UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

DESARROLLO DE UN SISTEMA DE EXTRACCIÓN LOCAL DE CARACTERÍSTICAS EN IMAGEN HISTOLÓGICA PARA LA IDENTIFICACIÓN AUTOMÁTICA DE CÁNCER DE PRÓSTATA

AUTORA: María Jesús García González

TUTORA: Valeriana Naranjo Ornedo

COTUTOR: Adrián Colomer Granero

Curso Académico: 2017-18

Resumen

El presente trabajo tiene como objetivo el desarrollo de un algoritmo automático para la clasificación de imágenes histológicas de próstata que sea capaz de distinguir entre tejido sano y tejido patológico de grado 3 en la escala de Gleason. Dicho TFG se enmarca en un proyecto de ámbito nacional denominado SICAP, cuyo propósito es proporcionar una herramienta de ayuda al profesional sanitario para el diagnóstico precoz del cáncer de próstata.

Para conseguir el objetivo del trabajo se propone una metodología que comienza por el acondicionamiento y preprocesado de las sesenta imágenes que conforman la base de datos proporcionada por el Hospital Clínico de Valencia. Tras este primer paso, se lleva a cabo una etapa de extracción de características, empleando descriptores de imagen que sean capaces de extraer la información relevante de la misma, sin necesidad de segmentar las distintas partes del tejido. Concretamente, los clasificadores empleados son filtros de Gabor, matrices de coocurrencia y granulometría. Tras la obtención de características, se realiza un análisis estadístico que acabe por descartar aquellas que no aportan información útil para la clasificación. En última instancia, dicha información se emplea para entrenar modelos predictivos empleando diversos algoritmos de aprendizaje automático. Los modelos resultantes son evaluados y comparados entre sí con el objetivo de establecer cuál de ellos presenta mayor precisión y robustez en la clasificación del problema bajo estudio.

Los resultados de la metodología propuesta se detallan y se discuten realizando además una comparación de estos con los obtenidos mediante otras metodologías presentes en el estado del arte. Tras analizar las fortalezas y debilidades del algoritmo propuesto, se exponen algunas mejoras y líneas futuras de investigación.

Palabras Clave: cáncer de próstata, imagen histológica, clasificación Gleason, filtros de Gabor, granulometría, matrices de coocurrencia, clasificación supervisada.

Resum

El present treball té com a objectiu el desenvolupament d'un algoritme automàtic per a la classificació d'imatges histològiques de pròstata que siga capaç de diferenciar entre teixit sa i teixit patològic de grau 3 en l'escala de Gleason. Aquest TFG s'emmarca en un projecte d'àmbit nacional denominat SICAP, el propòsit del qual és proporcionar una ferramenta d'ajuda al professional sanitari per al diagnòstic precoç del càncer de pròstata.

Per a aconseguir l'objectiu del treball es proposa una metodologia que comença pel acondicionament i preprocessat de les seixanta imatges que conformen la base de dades proporcionada per l'Hospital Clínic de València. Després d'este primer pas, es du a terme una etapa d'extracció de característiques, emprant descriptors d'imatge que siguen capaços d'extraure la informació rellevant de la mateixa, sense necessitat de segmentar les distintes parts del teixit. Concretament, els classificadors empleats són filtres de Gabor, matrius de coocurrencia i granulometria. Després de l'obtenció de característiques, es realitza una anàlisi estadística que acabe per descartar aquelles que no aporten informació útil per a la classificació. En última instància, la informació s'empra per a entrenar models predictius emprant diversos algoritmes d'aprenentatge automàtic. Els models resultants són avaluats i comparats entre si amb l'objectiu d'establir quin d'ells presenta més precisió i robustesa en la classificació del problema baix estudi.

Els resultats de la metodologia proposada es detallen i es discutixen realitzant a més una comparació d'estos amb els obtinguts per mitjà d'altres metodologies presents en l'estat de l'art. Després d'analitzar les fortaleses i debilitats de l'algoritme proposat, s'exposen algunes millores i línies futures d'investigació.

Paraules clau: càncer de próstata, imatge histològica, classificació Gleason, filtres de Gabor, granulometria, matrius de coocurrència, classificació supervisada.

Abstract

The aim of this TFG is to develop an automatic algorithm for the classification of prostate histological images that is able to distinguish between healthy tissue and pathological tissue of grade 3 on the Gleason scale. This TFG is part of a national project called SICAP, whose purpose is to provide a tool to help the health professionals for the early diagnosis of prostate cancer.

To achieve the objective of the work we propose a methodology that begins with the conditioning and preprocessing of the sixty images that makes up the database provided by the Hospital Clínico de Valencia. After this first step, a characteristic extraction stage is carried out, using image descriptors that are capable of extracting the relevant information from it, without the need to segment the different parts of the tissue. Specifically, the classifiers used are Gabor filters, co-occurrence matrices and granulometry. After obtaining characteristics, a statistical analysis that ends up discarding those that do not provide useful information for the classification is carried out. Ultimately, this information is used to train predictive models using various machine learning algorithms. The resulting models are evaluated and compared with each other in order to establish which of them presents greater precision and robustness in the classification of the problem under study.

The results of the proposed methodology are detailed and discussed, making a comparison of these with those obtained by other methodologies present in the state of the art. After analyzing the strengths and weaknesses of the proposed algorithm, some improvements and future lines of research are exposed.

Keywords: prostate cancer, histological image, Gleason score, Gabor filters, granulometry, co-occurrence matrices, supervised classification.

Índice general

Parte I

Memoria

Capítulo 1

Introducción

En las últimas dos décadas, el número de tumores diagnosticados en España ha sufrido un crecimiento constante. Este crecimiento se asocia no solo al aumento poblacional, si no también a las técnicas de detección precoz y al aumento de la esperanza de vida de la población.

Según el último informe emitido por la Sociedad Española de Oncología Médica [1], los tumores con más incidencia en España en el año 2017 fueron los de colorrecto, próstata, pulmón, mama, vejiga y estómago. Cuando se analizan los datos únicamente asociados al sexo masculino, es el cáncer de próstata el primero de la lista.

Este hecho es la motivación menester para la realización del presente trabajo. La elevada incidencia del cáncer de próstata hace que surja la necesidad de reacción en todos los niveles de la sociedad: campañas de concienciación, análisis obligatorios del PSA a varones mayores de 50 años para la detección precoz de la enfermedad, investigaciones clínicas sobre tratamientos, intervenciones e investigaciones científicas para el desarrollo de herramientas tecnológicas que sirvan de ayuda en estadios prematuros de la enfermedad.

El último punto es en el que se centra este trabajo: desarrollar un sistema de ayuda al profesional de la salud que sea capaz de detectar automáticamente los signos del cáncer a nivel histológico para el diagnóstico temprano de la enfermedad.

1.1 Cáncer de próstata

1.1.1 Aspectos fisiológicos de la glándula prostática.

La próstata es un órgano glandular interno que se encuentra en la pelvis de los varones, situado detrás del pubis, delante del recto e inmediatamente por debajo de la vejiga de la orina.

La glándula prostática aporta antígenos, fibrinógeno, espermina, fosfatasas ácidas (PSA principalmente). Las hormonas masculinas son las que estimulan el crecimiento de dicha glándula desde el desarrollo del feto. Si las hormonas masculinas desaparecen, la glándula no puede desarrollarse y reduce su tamaño.

