplea el voto de la mayoría (k>1).

Para el paso 2 del método, se calcula la distancia, que se define como d: X x $X \rightarrow \Re$. Teniendo en cuenta que una métrica ha de cumplir las siguientes propiedades para todo $x_i \in X$:

- No negativa: $d(x_1,x_2)>0$.
- Solo $d(x_1,x_2)=0$ si $x_1=x_2$.
- Simétrica: $d(x_1,x_2)=d(x_2,x_1)$
- Desigualdad triangular: $d(x_1,x_2) + d(x_2,x_3) \geq d(x_1,x_3)$

Por otra parte, aunque se utilizan diversas métricas de distancia como la del coseno o la cúbica, las tres más utilizadas son:

- Distancia del ajedrez: $d(x_1,x_2)=\text{máx}_{1\geq i\geq D}\ |(x_{1i}\ -\ x_{2i})|$. Ejemplo en la Figura 2.30.a.
- Distancia de Manhattan: $d(x_1,x_2)=\sum_{i=1}^{D}|(x_{1i}\ -\ x_{2i})|$. Ejemplo en la Figura 2.30.b.
- Distancia euclidea: $d(x_1,x_2)=\sqrt{\sum_{i=1}^{D}(x_{1i}-x_{2i})^2}$. Ejemplo en la Figura 2.30.c.



(a) d(a,b)=4 (b) d(a,b)=7 (c) d(a,b)=5

Figura 2.30: Ejemplo de las tres métricas más usadas: (a) distancia del ajedrez, (b) de Manhattan y (c) euclidea [31].

[12]Los prototipos de una clase son los datos observados de dicha clase.

Por tanto, el funcionamiento del método queda definido por el tipo de métrica (d) empleada y el número de vecinos (k) que se consideren para asignar una determinada observación a una clase.

En resumen, los modelos basados en KNN buscan para cada observación cuál de los datos de alrededor es el más cercano, y se clasifica en la misma clase que ese elemento. Cabe destacar que, para hacerlo más robusto se puede utilizar la media de los datos de alrededor o la media ponderada.

En este TFG se estudiarán diferentes tipos de clasificadores basados en el modelo KNN, y se obtendrán los resultados de precisión de todos ellos para el conjunto de datos de entrenamiento:

- *Fine KNN*. Emplea la distancia euclídea y el número de vecinos se establece en k=1.
- *Medium KNN*. Emplea la distancia euclídea y el número de vecinos se establece en k=10.
- *Coarse KNN*. Emplea la distancia euclídea y el número de vecinos se establece en k=10.
- *Cosine KNN*. Emplea la distancia del coseno y el número de vecinos se establece en k=10.
- *Cubic KNN*. Emplea la distancia cúbica y el número de vecinos se establece en k=10.
- *Weighted KNN*. Variante en la que la clase que se asigna a la nueva muestra depende de pesos asignados a cada voto en función de su cercanía al punto. De esta forma, una observación tendrá una probabilidad más o menos alta de pertenecer a una clase en función de la distancia a los puntos de alrededor. Emplea la distancia euclídea y el número de vecinos se establece en k=10.

En general, en todos estos tipos de técnicas la velocidad de predicción es media, al igual que el uso de memoria, y la interpretación es dificil.

Conforme se aumenta el valor de k, se reduce el error siempre y cuando se disponga de un número suficiente de muestras de entrenamiento etiquedas correctamente. En contraposición, cuantas más muestras se tienen, más tiempo se tardará en realizar la clasificación [28]. Para abordar este compromiso, surgen diferentes técnicas como los métodos de edición, que sirven para solucionar el problema de los patrones mal etiquetados y los métodos de condensado, que reducen el número de muestras sin afectar a la calidad del clasificador, pero consiguiendo reducir el problema de la complejidad computacional.

Técnicas combinadas

Se estudiarán técnicas avanzadas de clasificación que emplean múltiples clasificadores con el fin de obtener un rendimiento más alto.

- ***BAGGING: Bootstrap agregating***. Es una técnica de clasificación en la cual se promedian resultados de conjuntos de clasificadores, con el objetivo de reducir la varianza y, por tanto, minimizar el sobreajuste.

  Esta técnica consiste en dividir el conjunto de entrenamiento en varios subconjuntos con repetición y entrenar varios clasificadores, cada uno de ellos con un subconjunto diferente. Finalmente, se obtendrá la clasificación promediando todos los clasificadores (árboles de decisión, normalmente).

El método más conocido y el que se estudiará en este TFG es el llamado *Random Forest*. Es una técnica de clasificación en la que se usan tanto distintos subconjuntos de datos como un número diferente de características en cada uno de los árboles de decisión que componen el clasificador. La idea principal es hacer uso de *weak classifiers* (clasificadores débiles) para conseguir *strong classifiers* (clasificadores robustos). Esta técnica constituye hoy en día una de las más precisas y eficientes [28].

Una de las técnicas estudiadas en el presente TFG emplea el método combinado de *Random Forest Bag* usando, en concreto, 30 árboles de decisión.

- ***BOOSTING***. Se trata de una técnica de clasificación similar a la anterior, pero que, además de reducir la varianza del error, también reduce el sesgo y, por consiguiente, también minimiza el *overfitting*.

  Esta técnica consiste en entrenar varios clasificadores en serie, es decir, usando el resultado de un clasificador para obtener el siguiente. Los clasificadores más usados son los árboles de decisión.

  El método más conocido que emplea esta técnica y el que se usará en el presente TFG es el *Adaboost*. Es una técnica de clasificación en la cual se entrenan los clasificadores, de manera iterativa, intentando minimizar el error dando más peso a las muestras mal clasificadas en los pasos anteriores. Al igual que el anterior, se basa en el uso de *weak classifiers* para construir *strong classifiers*. La clasificación final es una media ponderada de los clasificadores. Esta técnica es una de las que mejores resultados presenta en términos de precisión y reducción del sobreajuste. Sin embargo, es sensible a los *outliers* y al ruido, por lo que deberán tenerse en cuenta dichas limitaciones en función del problema de clasificación que se trate.