Para comprender cómo se desarrolla el cáncer en este tejido, se ha de tener en cuenta que todas las células del organismo se dividen de forma regular con el fin de reemplazar aquellas envejecidas o muertas, manteniendo así su integridad y correcto funcionamiento.

Este proceso está regulado por mecanismos que indican cuándo deben dividirse y cuándo permanecer estables. Cuando se alteran dichos mecanismos, es cuando las células y sus descendientes inician una división incontrolada, que con el tiempo dará lugar a un tumor o nódulo.

Si estas células, además de crecer sin control, adquieren la capacidad de invadir los tejidos y órganos de alrededor y de trasladarse y proliferar en otras partes del organismo, el tumor anterior se denomina tumor maligno; es ahora cuando se puede comenzar a hablar de cáncer.

Desde el punto de vista histológico, una próstata sana se compone de unas 40-50 glándulas tubuloalveolares, que vacían su contenido en la uretra prostática a través de unos 20 conductos excretores largos independientes.

Como el propio nombre indica, las glándulas tubuloalveolares están constituidas por glándulas tubulares y alveolares. Las tubulares se originan por una invaginación en forma de tubo. Por otro lado, las alveolares se componen de dos partes, una porción proximal constituida por el conducto excretor y otra porción distal en forma de esfera, constituida por la porción secretora.

Las glándulas están incluídas en un estroma, compuesto en su mayor parte por células musculares lisas, mezcladas con tejido conjuntivo denso. En la Figura 1.1 están representadas estas estructuras. En la Sección 1.1.2 se explican con detalle los cambios que sufre este tejido conforme avanza la enfermedad.

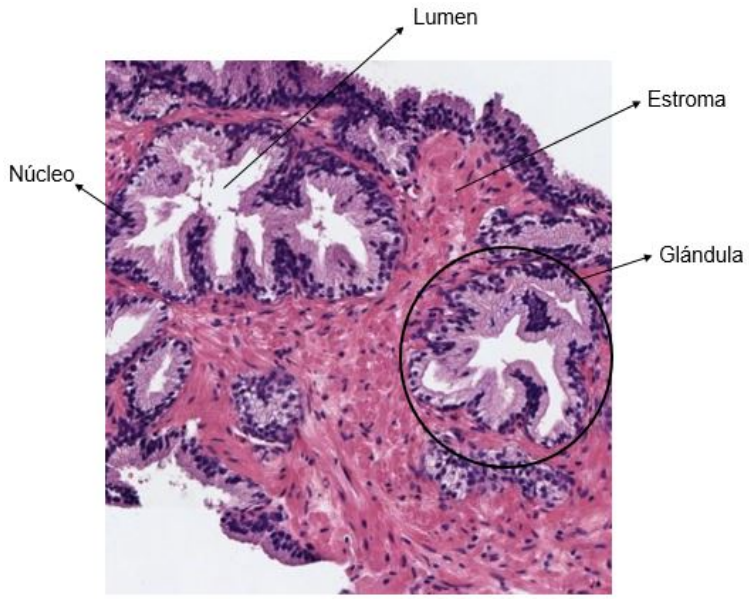


Figura 1.1: Partes básicas del tejido prostático sano: unidad glandular, núcleos, lúmenes y estroma.

1.1.2 Factores de riesgo, prevención y diagnóstico del cáncer de próstata

Un factor de riesgo se define como cualquier agente que incrementa el riesgo de padecer una enfermedad determinada.

En conreto, para el cáncer de próstata aparecen como factores de riesgo [2]:

- **Herencia**. Se ha estimado que el 10 % de los cánceres de próstata tienen un componente hereditario. El riesgo de padecer la enfermedad cuando el paciente posee antecedentes familiares de primer grado es dos veces superior.
- **Raza**. España, Grecia o Italia son los países europeos con menor tasa de aparición de la enfermedad. Por otro lado, en los varones afroamericanos es más frecuente que en el resto de la población.
- **Edad**. El cáncer de próstata afecta fundamentalmente a varones con edad avanzada (75 % de los casos se dan en mayores de 65 años).
- **Hormonas**. La enfermedad está influenciada por los andrógenos (testosterona). Los tumores disminuyen conforme lo hacen los niveles de esta hormona.
- **Dieta**. Existen evidencias de que dietas con alto contenido en grasas aumentan el riesgo de cáncer de próstata.
- Aunque no se hayan podido encontrar evidencias sólidas, la **obesidad y el consumo de alcohol y de tabaco** constituyen un factor de riesgo en la mayoría de los cánceres.

No existe una guía para prevenir el cáncer de próstata, ya que este está marcado por los factores de riesgo mencionados. Sin embargo, es obvio que se puede reducir el impacto de alguno de ellos, como por ejemplo los detallados en los dos últimos puntos.

Uno de los problemas que presenta el cáncer de próstata a la hora de su detección temprana es que no causa ninguna alteración perceptible por el paciente en las fases iniciales de la enfermedad. Es por este motivo por el que se recomienda a los varones mayores de 50 años realizarse análisis de sangre que busquen específicamente alteraciones en hormonas secretadas por la próstata.

Para tratar de detectar el cáncer de próstata en pacientes con un alto riesgo de padecer la enfermedad, se realizan los siguientes exámenes:

- Análisis del PSA (Prostate-Specific Antigen). PSA es una proteína que libera el tejido prostático a la sangre; su concentración se ve aumentada cuando hay actividad anormal en la próstata.
- Tacto rectal. Se detectan las anomalías macroscópicas del tejido, es decir, se palpa el tumor en sí.

Sin embargo, las pruebas anteriores no se pueden utilizar como método diagnóstico cuando sus resultados son anómalos. Siempre se complementan con una biopsia del tejido, que es la prueba determinante para el diagnóstico de la patología.

Una biopsia consiste, básicamente, en la extracción de una pequeña porción de tejido para que, después de su procesado, un patólogo la analice bajo el microscopio con el objetivo de determinar si el tejido extraído es normal o patológico.

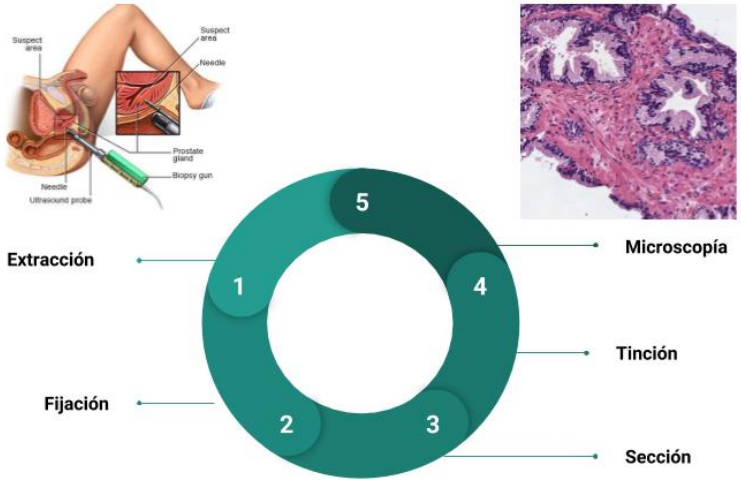


Figura 1.2: Esquema del procedimiento habitual para preparar muestras biológicas. (Imagen extracción obtenida de [4])

El procedimiento habitual [3] para preparar muestras que un patólogo pueda analizar de manera microscópica está recogido en la Figura 1.2 y consiste en:

Una vez extraída la biopsia, se incluye en un fijador (generalmente formol bufferado al 10 %) para evitar los cambios histológicos y no alterar los tejidos: hace precipitar las proteínas, aumenta la consistencia de los tejidos, inactiva las enzimas proteolíticas e inhibe el crecimiento bacteriano. El primer paso del análisis de la muestra, realizado por un patólogo, es una descripción macroscópica de la misma.