  En este proyecto se estudian varios algoritmos que emplean esta técnica:

  - *AdaBoost*. Combina 30 árboles de decisión; cada uno con un número máximo de 20 *splits*. Presenta una nivel de flexibilidad medio o alto.
  - *RUSBoosted*. Modificación del anterior. Consiste en una alternativa para mejor el rendimiento cuando los datos de entrenamiento están desequilibrados entre clases. Aplica el método RUS, que es una técnica que elimina aleatoriamente ejemplos de la clase mayoritaria [35]. Combina 30 árboles de decisión; cada uno con un número máximo de 20 *splits*.

- **Subespacios aleatorios**. Se especifica el número de predictores que se quiere usar y, a continuación, el algoritmo crea subconjuntos de características independientes empleando uno para cada clasificador, pero teniendo en cuenta todas las observaciones.

  En este proyecto se estudian varios clasificadores que hacen uso de esta técnica:

  - *Subspace Discriminant*. Emplea 30 clasificadores de análisis discriminante y 25 subespacios.
  - *Subspace KNN*. Emplea 30 clasificadores KNN y 25 subespacios.

  Ambos métodos combinados presentan una velocidad de predicción media, son difíciles de interpretar y tienen una flexibilidad media.

2.2.6.2 Classification Learner

Una vez se han extraído las características y se han estudiado y seleccionado las que van a aportar más información y son más discriminantes para el objetivo final del proyecto, se estudiarán los diferentes clasificadores supervisados para seleccionador cuál de ellos tendrá un mejor comportamiento a la hora de asociar una clase a cada observación. Finalmente, se procederá a la etapa de clasificación.

Para esta etapa, se ha decidido emplear la aplicación de MATLAB® llamada *Classification Learner App*. Gracias a esta, se pueden entrenar todos los clasificadores detallados en el apartado anterior para clasificar los datos de entrenamiento llevando a cabo un estudio pormenorizado. Además de analizar diferentes clasificadores de *machine learning*, la *app* también permite examinar los datos, seleccionar características, especificar el tipo de validación, entrenar modelos y evaluar resultados. Incluso, es posible realizar un entrenamiento automatizado para buscar el mejor modelo de clasificación, incluyendo todos los clasificadores explicados en el apartado 2.2.6.1.

En conclusión, mediante esta *app* se realizará un aprendizaje automático supervisado suministrando un conjunto conocido de datos de entrada (observaciones) y unas respuestas conocidas a dichos datos (etiquetas o clases). Se usarán estos datos para entrenar un modelo que generará predicciones para la respuesta de nuevos datos.

En primer lugar, se elegirá el método de validación para examinar la precisión predictiva de los modelos ajustados. La validación estimará el rendimiento del modelo con datos nuevos en comparación con los datos empleados en el entrenamiento y ayudará a seleccionar el mejor modelo de clasificación. Una ventaja de la validación es que ayuda a evitar el sobreajuste del modelo. Los métodos de validación que ofrece el modelo son [36]:

- ***Cross-Validation***. En este método, primero hay que seleccionar el número de *folds* (o divisiones) en las que se desea dividir el conjunto de datos. Siendo k = número de folds, se crean k modelos donde cada uno de ellos es entrenado con k-1 conjuntos, dejando uno de los subconjuntos para la validación. En cada iteración, se escogen diferentes datos para el entrenamiento y para el test, de manera que todas las observaciones son seleccionadas para entrenar y validar, teniendo en cuenta que en cada iteración, un determinado número de datos únicamente puede ser usado para una de las dos opciones. A continuación, se calcula el error medio cometido sobre todos los conjuntos y el que mejor resultados proporcione será el utilizado para predecir las etiquetas de nuevas muestras.
- ***Holdout Validation***. En este método, se selecciona el porcentaje de los datos que se va a usar como conjunto de test y se realiza una única partición que se lleva a cabo de forma aleatoria. La aplicación entrena un modelo con el conjunto de datos que no ha sido seleccionado y evalúa el rendimiento con el conjunto de test seleccionado. Por tanto, para la validación se utiliza solo un parte de los datos, por lo que este método se recomienda solo para grandes conjuntos de datos.
- ***No Validation***. Este método no protege contra el sobreajuste. La aplicación utiliza todos los datos para el entrenamiento y calcula la tasa de error con los mismos. No hay datos de test, por ello, la estimación del rendimiento es poco realista para datos nuevos. Es probable

que la precisión de entrenamiento sea irrealmente alta y la precisión predictiva sea mucho menor.

En el presente TFG se empleará el método del *cross-validation* con k=5 *folds* o divisiones de los datos.

Una vez se ha seleccionado el tipo de validación y se han proporcionado los datos y la respuesta con la clase, se seleccionarán los clasificadores que se quieren evaluar. En este TFG se ha hecho uso de la *Classification Learner App* que permite el estudio de un total de 23 clasificadores, que son los detallados en el subapartado anterior.

Después de entrenar los diferentes modelos de cada familia de clasificadores, aparecerá una lista con la *accuracy* en porcentaje asociada a cada modelo de entrenamiento. La mejor *accuracy* se resalta con un recuadro haciendo referencia al clasificador que ha prorporcionado los valores de precisión más altos durante la etapa de validación. Este valor estima el desempeño de un modelo con datos nuevos en comparación con los datos de entrenamiento.

La mejor puntuación no tiene porque ser el mejor modelo para un objetivo dado, por ello, además de la *accuracy*, la aplicación también proporciona más medidas de calidad como la matriz de confusión o la curva ROC que serán explicadas más adelante. También, presenta un diagrama de dispersión en el cual se pueden analizar visualmente las muestras que han sido mal etiquetadas con cada modelo.

Una vez se ha escogido el modelo que se quiere usar para el proyecto, se guardará y se generará el código, el cual corresponde con la función *trainClassifier* del diagrama de la Figura 2.3. Dicho modelo contendrá la información necesaria para predecir futuras muestras, nunca antes vistas.

Para realizar la predicción se empleará el modelo compacto del clasificador, pero será necesario hacer una normalización previa de los datos de *test* mediante los valores *mu* y *sigma* obtenidos en la selección de características del apartado 2.2.5, así como eliminar las características que se descartaron en dicho apartado. El motivo de esto es que para proceder a la clasificación, los datos de *test* tendrán que ser analizados con las mismas características, estar ordenados de la misma manera y mantener la misma estructura que los datos de entrenamiento empleados. La diferencia radica en que, en esta ocasión, no se le pasará la información de la clase (vector *groundtruth*) al clasificador.

2.2.6.3 Indicadores de resultados

El clasificador seleccionado será evaluado mediante una serie de medidas de calidad que se recopilarán y se compararán más adelante con los resultados de predicción para comprobar que funciona como se espera.

Por tanto, se obtendrán los mismos indicadores tanto para la evaluación del clasificador durante la etapa de entrenamiento como para la evaluación de los resultados durante la fase de predicción. Los indicadores implementados en este TFG son los siguientes:

- **Matriz de confusión**. Se trata de una matriz cuadrada donde el número de filas y columnas equivale al número de clases. Esta matriz muestra la relación entre las clases de referencia y la categorización de las observaciones por el clasificador. Además, a partir de la matriz de confusión pueden obtenerse otros índices relativos a la sensibilidad, la especificidad, el

valor predictivo positivo (PPV), el valor predictivo negativo (NPV) y el porcentaje de éxito, entre otros. Los valores contenidos en la matriz hacen referencia a los datos clasificándolos como:

- Verdaderos negativos (VN). Número de muestras que el clasificador etiqueta como sanas y realmente lo son.
- Verdaderos positivo (VP). Número de muestras que son etiquetadas correctamente como patológicas de grado 3.
- Falsos negativos (FN). Número de muestras que son clasificadas erróneamente como sanas.
- Falsos positivos (FN). Número de muestras que se clasifican como patológicas cuando realmente eran sanas.

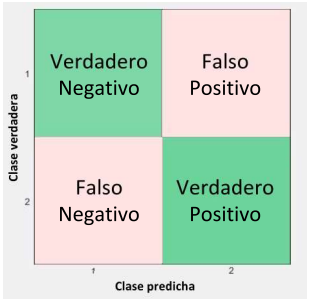


Figura 2.31: Estructura de una matriz de confusión donde 1 se refiere a la clase negativa y 2, la positiva.

- **Sensibilidad**. Indica la capacidad del clasificador para identificar las etiquetas positivas. Es decir, mide la capacidad para etiquetar como patológica una muestra cuando realmente lo es.

$$Sensibilidad = \frac{VP}{VP + FN} = FVP \tag{2.29}$$

donde FVP es la fracción de verdaderos positivos.

- **Especificidad**. Indica la capacidad del clasificador para identificar las etiquetas negativas. O sea, mide la bondad del clasificador a la hora de etiquetar como sana una muestra sana.

$$Especificidad = \frac{VN}{VN + FP} = FVN \tag{2.30}$$

donde FVN es la fracción de verdaderos negativos.

- **Ratio de acierto**. Indica el número o porcentaje de muestras del total que están clasificadas correctamente. También se le conoce como *accuracy*.

$$RatioAcierto = \frac{VN + VP}{T} \tag{2.31}$$

donde T es el total de muestras.

- **Curva ROC**. Se trata de una gráfica en la que se representa la sensibilidad en función de la fracción de falsos positivos (FFP) o 1-especificidad. Ambos ejes incluyen valores entre 0 y 1. El punto óptimo se encuentra en la esquina superior izquierda, de forma que se intenta maximizar conjuntamente tanto la sensibilidad como la especificidad.

  Se trata de un indicador excelente para comparar distintos clasificadores. Idealmente, la curva describiría un ángulo de 90º donde la sensibilidad sería 1 y la FFP sería 0. Además, el valor óptimo ideal sería (0,1).

  Cabe destacar que, lo realmente interesante de esta gráfica es el parámetro AUC, detallado a continuación.

- **AUC** (del inglés, *Area Under Curve*). Se trata del área bajo la curva ROC. Es un parámetro que evalúa la bondad de una prueba diagnóstica de forma continua tal que, cuanto mayor sea el área bajo la curva ROC, mayor será la capacidad discriminatoria del clasificador. Idealmente, el área bajo la curva ROC sería 1.

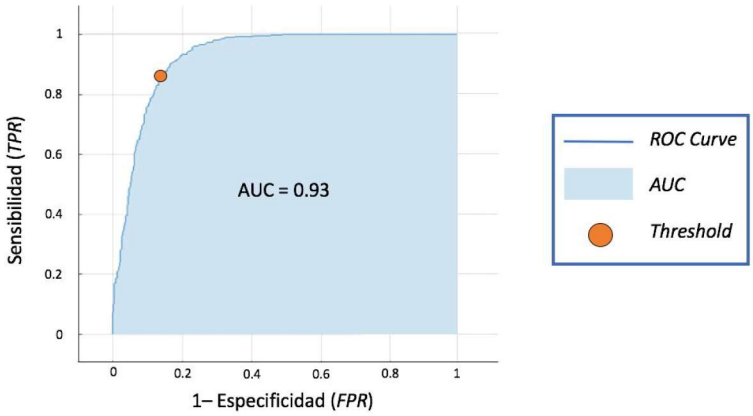


Figura 2.32: Ejemplo de Curva ROC y área bajo la curva (AUC) [34].