Después, se continúa con la preparación de la muestra: se secciona de forma que se seleccionen las partes representativas; se debe incluir parte del tejido sano vecino a la lesión para poder buscar infiltraciones en caso de tumores malignos. El tamaño de los bloques tisulares no debe exceder de 3 mm de espesor, 3 cm de largo y 2 cm de ancho.

El último paso consiste en teñir las muestras para resaltar y crear contraste entre las diferentes estructuras del tejido a estudiar. Para la gran mayoría de las muestras, la tinción más utilizada es Hematoxilina y Eosina; para el cáncer de próstata en concreto también.

Una vez se ha finalizado la preparación, el patólogo realiza una descripción microscópica. Trata de encontrar alteraciones en el tejido que sean signo de cáncer. Desde el año 1973, patólogos de todo el mundo emplean el mismo sistema de clasificación: la escala de Gleason [5].

Como idea general, este sistema asigna al tejido una puntuación de 1 a 5 en función del patrón arquitectónico de las glándulas del tumor, como se puede visualizar en la Figura 1.3. Es decir: Gleason 1 califica tejidos en estadios iniciales de la patología y Gleason 5 representa el estadio más agresivo (Tabla 1.1).

En la práctica, la clasificación comienza en Gleason 3, ya que los patrones 1 y 2 son poco frecuentes y su repercusión clínica es mínima. Por otro lado, Gleason 5 es muy fácil de distinguir, por lo que los esfuerzos en mejora del diagnóstico se centran en distinguir entre 3 estadios posibles: tejido sano, grado 3 y grado 4.

Clasificación	Características principales
Grado 1	Glándulas pequeñas, compactas y discretas.
Grado 2	Glándulas diferenciadas pero con variaciones de tamaño. Incremento del estroma entre glándulas.
Grado 3	Glándulas agrupadas o individuales. Las células comienzan a invadir el tejido circundante.
Grado 4	Pocas glándulas reconocibles. Muchas células invaden el tejido circundante.
Grado 5	No se reconocen glándulas. Solo se ven células, que no forman ninguna estructura propia del tejido.

Tabla 1.1: Principales características histológicas del tejido prostático según el sistema de clasificación de Gleason.

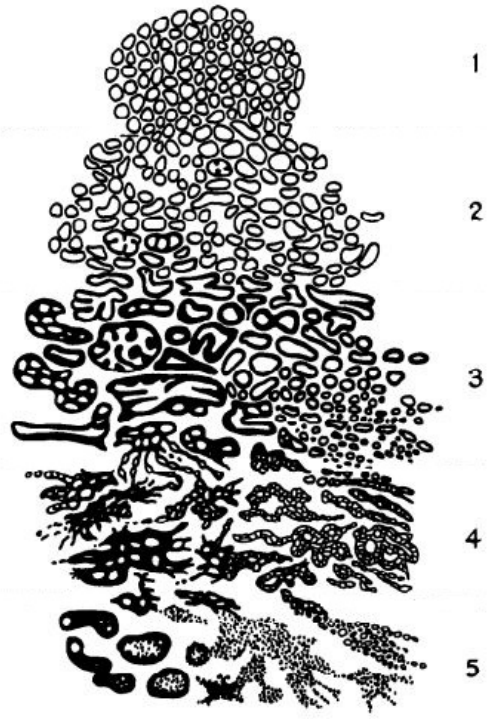


Figura 1.3: Patrones histológicos de tejido prostático clasificados según el sistema de clasificación Gleason [5].

El análisis microscópico de las muestras es muy costoso, ya que el patólogo ha de recorrer visualmente toda la muestra al microscopio óptico en busca de las anomalías propias del cáncer. Por otro lado, cuando se realiza un diagnóstico, este suele ser subjetivo, basado en la propia experiencia del profesional; no es capaz de proporcionar datos medibles, contrastables y reproducibles para la clasificación del tejido.

Estos hechos hacen patente la necesidad de crear herramientas automáticas de ayuda al diagnóstico clínico. Este objetivo es el que persigue el proyecto en el que se enmarca el presente trabajo fin de grado. SICAP (Sistema de interpretación de Imágenes histopatológicas para la detección del CÁncer de Próstata) es un proyecto financiado por el Ministerio de Economía, Industria y Competitividad [DPI2016-77869-C2-1-R] en el que participa la Universitat Politècnica de València (UPV) junto con la Universidad de Granada y el Servicio de Anatomía Patológica del Hospital Clínico Universitario de Valencia.

1.2 Estado del arte en la detección automática de cáncer y cáncer de próstata

El objetivo principal de este trabajo es desarrollar un algoritmo que lleve a cabo una clasificación a nivel local a partir de los diferentes patrones tisulares que se manifiestan en la imagen histológica de próstata. En concreto, se pretende crear un modelo predictivo que sea capaz de discernir entre localidades sanas y localidades que presenten signos típicos de cáncer de grado 3.

Para ello, se hace necesario pasar de muestras físicas a muestras digitales que permitan extraer la información relevante de las mismas mediante técnicas de tratamiento digital de imagen.

Hace algunos años se comenzaron a desarrollar escáneres digitales de microscopía (Whole Slide Scanners), con los que se generan imágenes digitales de la muestra completa con una gran calidad y resolución. La información de una muestra se adquiere en imágenes con gran aumento (40x-100x), realizando múltiples tomas secuenciales adyacentes hasta barrer toda la preparación, gracias a la cámara acoplada al microscopio. Para realizar este proceso, se hace uso de un microscopio robotizado, que es un sistema automático conectado al microscopio convencional. Este controla los movimientos de la platina del microscopio óptico convencional y automatiza la adquisición y almacenamiento de un patrón predeterminado de campos microscópicos con la ayuda de un computador. Cuanto mayor aumento se utilice, más pequeño será el campo visible y será entonces necesario un mayor número de tomas [6].

A partir de la muestra digitalizada es posible extraer información relevante:

- **Segmentación o división de la imagen en diferentes regiones.** Este conjunto de métodos se basan en que cada una de las regiones en las que se puede delimitar una imagen está caracterizada por un conjunto de parámetros, que tendrán valores diferentes de los del resto. El principio básico es que estos parámetros sean homogéneos dentro de la región y heterogéneos fuera de ella. Como técnicas de segmentación podemos encontrar la umbralización, el clustering, el crecimiento de regiones entre muchas otras.
- **Clasificación de patrones**. El objetivo, en este caso, es obtener la información relevante contenida en la imagen mediante el uso de descriptores apropiados que sean capaces de definirla y diferenciarla de otras. Es útil para determinar la posición, tamaño y orientación de objetos dentro de la imagen, para seleccionar objetos que cumplan con ciertos atributos o para reconocer un objeto de entre un conjunto.