Capítulo 3

Resultados y discusión

3.1 Mejor clasificador

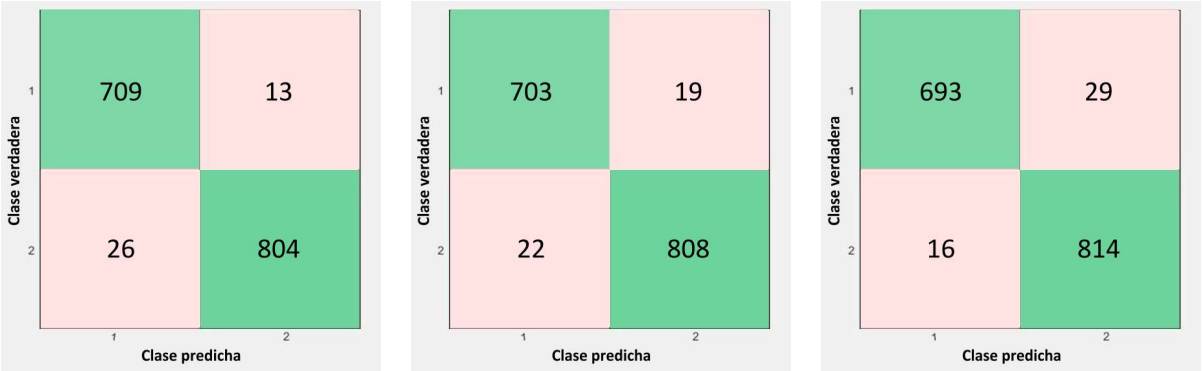
El proceso de selección del clasificador empezará con el entrenamiento con los datos de *train* de varios clasificadores, concretamente, los 23 explicados en la Sección 2.2.6.1. De todos ellos, se calculará la *accuracy* durante la etapa de validación y se seleccionará el que presente mejores resultados para abordar la fase de predicción.

Cabe destacar que la selección del mejor clasificador se realizará con la aplicación *Classification Learner*, como se explica en la subsección 2.2.6.2. Se obtendrán los resultados expuestos en la Tabla 3.1, donde se pueden ver remarcados los valores más altos, correspondientes a los clasificadores que se detallarán con mayor profundidad.

Simple Tree	90.7 %	*Medium Tree*	93.6 %
Complex Tree	93.6 %	*Linear Discriminant*	83.8 %
Quadratic Discriminant	86.9 %	*Logistic Regression*	94.7 %
Linear SVM	94.5 %	*Quadratic SVM*	**97.4 %**
Cubic SVM	**97.5 %**	*Fine Gaussian SVM*	73.8 %
Medium Gaussian SVM	**97.1 %**	*Coarse Gaussian SVM*	93.6 %
Fine KNN	92.3 %	*Medium KNN*	93.0 %
Coarse KNN	88.7 %	*Cosine KNN*	92.7 %
Cubic KNN	91.5 %	*Weighted KNN*	93.1 %
Boosted Trees: AdaBoost	95.7 %	*RusBoosted Trees*	93.9 %
Bagged Trees	96.1 %	*Subspace Discriminant*	93.2 %
Subspace KNN	94.4 %		

Tabla 3.1: Valores de precisión obtenidos para los diferentes modelos de clasificción entrenados.

Como se puede observar en la Tabla 3.1, la familia de clasificadores basados en los modelos de *Support Vector Machine* destaca considerablemente por encima del resto, siendo los tres mejores: el SVM cuadrático, el cúbico y el gaussiano medio. A continuación, se observan las matrices de confusión de dichos clasificadores con las se corroborarán los resultados obtenidos.



(a) (b) (c)

Figura 3.1: Matrices de confusión de: (a) SVM cúbico, (b) SVM cuadrático y (c) SVM gaussiano medio.

A partir de las matrices de confusión anteriores, se calculará el número de muestras correctamente clasificadas como VP + VN, obteniéndose los resultados de la Tabla 3.2.

SVM	cúbico	cuadrático	gaussiano medio
nº muestras correctamente clasificadas	1513	1511	1507

Tabla 3.2: Número de muestras correctamente clasificadas para cada clasificador.

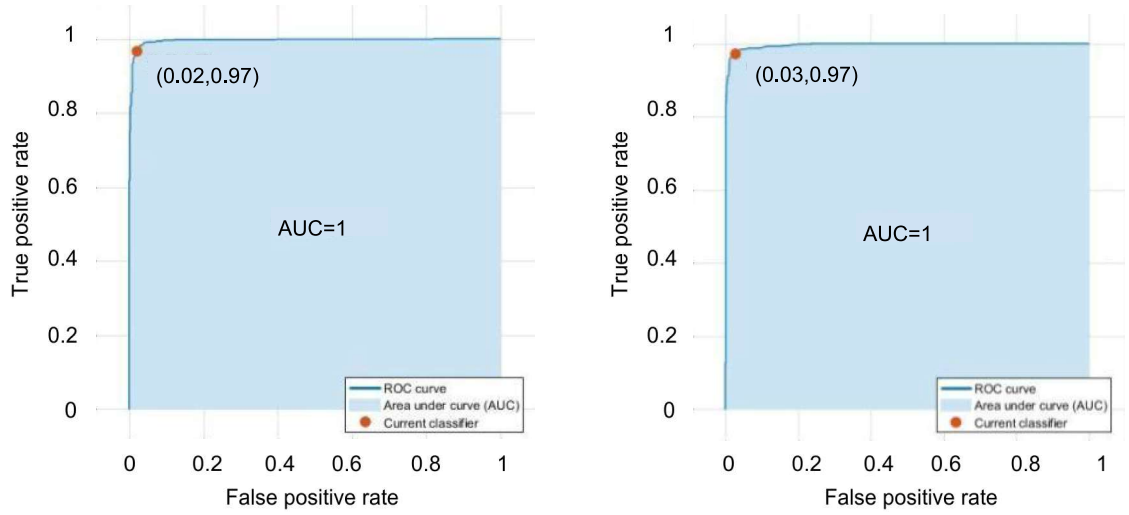
Una vez obtenidos estos resultados, se afirmará con seguridad que el clasificador *Medium Gaussian SVM* no es el más adecuado. El motivo de esto es que no solo posee el menor número de muestras correctamente clasificadas, si no que, además, es el que mayor diferencia tiene entre VP y VN; será interesante que el clasificador esté equilibrado en la medida de lo posible.