En el presente trabajo fin de grado se seguirá la estrategia de clasificación de patrones para la consecución del objetivo principal. Dicha metodología se basa en dos grandes etapas: la extracción de características y la clasificación.

La etapa de extracción de características tiene como objetivo convertir la información relevante de una imagen en datos numéricos para poder encontrar patrones que permitan llevar a cabo una diferenciación automática entre las distintas clases objetivo. Esta conversión se lleva a cabo mediante descriptores de imagen capaces de caracterizar los diferentes objetos que aparecen en una imagen atendiendo a su morfología (tamaño y forma de los objetos), contexto, color (histograma), estructura, forma o textura, entre otros.

A partir de los datos numéricos obtenidos de la etapa de extracción de características, es posible llevar a cabo la generación de modelos automáticos capaces de predecir la pertenencia a cierta clase de futuras instancias. Para generar dichos modelos, se hace necesaria una etapa de entrenamiento empleando algoritmos de clasificación o aprendizaje automático (*machine learning* en inglés). Estos algoritmos pueden ser clasificados en dos grandes grupos atendiendo a la naturaleza del entrenamiento:

- **Apredizaje supervisado**. El sistema aprende de ejemplos, es decir, analiza una imagen y su etiqueta para encontrar una función que mapee entre las entradas y las salida deseadas.

Algunos ejemplos de algoritmos que utilizan esta aproximación son Support Vector Machine (SVM), árboles de decisión, K-Nearest Neighbour (KNN), Bayes, etc.

- **Apredizaje no supervisado**. En este caso, el sistema aprende modelos cuando solo se encuentran disponibles los datos, sin etiquetas o salidas explícitas. Por tanto, tratan de descubrir la estructura oculta de los datos para formar las salidas. Como ejemplos de esta clase de algoritmos encontramos los dendogramas, K-medias, DBSCAN, Expectation-maximization, etc.

En la literatura se pueden encontrar varios ejemplos de implementación de este tipo de técnicas para la detección automática de cáncer. Todos ellos emplean imagen histológica digitalizada para realizar la clasificación; sin embargo, emplean metodologías diferentes.

En [7] emplean el espacio de color de las imágenes para detectar los componentes básicos del tejido prostático; además de estos componentes, detectan mucina[1] azul que aparece en ocasiones en los lúmenes de las glándulas al avanzar la enfermedad. Una vez identificados, son empleados para la segmentación de las unidades glandulares completas. Las glándulas son empleadas en la clasificación; se les aplican descriptores de imagen para extraer 15 características estructurales. Con estas características realizan la clasificación de las imágenes en tres clases: tejido benigno, maligno con grado 3 o con grado 4 en la escala de Gleason.

En [8] proponen una nueva metodología para la segmentación de las unidades glandulares basada en la asociación de cada núcleo celular al lúmen más cercano. Se extraen 22 características que describen la información estructural y contextual de cada segmento. En este caso, clasifican las imágenes entre artefacto, glándulas sanas o glándulas cancerígenas. Para esta clasificación, han realizado varios experimentos para determinar cuál ofrece mejores resultados: clasificación binaria entre las clases o clasificación de las tres clases al tiempo.

En [9] plantean un método multi-resolución para el diagnóstico del cáncer de próstata empleando *microarrays* de tejido. De cada especimen de tejido se segmentan sus glándulas empleando características de intensidad y de textura. A partir de los resultados de la segmentación, se extraen características morfológicas de los lúmenes y de los núcleos para clasificar entre sanas y malignas.

En [10] segmentan automáticamente las unidades glandulares y su vecindario empleando redes neuronales convolucionales. Tras esto, extraen características de color, forma y textura de los superpíxeles que se corresponden con el interior y el exterior de las regiones glandulares. En la etapa de clasificación distinguen entre tejido maligno con grado 3 y tejido maligno con grado 4 en la escala de Gleason.

Otra forma de abordar este problema es sin segmentar las unidades glandulares. Ejemplo de esta metología se encuentra en [11]. Emplean conjuntamente *Local Binary Patterns (LBP)* y *One Dimensional Scale Invariant Feature Transform (1-D SIFT)* para obtener el vector de caractísticas que permita clasificar imágenes histológicas de hígado en sanas y cancerígenas.

En [12] han empleado los filtros de Gabor y LBP como descriptores de textura de las imágenes. Con estas características, clasifican las imágenes en benignas, grado 3, 4 o 5 en la escala de Gleason.

[1] En muchos adenocarcinomas se presenta una producción aumentada de mucinas, que son proteínas producidas por las células epiteliales.

Capítulo 2

Objetivos

El objetivo principal del presente trabajo es mejorar el rendimiento diagnóstico del cáncer de próstata en forma de algoritmo que sea capaz de detectar en imágenes histológicas de próstata los signos de cáncer de grado 3.

Este proceso de diagnóstico se traduce en un problema de clasificación entre dos posibles clases, en este caso: tejido sano y tejido patológico de grado 3. Se plantean varios objetivos parciales que conduzcan a este fin:

- Realizar un estudio del estado del arte para conocer las diferentes técnicas y procedimientos que se han empleado en los últimos años para resolver este mismo problema u otros similares por investigadores en todo el mundo.
- Crear dos bases de datos de imágenes histológicas. Una de ellas contendrá imágenes de tamaño 512×512 píxeles y la otra contendrá imágenes de tamaño 1024×1024 píxeles.
- Dividir las imágenes en dos conjuntos. Se realizará una partición de los datos para crear un conjunto de entrenamiento empleado para entrenar el modelo y un conjunto de test que se utilizará para evaluar la precisión y robustez del mismo. Cabe reseñar que ninguna imagen podrá estar contenida en ambos sets de datos, garantizando así la fiabilidad del método.
- Emplear diferentes descriptores de imagen. Se estudiarán y aplicarán novedosas técnicas de descripción de imagen que sean capaces de extraer la mayor cantidad de información relevante de la imagen histológica. Estos parámetros extraídos se emplearán en la etapa de clasificación para diferenciar entre ambas clases.
- Realizar un análisis estadístico de las características anteriores para seleccionar solo aquellas que aportan información útil para la clasificación, eliminando aquellos datos que vayan a resultar redundantes e imprecisos.
- Comparar el rendimiento de diferentes tipos de clasificadores seleccionando finalmente el que mejores resultados proporcione.

- Hacer una evaluación de la clasificación realizada con las muestras de test anteriormente mencionadas. Las figuras de mérito a maximizar por el modelo desarrollado serán la precisión, la sensibilidad y la especificidad, que se definirán a lo largo de la presente memoria.
- Extraer conclusiones a todos los niveles del proceso. Comprobar si el objetivo principal del trabajo se ha alcanzado con el éxito esperado y analizar el comportamiento del sistema determinando sus fortalezas y debilidades.
- Por último, resulta muy interesante establecer, en base a las conclusiones extraídas, posibles lineas futuras de trabajo que propicien una mejora de las prestaciones del sistema y permitan abarcar nuevos ámbitos más allá de los perseguidos en el presente trabajo.

Capítulo 3

Material y métodos

3.1 Material

Se va a proceder a detallar las diferentes herramientas que se han empleado para la realización de este trabajo.