A continuación, se comprobarán los resultados de otra de las medidas de calidad de la clasificación que ofrece la aplicación, la curva ROC. Se pueden observar las curvas ROC de los clasificadores *Quadratic SVM* y *Cubic SVM* en la Figura 3.2.

Tras observar los resultados de las curvas ROC de ambos clasificadores, se comprobará que el mejor clasificador es el ***Cubic SVM***. Debido, principalmente, a los siguientes motivos:

- Posee una curva ROC que se asemeja en mayor medida a la curva ROC ideal.
- Posee un valor óptimo más cercano al ideal.

Asimismo, a pesar de poseer el mismo valor AUC, no se tendrá en cuenta ya que probablemente sea una aproximación y los valores reales sean del orden de 0'99.



(a) (b)

Figura 3.2: Curva ROC y valor AUC de: (a) SVM cúbico y (b) SVM cuadrático.

3.2 Validación

Tras la selección del mejor clasificador, se implementará la función *Validation* y se obtendrán una serie de indicadores de calidad para analizar los valores de otras métricas de interés que proporciona el clasificador seleccionado.

Indicador	**Resultados**
Sensibilidad	0.9848
Especificidad	0.9699
Ratio de acierto	0.9768
AUC	0.9963

Tabla 3.3: Indicadores de calidad de la validación con el clasificador SVM cúbico.

Los valores obtenidos son considerablemente altos, lo cual indica que el SVM cúbico es un prometedor clasificador para obtener interesantes resultados durante la etapa de predicción. El ratio de acierto y la *accuracy* son lo mismo a nivel matemático y se ha obtenido en la validación un 97.68 %.

3.3 Predicción

Una vez validado nuestro clasificador a partir de los datos del conjunto de entrenamiento, se procederá a comprobar los resultados que se obtendrían a la hora de analizar y predecir la clase de muestras totalmente nuevas. De esta forma, se podrá predecir el comportamiento del clasificador en un futuro cuando se analicen muestras de nuevos pacientes. Para ello, se obtendrán los mismos indicadores que en la validación:

Indicador	Resultados
Sensibilidad	0.9791
Especificidad	0.9694
Ratio de acierto	0.9742
AUC	0.9963

Tabla 3.4: Valores obtenidos, para diferentes indicadores, durante la etapa de predicción con el clasificador SVM cúbico entrenado.

Como se puede observar, al comparar la Tabla 3.3 y la 3.4, los valores son bastante similares, lo cual indica que el modelo de clasificación entrenado es muy robusto, pues los resultados no varían en gran medida cuando se analizan muestras nunca vistas anteriormente. Aún así, este hecho era algo esperado ya que la validación se realiza con los datos de entrenamiento y siempre existirá algo de ajuste de los datos empleados al clasificador entrenado.

Asimismo, se obtendrá la matriz de confusión de los datos de *test* para comprobar otros resultados, esta se muestra en la Figura 3.3.

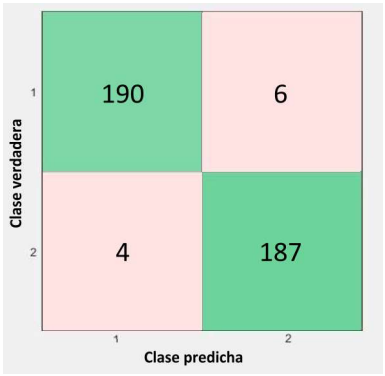


Figura 3.3: Matriz de confusión de los datos de *test* con el clasificador SVM cúbico entrenado.

Los resultados obtenidos incitan a pensar que el modelo de clasificación construido es bastante robusto, ya que los valores proporcionados para las diferentes métricas son considerablemente elevados, en comparación con los propuestos en los diferentes estudios del estado del arte. Aún así, el modelo es susceptible de mejora y, por ello, en futuros estudios se intentará profundizar en esta prometedora línea de investigación.

Cabe señalar también que este proyecto se inicia a partir de una segmentación de las unidades glandulares llevada a cabo de forma manual, por lo que los resultados serán, a priori, mejores que los enfoques end-to-end, ya que este no cuenta con el error añadido de la etapa de segmentación.

El resultado obtenido es significativamente alto pero se procederá a compararlo con los resultados de la literatura para tener una visión más objetiva de la calidad obtenida (véase en la Tabla 3.5).

Como vemos el presente clasificador proporciona muy buenos resultados, siendo solo superado por el primer autor y con una diferencia bastante pequeña. Si se observan las características extraídas por el primer autor en la Tabla 1.3, se podrá comprobar que los descriptores estudiados en dicho artículo son principalmente de forma. Ello hace pensar que, probablemente, la alta calidad del clasificador sea debido, principalmente, a los descriptores de forma.

Autores	**Resultados**
Nguyen, Sabata y Jain [13]	98.3 %
Nguyen, Sarkar y Jain [15]	79.0 %
Jian Ren [12]	83.0 %
Kwak y Hewitt [16]	95.0 %
TFG	97.4 %

Tabla 3.5: Comparación de la *accuracy* obtenida en los diferentes estudios con respecto a la obtenida en el presente TFG.