3.1.1 Base de datos

La base de datos con la que se ha trabajado se ha construido a partir de imágenes histológicas de biopsia de próstata procedentes del departamento de anatomía patológica del Hospital Clínico Universitario de Valencia. En la Subsección 1.1.2 se ha explicado el procedimiento general de procesado de las biopsias y la técnica de escaneado para la digitalización de las mismas. En la Figura 3.1 se ilustra todo el proceso de creación de esta base de datos. Para este trabajo en concreto:

- Se trabaja con 60 imágenes digitalizadas a 40x. De estas imágenes, 17 fueron etiquetadas por el patólogo como sanas y 43 como patológicas (con distintos grados en la escala de Gleason).
- Estas imágenes fueron divididas en patches con un solapamiento del 50 % con el objetivo de conformar una amplia base de datos que permita el aprendizaje de modelos predictivos. El tamaño original de las imágenes es del orden de $[50000, 90000] \times [90000, 190000]$ píxeles, ocupando desde 300 MB hasta 1.5 GB. Los tamaños escogidos para estos patches son: 1024×1024 píxeles y 512×512 píxeles. Con ambos grupos de imágenes se divide la base de datos en otras dos, una para cada tamaño: BBDD_512 y BBDD_1024. El objetivo de disponer de dos tamaños diferentes de imagen es estudiar la influencia que tiene este parámetro en la construcción de los modelos.
- Cabe destacar que por cada región extraída de la imagen original, se dispone de una máscara binaria que delimita el tejido del fondo. Asimismo, también se dispone de una máscara

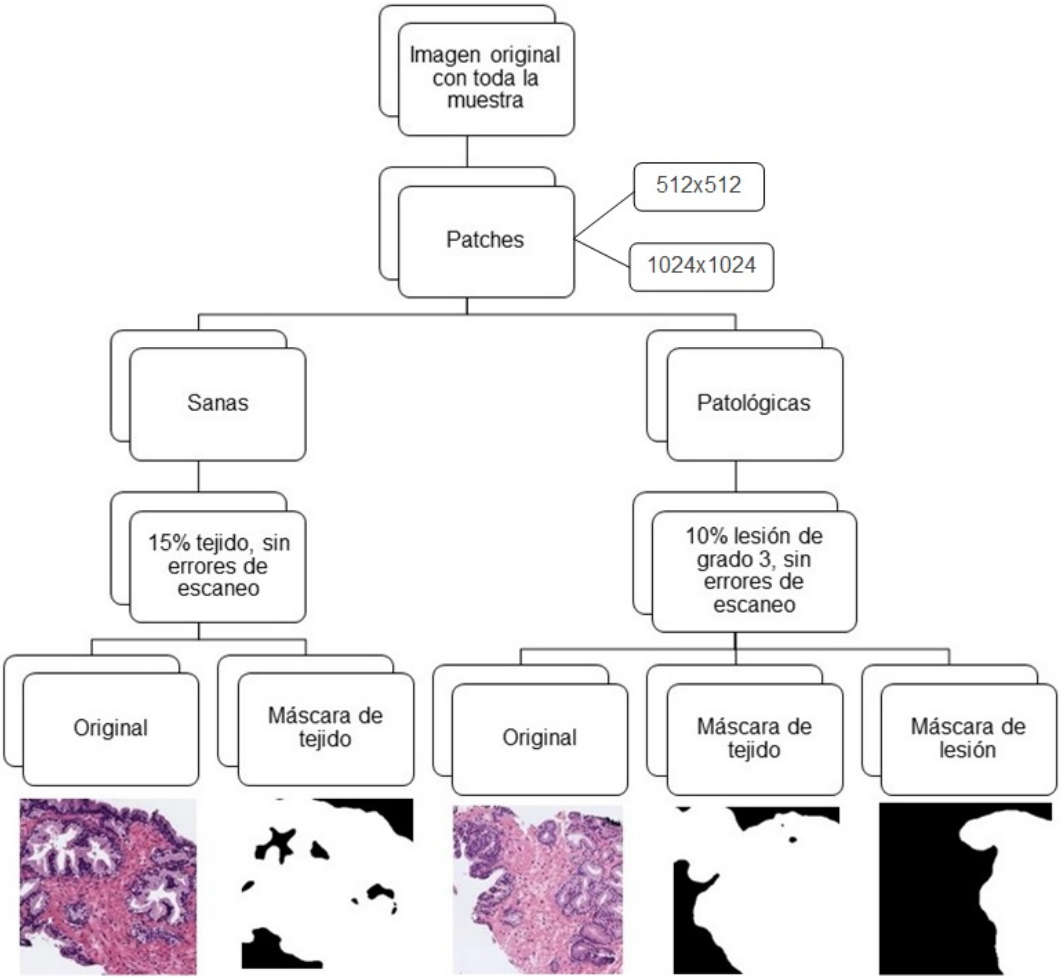


Figura 3.1: Esquema resumen del proceso de creación de la base de datos.

binaria que contiene las anotaciones de grado 3 (en la escala Gleason) que realizaron los patólogos sobre las imágenes originales mediante la aplicación web microDraw [13].

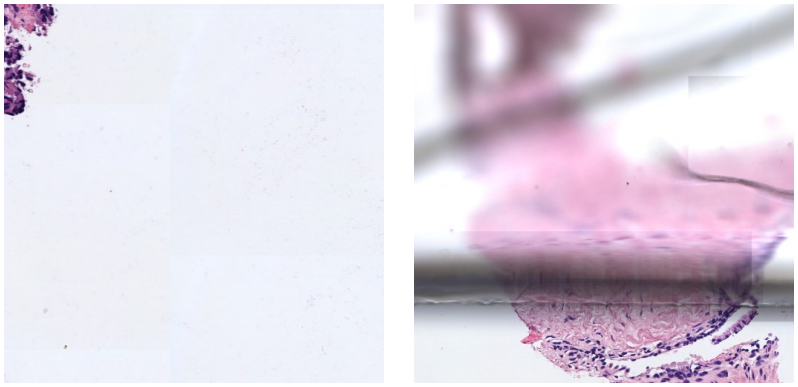
El resultado de la extracción de patches empleando diferentes tamaños de región proporciona: BBDD_512, que cuenta con 28472 imágenes sanas y 99675 imágenes patológicas y BBDD_1024, que contiene 6681 imágenes sanas y 26716 imágenes patológicas. Para que estas regiones resulten útiles y acaben por dar un buen resultado en la clasificación final, han de cumplir una serie de criterios:

- Las imágenes patológicas tienen, como ya se ha mencionado, una máscara de lesiones asociada. De la totalidad de estas imágenes, solo se utilizarán aquellas regiones que presenten un porcentaje de píxeles marcados como grado 3 superior al 10 %. El resto de la imagen, que será por tanto tejido "normal", se desecha. En la Figura 3.2 (a) se muestra la imagen original, en la Figura 3.2 (b) la máscara de la lesión asociada a esa imagen y en la Figura 3.2 (c) la región patológica que entra a formar parte de la base de datos finalmente.
- Para poder extraer información suficiente de una imagen sana se ha determinado que esta posea al menos un 15 % de tejido. En la Figura 3.3 (a) se muestra un ejemplo.

(a) (b) (c)

Figura 3.2: Imagen original (a), máscara de la lesión de grado 3 (b) y región patológica final (c).

- Alguna de estas imágenes tenía errores en el escaneado. También han sido eliminadas, como en el caso de la Figura 3.3 (b).