Es necesario hacer hincapié en que los resultados obtenidos del estado del arte no son objetivamente comparables con los resultados del presente TFG, pues existen numerosas diferencias con respecto al tipo de clasificador empleado y las clases que se pretenden separar. Además, se han empleado bases de datos diferentes debido a que no existen públicas y esto, también afectará a los resultado.

Capítulo 4

Conclusiones y líneas futuras

4.1 Conclusiones

Se ha propuesto un novedoso método de detección de cáncer de grado 3 mediante la extracción avanzada de características de imágenes histológicas de próstata, empleando herramientas de procesamiento de imagen digital. El método compagina, de forma novedosa, diferentes descriptores para lograr unos resultados más precisos a la hora de discriminar entre glándulas benignas y patológicas de grado 3 en tejido prostático. Para ello, se han combinado ideas sugeridas tanto en la literatura científica del estado del arte, como ideas propias tras realizar una exhaustiva revisión bibliográfica.

En primer lugar, se ha conseguido crear una base de datos de glándulas histológicas (de carácter benigno y patológico de grado 3) y se ha evidenciado que el análisis individual de las estructuras glandulares es una prometedora línea de investigación de cara a la distinción del cáncer de grado 3 en imágenes histológicas de próstata. Además, se han implementado técnicas de *clustering* para la importante tarea de separar con precisión los principales elementos que constituyen una unidad glandular (núcleos, estroma, citoplasma y lumen). Por otra parte, se ha reivindicado la utilidad de adquirir distintos canales de color para resaltar determinados componentes.

En segundo lugar, se ha conseguido hacer uso de un total de 49 características diferentes, tras exhaustivos procesos de extracción y selección de características, que han permitido obtener un valor de 97.4 % de precisión durante la etapa de predicción. Cabe destacar que los resultados propuestos presentan una gran fiabilidad, pues se han implementado técnicas de partición de datos como el *cross-validation* con 5 *folds* para la etapa de validación y como el *holdout method* (80 %-20 %) para la fase de predicción, con el objetivo de dotar de gran robustez al modelo entrenado. Asimismo, se ha demostrado que la combinación de descriptores de forma, textura y color resulta de gran interés.

En tercer lugar, se ha llevado a cabo una comparación con otros estudios publicados en el estado del arte para ensalzar la bondad del clasificador, así como el interés de los métodos y las estrategias propuestas en este proyecto.

En resumen, se ha conseguido diseñar y desarrollar un método de extracción de características de glándulas prostáticas para el diagnóstico de cáncer de grado 3, cuyos resultados han sido evaluados aportando valores sobresalientes que podrán ser empleados en un futuro.

4.2 Líneas futuras

El proyecto se ha realizado con éxito, pero existen posibles mejoras que supondrían un incremento en la calidad del proyecto. A continuación, se exponen algunas de las limitaciones encontradas durante la realización del trabajo, así como sus posibles soluciones. Además, también se exponen diversas ideas de interés de cara a la continuidad de esta línea de investigación.

Por una parte, resultaría interesante emplear un algoritmo más elaborado para la etapa del *clustering*, ya que es una de las fases más importantes de las que depende en gran medida el resto del proyecto. Además, a parte de los cuatro elementos ya obtenidos (núcleos, estroma, citoplasma y lúmenes), se propone contemplar más grupos a separar, como por ejemplo la mucina, que está cobrando importancia en artículos recientes de la literatura médica. De esta forma, se podrían emplear conjuntamente características de dicho componente junto con el resto con la finalidad de obtener mejores resultados en la precisión del clasificador. Igualmente, se plantea tener en cuenta parámetros de la glándula para hacer el algoritmo más personalizado y exacto.

Por otra parte, en este proyecto se ha requerido de una segmentación previa de las glándulas como punto de partida. Esto ha conllevado un trabajo manual que debe ser automatizado con la mayor precisión posible de cara a futuros estudios, ya que el objetivo final reside en proporcionar un sistema de ayuda diagnóstica a los patólogos para que no necesiten implementar tareas manuales y tediosas.

Además, en este proyecto sólo se han empleado clasificadores tradicionales de *machine learning*, se propone utilizar otros clasificadores como redes neuronales basadas en el perceptrón multicapa. También, se propone abordar el problema con diferentes arquitecturas de *deep learning*. Destacar que para ello sería necesario un número de muestras muy superior del que se dispone hoy en día. No obstante, existe una colaboración entre los patólogos del Hospital Clínico de Valencia y el grupo CVBLab de la UPV y se esperan nuevas muestras para trabajar con ellas. Por lo que, estas líneas de futuro serán plausibles próximamente.

Finalmente, en este TFG solo se ha contemplado la distinción de glándulas sanas y patológicas de grado 3 debido a que este último es el primer grado considerado como cáncer y el objetivo del proyecto ha sido conseguir un diagnostico más precoz para aportar un tratamiento temprano y personalizado. Por tanto, se propone como posible línea de futuro ampliar el área de conocimiento y desarrollar un método que detecte todos los grados del *Gleason score*.

Parte II

Presupuesto

Capítulo 5

Presupuesto

En el presente capítulo se presentará una valoración económica del proyecto realizado. Se detallarán las diferentes partes que conforman el presupuesto del presente TFG. Además, se comentarán algunas consideraciones que se han tenido en cuenta.

5.1 Presupuestos parciales

Este apartado del informe se desglosa en tres cuadros de precios: mano de obra, maquinaria y materiales.