(a) (b)

Figura 3.3: (a) ejemplo de una imagen que no cumple con el porcentaje de tejido requerido para su análisis. (b) ejemplo de un error de escaneo.

Teniendo todo esto en cuenta, BBDD_512 contiene finalmente: 766 imágenes de grado 3 y 7783 imágenes sanas; BBDD_1024 contiene: 228 imágenes de grado 3 y 2366 imágenes sanas.

Una vez obtenidas las dos bases de datos anteriores se construye una tercera que contenga imágenes de ambas para poder analizar el efecto del tamaño de bloque empleado de manera que se realice un estudio multiresolución. Las imágenes se seleccionaron aleatoriamente de las bases de datos, conteniendo un 50 % de imágenes patológicas y un 50 % de imágenes sanas. Dentro de cada clase, la proporción de imágenes de cada base de datos también es el 50 %. Con todo esto, se ha construido BBDD_multi, que contiene: 456 muestras sanas y 456 muestras de grado 3.

3.1.2 Otros materiales

- Para la implementación de los algoritmos utilizados y el cálculo de todos los resultados del TFG se ha empleado el programa MATLAB® v.R2016b, The MathWorks, Inc. (Natick, Massachusetts, Estados Unidos). Este programa 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 arrays directamente. Resulta una herramienta muy útil para el análisis de datos, las comunicaciones inalámbricas, la visión artificial y el aprendizaje profundo, entre otros [14].
- Códigos públicos. Más adelante se hará referencia a distintos algoritmos que han sido creados por grupos de investigación o desarrolladores de uso público.
- Todo el TFG ha sido desarrollado y ejecutado en un ordenador personal. Concretamente se trata de un computador Acer Aspire E1 con procesador Intel Core i5 @2.6 GHz y memoria RAM de 8 GB. Estos datos se han de tener en cuenta si se quisieran reproducir los resultados obtenidos en cuanto al coste computacional que supone conseguirlos.

3.2 Metodología.

Una vez comentados todos los materiales que se han requerido para realizar el trabajo, se procede a describir la metodología empleada para la consecución de los resultados que se exponen en el siguiente capítulo de esta memoria.

En la Figura 3.4, el proceso comienza con un preprocesado de la base de datos, concretamente la deconvolución de color de las imágenes para obtener la aportación de cada tinción de manera independiente. Una vez las imágenes tienen la estructura deseada, se lleva a cabo una extracción de características a partir de descriptores de textura y morfología (filtros de Gabor, granulometría y matrices de coocurrencia). Antes de proceder a la etapa final de clasificación, se balancea la base de datos y se divide en dos grupos (train y test). Con las características extraídas de las imágenes de train se realiza la etapa de entrenamiento: se selecciona cuál es el mejor clasificador y se entrena con esos datos concretos. Finalmente, se le proporcionan al clasificador entrenado los datos extraídos de las imágenes de test para evaluar cómo se comporta frente a casos nuevos.

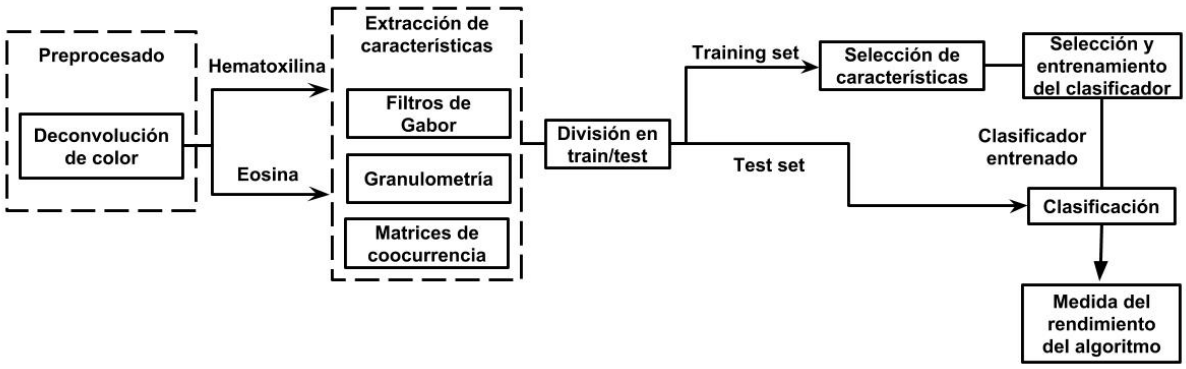


Figura 3.4: Diagrama de bloques de la metología seguida en este trabajo.

A continuación se procede a detallar cada uno de los bloques del diagrama expuesto en la Figura 3.4 en profundidad.

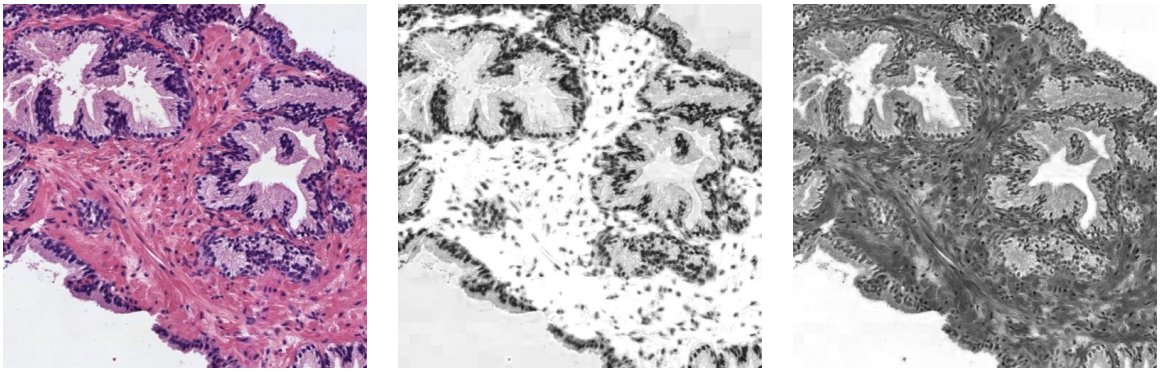
3.2.1 Deconvolución de color.

La técnica de preprocesado elegida para este trabajo es la deconvolución de color. Esta técnica obtiene dos imágenes a partir de cada imagen de entrada; cada una de ellas posee la información relativa a cada componente de la tinción empleada (Hematoxilina y Eosina). La idea básica de este procedimiento es que, debido a sus características químicas, cada componente tiene más afinidad por unas estructuras del tejido. Por tanto, si se separan ambas contribuciones, se apreciarán con mayor contraste unas estructuras de otras.

La tinción hematoxilina-eosina es la comúnmente empleada en el tratamiento de muestras histológicas en general y de muestras de cáncer de próstata en particular. La hematoxilina es básica y, por tanto, tiñe en tonos azul y púrpura estructuras ácidas (que son basófilas), como por ejemplo los núcleos celulares. La eosina, por el contrario, es una tinción ácida y tiñe de rosa componentes básicos de las muestras (que son acidófilos), como por ejemplo el citoplasma.

Aprovechando estas propiedades químicas de la tinción, se puede emplear la deconvolución de color para obtener imágenes de grises que realcen las características representativas de las estructuras clave en la identificación de glándulas malignas.