5.1.1 Coste mano de obra

A continuación, se describirán los recursos humanos que han sido necesarios para el desarrollo del presente proyecto. Se realizará una estimación de los costes en función del tiempo dedicado al trabajo. Se considerará la contribución de: D^{a}. Valery Naranjo Ornedo (como directora del proyecto), D. José Gabriel García Pardo (como cotutor del proyecto) y D^{a}. Elena Payá Bosch (como alumna y autora del proyecto).

Denominación	**Uds.**	**Cantidad**	**Precio Unitario (€)**	**Total (€)**
Tutora (Catedrática)	h	15	42,00	630,00
Cotutor (Doctorando)	h	45	17,20	774,00
Autora (Estudiante GIB)	h	400	12,50	5.000,00
			Total	**6.404,00**

Tabla 5.1: Cuadro de precios de la mano de obra.

5.1.2 Coste maquinaria

En esta subsección se detallará el cuadro de precios de los recursos correspondientes al *hardware* y al *software* que han sido necesarios para el desarrollo del presente proyecto.

Cabe destacar que para la realización del presente TFG, las herramientas *software* empleadas han sido MATLAB®, concretamente, *Image Processing Toolbox* y *Statistics and Machine Learning Toolbox* y, para el desarrollo del trabajo escrito, se ha empleado la aplicación conocida como *Overleaf* (una herramienta de escritura, edición y publicación en línea de LaTex).

Asimismo, se tendrá en cuenta que las herramientas *hardware* no se han empleado específicamente para la elaboración del TFG. Por ello, se calculará el periodo de amortización de las mismas.

Denominación	**Uds.**	**Cantidad**	**Precio Unitario (€)**	**Periodo de amortización (años)**	**Intervalo amortizado (meses)**	**Total (€)**
Overleaf	u	1	0,00	1	6	0,00
Licencia MATLAB®	u	1	800,00	1	6	800,00
Image Processing Toolbox	u	1	400,00	1	6	400,00
Statistics and Machine Learning Toolbox	u	1	400,00	1	6	400,00
HP Pavilion TS 14 Notebook PC	u	1	800,00	4	6	100,00
					Total	**1.700,00**

Tabla 5.2: Cuadro de precios de las herramientas *hardware* y *software*.

5.1.3 Coste materiales

Finalmente, para llevar a cabo el proyecto se ha requerido la realización de pruebas de biopsia, a partir de las cuales, se han obtenido las muestras correspondientes a las imágenes utilizadas durante el proyecto. Además, la adquisición de cada muestra tiene un coste asociado al *slide*.

Denominación	**Uds.**	**Cantidad**	**Precio Unitario (€)**	**Total (€)**
Biopsia	u	25	600,00	15.000,00
Muestras	u	35	10,00	350,00
			Total	**15.350,00**

Tabla 5.3: Cuadro de precios de los materiales.

5.2 Presupuesto total

Para el cálculo del presupuesto total del proyecto, será necesario tener en cuenta los cuadros de precios de los presupuestos parciales definidos anteriormente. Además, se añadirán el porcentaje de gastos generales (13 %) y el asociado al beneficio industrial (6 %). A continuación, se añadirá al precio total bruto el impuesto del IVA (21 %), obteniendo como resultado el presupuesto total que supondría la realización del presente TFG.

CAPÍTULOS	**IMPORTE (€)**
1. Coste de la mano de obra	6.404,00
2. Coste de las herramientas	1.700,00
3. Coste de los materiales	15.350,00
PRESUPUESTO DE EJECUCIÓN DE MATERIAL	**23.454,00**
13 % de gastos generales	3.049,02
6 % de beneficio industrial	1.407,24
PRESUPUESTO DE EJECUCIÓN POR CONTRATA	**27.910,26**
21 % de IVA	5.861,15
PRESUPUESTO TOTAL	**33.771,41**

Tabla 5.4: Presupuesto total.

Bibliografía

[1] Organización Mundial de la Salud, *Cáncer*, http://www.who.int/es/news-room/fact-sheets/detail/cancer, 2018 (vid. págs. 3, 4).

[2] SEOM, "Las Cifras del Cáncer en España", Sociedad Española de Oncología Médica (SEOM), inf. téc., 2018 (vid. págs. 3-5).

[3] G.R. F. Collaborators, *Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015.* Lancet., 2016, vol. 388 (vid. pág. 4).

[4] Asociación Española Contra el Cáncer, *¿QUÉ ES EL CÁNCER DE PRÓSTATA?*, https://www.aecc.es/es/todo-sobre-cancer/tipos-cancer/cancer-prostata/que-es-cancer-prostatar, 2018 (vid. pág. 5).

[5] American Society of Clinical Oncology, *Cáncer de próstata*, https://www.cancer.net/es/tipos-de-c%C3%A1ncer/c%C3%A1ncer-de-pr%C3%B3stata, 2017 (vid. págs. 5, 6).

[6] Instituto Nacional del Cáncer, ed., *Diccionario de cáncer*, 2017 (vid. pág. 6).

[7] D. J. P. Burgués., "¿En qué consiste una biopsia de próstata? ¿Cuándo hay que hacerla?", 2016 (vid. pág. 6).

[8] MRI Imaging Center of Fresno, Inc., *Que es MRI*, http://mrifresno.com/que-es-mri/, 2015 (vid. pág. 7).

[9] Wikipedia, *Técnica histológica — Wikipedia, La enciclopedia libre*, https://es.wikipedia.org/w/index.php?title=T%C3%A9cnica_histol%C3%B3gica&oldid=104583300, [Internet; descargado 24-junio-2018], 2017 (vid. pág. 7).

[10] Antonio Félix Conde Martí, coordinador del club, *Libro Blanco de la Anatomía Patológica en España*. 2015, cap. Recomendaciones del Club de Patología Digital de la SEAP (vid. pág. 8).

[11] PCEC, "Gleason Score, Prostate Cancer Grading & Prognostic Scoring", Prostate Conditions Education Coincil, inf. téc. (vid. pág. 9).