Como se puede apreciar en la Figura 3.5, a partir de una imagen histológica estándar (a), teñida con hematoxilina y eosina, se obtienen dos imágenes de grises: una que contiene la información relativa a la hematoxilina (b), en la que destacan principalmente los núcleos celulares, y otra que contiene la informació relativa a la eosina (c), en la que se ven con mayor claridad los lúmenes y citoplasmas.



(a) (b) (c)

Figura 3.5: Deconvolución de color de una imagen histológica de cáncer de próstata: (a) imagen original, (b) canal hematoxilina y (c) canal eosina.

La deconvolución de color mencionada se basa en que cada tinción pura estará caracterizada por un factor específico de absorción de la luz, c, en cada uno de los canales RGB [15].

La Ecuación 3.1 se corresponde con la Ley de Lambert-Beer, que describe la relación entre las intensidades de luz detectadas y la cantidad (A) de tinción y el factor de absorción c:

$$I_C = I_{0,C} \exp\left(-Ac_C\right) \tag{3.1}$$

siendo $I_{0,C}$ la intensidad de luz que incide sobre la muestra, I_C la intensidad de luz que se detecta y el subíndice C el canal de detección.

Es decir, la transmisión de la luz y, por tanto, el nivel de gris de cada canal, depende de forma no lineal de la concentración de tinción que se encuentre en la muestra. Debido a esta no-linealidad, no se puede emplear la Ecuación 3.1 para obtener la contribución de cada tinción en la imagen final.

El parámetro Densidad Óptica (OD) para cada canal es:

$$OD_C = -\log\left(\frac{I_C}{I_{0,C}}\right) = A * c_C \tag{3.2}$$

que, como se puede apreciar en la Ecuación 3.2, es lineal con la concentración de tinción absorbida por parte del material. Por tanto, este parámetro sí que se puede utilizar para separar las contribuciones de cada tinción en una muestra.

Cada tinción pura está relacionada con un OD específico para cada canal RGB; se puede caracterizar con un vector de dimensiones 1x3. Cuando existe una mezcla de tinciones, esta se puede caracterizar con una matriz, en la que cada fila representa una tinción y cada columna el valor de OD para cada canal de detección:

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

Para conseguir la información sobre la contribución independiente de cada tinción, hay que realizar una transformación ortonormal de la matriz anterior. Esta transformación se ha de normalizar para conseguir el peso correcto del factor de absorción de cada tinción por separado:

$$\begin{aligned} \hat{p}_{11} &= p_{11}/\sqrt{p_{11}^2 + p_{12}^2 + p_{13}^2} \\ \hat{p}_{21} &= p_{21}/\sqrt{p_{21}^2 + p_{22}^2 + p_{23}^2} \end{aligned} \tag{3.3}$$

Si $\boldsymbol{x}$ es el vector 2x1 para las cantidades de las dos tinciones en un píxel en particular y $\boldsymbol{M}$ la matriz de OD normalizada según la Ecuación 3.3, entonces el vector de niveles de OD detectados en ese píxel es:

$$y = \mathbf{xM} \longrightarrow \mathbf{x} = \mathbf{M}^{-1}[y] \tag{3.4}$$

La inversa de la matriz M se conoce como *Matriz de deconvolución de color*. Su diagonal tiene valores superiores a 1, mientras que el resto de elementos son negativos. Esto quiere decir que los valores OD corregidos para cada tinción se obtienen sustrayendo una porción de OD del resto de canales. De esta forma:

- Para obtener la contribución de hematoxilina en la muestra: se resta una porción de los canales verde y azul al canal rojo, que es el que se encuentra aumentado.
- Para obtener la contribución de eosina: se resta una porción de los canales rojo y azul al canal verde, que es el que se encuentra aumentado.

En el caso de que las tinciones utilizadas fueran puras, la matriz de deconvolución de color se correspondería con la matriz unitaria.

A partir del artículo científico que planteó el conocimiento teórico anterior [15], Tom Macura tradujo la función para su implementación, desarrollada en ImageJ Java Plugin, a código C compilable en Matlab mediante una función pasarela o MEX [16]. Esta función es la que se ha empleado para llevar a cabo la deconvolución de color en el presente trabajo.

Este paso del preprocesado es vital para la precisión y robustez del modelo predictivo entrenado, ya que va a conseguir realzar las estructuras que se ven alteradas conforme la enfermedad avanza; es decir, hará mucho más evidentes las diferencias entre imágenes sanas e imágenes de grado 3.

3.2.2 Extracción de características.

Como ya se ha señalado en la Sección 1.2, cuando se quiere caracterizar una imagen para ser capaces de extraer información útil de ella se emplean descriptores de imagen. Estos descriptores, básicamente, analizan la imagen y obtienen de ella una serie de parámetros, llamados características.

Para que la descripción resulte útil, ha de ser: única, completa e invariante frente a transformaciones geométricas y/o de intensidad.

Los descriptores empleados en este trabajo son de dos tipos: unos que describen la textura de la imagen (*matrices de coocurrencia* y *filtros de Gabor*) y otros que caracterizan el tamaño, forma e intensidad de los objetos en la imagen (*granulometría*).

Han sido elegidos estos descriptores en concreto porque la intención es analizar la imagen completa, sin segmentar sus partes para analizarlas por separado; en ese tipo de análisis es más apropiado describir las imágenes con características morfológicas, por ejemplo. También se ha considerado interesante comprobar el rendimiento que estos descriptores pueden ofrecer porque son muy pocos los ejemplos de su aplicación en el estado del arte.

3.2.2.1. **Matrices de coocurrencia.**

Las *matrices de coocurrencia* son descriptores que capturan la información de la textura de una imagen. Como idea general, son matrices cuadradas con dimensiones iguales al número de niveles de gris que tiene la imagen a analizar. Cada entrada de la matriz P(i,j) se corresponde con la probabilidad de que dos píxeles cualesquiera, separados una distancia d y con un ángulo θ concretos, tengan niveles de gris i y j, respectivamente. En otras palabras, la matriz de coocurrencia contiene la frecuencia relativa con la que dos píxeles de la imagen, con intensidades i y j respectivamente, aparecen separados una distancia d y un ángulo θ determinados.

La Figura 3.6 muestra un ejemplo de cómo, a partir de una imagen, se va creando la matriz de coocurrencia. Este caso se corresponde con d=1 y $\theta = 0^{o}$.