[12] D. J.F.X. Q. Jian Ren Evita Sadimin, *Computer aided analysis of prostate histopathology images to support a refined Gleason grading system*, 2017. DOI: 10.1117/12.2253887 (vid. págs. 10, 11, 61).

[13] K. Nguyen, B. Sabata y A. K. Jain, "Prostate cancer grading: Gland segmentation and structural features", *Pattern Recognition Letters*, vol. 33, n.º 7, págs. 951 -961, 2012, Special Issue on Awards from ICPR 2010, ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2011.10.001 (vid. págs. 10, 11, 61).

[14] Wikipedia, *Mucina — Wikipedia, La enciclopedia libre*, https://es.wikipedia.org/w/index.php?title=Mucina&oldid=101558629, [Internet; descargado 24-junio-2018], 2017 (vid. pág. 10).

[15] K. Nguyen, A. Sarkar y A. K. Jain, "Structure and Context in Prostatic Gland Segmentation and Classification", en *Proceedings of the 15th International Conference on Medical Image Computing and Computer-Assisted Intervention - Volume Part I*, ép. MICCAI'12, Nice, France: Springer-Verlag, 2012, págs. 115-123, ISBN: 978-3-642-33414-6. DOI: 10.1007/978-3-642-33415-3_15 (vid. págs. 11, 61).

[16] J. T. Kwak y S. M. Hewitt, "Multiview Boosting Digital Pathology Analysis of Prostate Cancer", *Comput. Methods Prog. Biomed.*, vol. 142, n.º C, págs. 91-99, abr. de 2017, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2017.02.023 (vid. págs. 11, 61).

[17] *MicroDraw*, http://microdraw.pasteur.fr/ (vid. pág. 15).

[18] The MathWorks, Inc., *MATLAB*, https://es.mathworks.com/products/matlab.html (vid. pág. 17).

[19] The MathWorks, Inc., *Kmeans clustering*, https://es.mathworks.com/help/stats/kmeans.html (vid. pág. 22).

[20] Wikipedia, *Tinción hematoxilina-eosina — Wikipedia, La enciclopedia libre*, https://es.wikipedia.org/w/index.php?title=Tinci%C3%B3n_hematoxilina-eosina&oldid=107911117, [Internet; descargado 28-junio-2018], 2018 (vid. pág. 23).

[21] A. C. Ruifrok y D. A. Johnston, "Quantification of histochemical staining by color deconvolution", *Analytical and quantitative cytology and histology*, vol. 23, n.º 4, 2001 (vid. pág. 23).

[22] J. M. P. J. Ríos Díaz y M. B. Aledo, "El análisis textural mediante las matrices de co-ocurrencia (GLCM) sobre imagen ecográfica del tendón rotuliano es de utilidad para la detección cambios histológicos tras un entrenamiento con plataforma de vibración.", *Cultura, Ciencia y Deporte*, págs. 91-102, 2009 (vid. págs. 25, 29).

[23] Óscar Boullosa García, "Estudio comparativo de descriptores visuales para la detección de escenas cuasi-duplicadas", Escuela Politécnica Superior Universidad Autónoma de Madrid, inf. téc., 2011 (vid. págs. 25, 36).

[24] I. Claudia Ximena MAzo, *Descriptores de Textura*, =https://es.slideshare.net/mmv-lab-univalle/descriptores-de-textura (vid. pág. 28).

[25] The MathWorks, Inc, *graycomatrix*, `https://es.mathworks.com/help/images/ref/graycomatrix.html` (vid. pág. 30).

[26] E. A. Oscar García-Olalla, *Conceptos y métodos en Visión por Computador, Cápitulo 7. DESCRIPCIÓN DE TEXTURA EN IMÁGENES UTILIZANDO LOCAL BINARY PATTERN (LBP)* (vid. pág. 31).

[27] T. Ojala, M. Pietikäinen y T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, págs. 971-987, 2002 (vid. págs. 31-34).

[28] V. Naranjo, *Clasificación de patrones* (vid. págs. 33, 44, 46, 49-51).

[29] ATGC Grupo de Tecnologías Avanzadas en Computación – Loxa Ecuador, *Clasificación supervisada y no supervisada*, `https://advancedtech.wordpress.com/2008/04/14/clasificacion-supervisada-y-no-supervisada/`, 2018 (vid. pág. 44).

[30] C. H. L., *Arboles de Decisión (I)* (vid. pág. 45).

[31] E. F. i Garcia, *Fundamentos de los Sistemas de Ayuda a la Decisión Basados en Datos Biomédicos: Métodos de Aprendizaje*, 2017-2018 (vid. págs. 46, 49).

[32] C. A. González, *SVM: Máquinas de Vectores Soporte* (vid. pág. 47).

[33] O. C.S.S.S.S.B.R. G. Ben-Hur A., "Support Vector Machines and Kernels for Computational Biology", *PLoS Computational Biology*, vol. 10, n.º 4, 2008, ISSN: e1000173. DOI: `http://doi.org/10.1371/journal.pcbi.1000173` (vid. pág. 48).

[34] J. G. G. Pardo, *Diseño y desarrollo de un sistema automático de clasificación de estructuras glandulares en imágenes histológicas de próstata*, 2017-18 (vid. págs. 48, 55).

[35] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse y A. Napolitano, "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance", *IEEE Transactions on Systems, Man, and Cyber-*

netics - Part A: Systems and Humans, vol. 40, n.º 1, págs. 185-197, 2010, ISSN: 1083-4427. DOI: 10.1109/TSMCA.2009.2029559 (vid. pág. 51).

[36] The MathWorks, Inc., *Select Data and Validation for Classification Problem*, https://es.mathworks.com/help/stats/select-data-and-validation-for-classification-problem.html (vid. pág. 52).