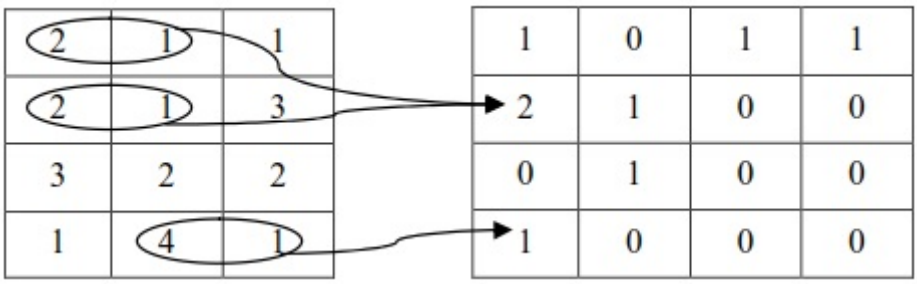
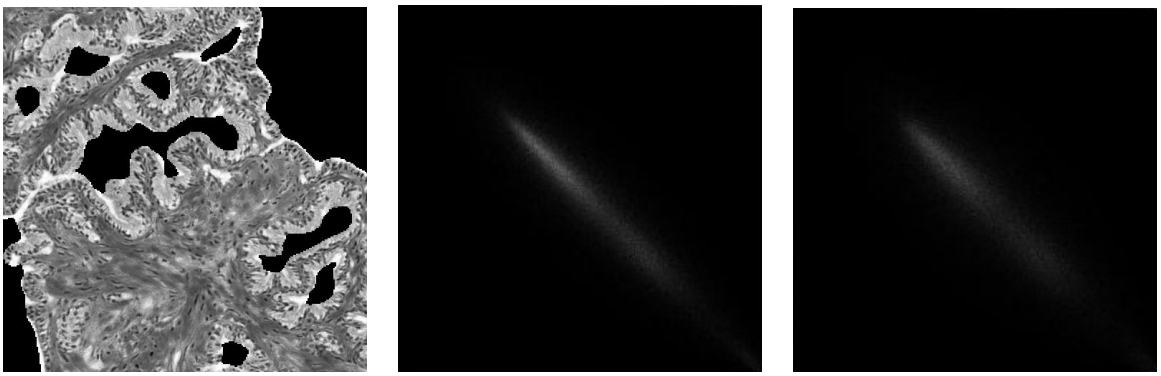


Figura 3.6: Ejemplo sencillo de construcción de una matriz de coocurrencia [17].

Para detectar patrones de textura en la imagen, suele ser necesario calcular varias matrices, correspondientes a distintas combinaciones de d y θ.

Cuando una matriz de coocurrencia contiene valores elevados a lo largo de la diagonal, esta está describiendo una textura suave, ya que los pares de píxeles tendrán un nivel de gris similar. Sin embargo, valores de la diagonal bajos ponen en evidencia la existencia de una macrotextura en la imagen, ya que los niveles de gris separados por d serán diferentes. Un ejemplo con una imagen real de la base de datos del trabajo se puede ver en la Figura 3.7: (a) es la imagen original a partir de la cual se han calculado dos matrices de coocurrencia: d=1 y $\theta = 0^o$ (b) y d=1 y $\theta = 45^o$ (c).



(a) (b) (c)

Figura 3.7: Matrices de coocurrencia calculadas a partir de (a). La imagen central se corresponde con $(d, \theta) = (1, 0^o)$; la imagen de la derecha se corresponde con $(d, \theta) = (1, 45^o)$.

Lo habitual es expresar las matrices de forma normalizada; los valores calculados se dividen entre el número total de píxeles contenidos en la imagen original.

Las matrices de coocurrencia tienen como ventaja que conservan la información espacial de las imágenes. Por ejemplo, un estadístico basado en el histograma daría el mismo resultado para un tablero de ajedrez con los cuadros blancos y negros intercambiados.

En [18] se desarrolló un método con el que poder extraer características que definan completamente una textura a partir de matrices de coocurrencia. Estas características están calculadas en el dominio espacial, a partir del hecho de que las texturas tienen naturaleza estadística: se asume que la información textural de una imagen está explicada con totalidad en las relaciones espaciales que unos tonos de gris tienen con otros en la imagen.

Se determinaron un conjunto de 14 características que demostraron ser aptas para describir completamente las texturas de la imagen de entrada. Algunas de estas medidas están relacionadas con características específicamente de la textura de la imagen y otras caracterizan la complejidad y la naturaleza de las transiciones en los niveles de gris.

- p(i,j). Es la intrada (i,j)-ésima en la matriz normalizada.
- $p_x(i)$. Es la entrada i-ésima de la matriz de probabilidad marginal obtenida sumando las filas de p(i,j):

$$p_x(i) = \sum_{j=1}^{N_g} P(i,j) \tag{3.5}$$

- N_g. Es el número de niveles de gris que hay en la imagen (y, por tanto, el tamaño de la matriz resultante).
- $\sum_i = \sum_{i=1}^{N_g}$
- $\sum_j = \sum_{j=1}^{N_g}$
- $p_y(j) = \sum_i p(i,j)$
- $p_{x+y}(k) = \sum_{\substack{i \\ |i+j|=k}} \sum_j p(i,j)$, con k=2,3,...,$2N_g$.
- $p_{x-y}(k) = \sum_{\substack{i \\ |i-j|=k}} \sum_j p(i,j)$, con k=0,1,...,$N_g$-1.

Estas características son:

- **Contraste**. Es una medida de las variaciones locales de los niveles de gris presentes en la imagen. Cuando hay grandes cambios, el contraste es alto. Este parámetro caracteriza también la dispersión de los valores de la matriz con respecto a la diagonal principal.

$$Contraste = \sum_{n=0}^{N_{g-1}} n^2 \left\{ \sum_{\substack{i=1 \\ |i-j=n|}}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right\} \tag{3.6}$$

- **Homogeneidad** o diferencia inversa del momento. Mide la homogeneidad local de la imagen; asigna valores más altos a las diferencias más pequeñas entre pares de píxeles. Por tanto, tiene el comportamiento contrario al contraste.

$$Homogeneidad = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j) \tag{3.7}$$

- **Energía** o segundo momento angular. Es una medida de la homogeneidad de la imagen. Imágenes muy homogéneas tienen pocas transiciones dominantes de niveles de gris, lo que resulta en una energía más alta.

$$Energia = \sum_i \sum_j \{p(i,j)\}^2 \tag{3.8}$$

- **Entropía**. Mide la no uniformidad de la imagen. Si la imagen es heterogénea, muchos elementos de la matriz de coocurrencia tendrán valores pequeños, lo que se traduce en una entropía muy grande. Está correlada de manera inversa con la energía.

$$Entropia = -\sum_i \sum_j p(i,j) \log(p(i,j)) \tag{3.9}$$

- **Varianza** o suma de cuadrados. Mide la heterogeneidad de la imagen y está fuertemente correlacionado con la desviación típica de esta. Caracteriza la distribución de los niveles de gris alrededor de la media. Por tanto, la varianza se ve aumentada cuando los niveles de gris difieren de la media.

$$Varianza = \sum_i \sum_j (i-\mu)^2 p(i,j) \tag{3.10}$$

- **Correlación**. Indica la fuerza de la relación lineal y la proporcionalidad entre dos variables. Se considera que dos variables están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores de la otra.

$$Correlacion = \frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3.11}$$

- **Información de la medida de la correlación 1 y 2**:

$$IMC1 = \frac{HXY - HXY1}{\text{máx}\{HX, HY\}} \tag{3.12}$$

$$IMC2 = (1 - \exp\left[-2{,}0(HXY2 - HXY)\right])^{1/2} \tag{3.13}$$

Donde HX y HY son las entropías de p_x y p_y (ecuación 3.9) y:

$$HXY = -\sum_i \sum_j p(i,j) \log(p(i,j)) \tag{3.14}$$

$$HXY1 = -\sum_i \sum_j p(i,j) \log\{p_x(i) p_y(j)\} \tag{3.15}$$

$$HXY2 = -\sum_i \sum_j p_x(i) p_y(j) \log\{p_x(i) p_y(j)\} \tag{3.16}$$

En la Tabla 3.1 se muestran el resto de características y su expresión matemática.

Algunos años más tarde, en [19] se propusieron nuevas características a extraer de las matrices de coocurrencia . Estas están recogidas en la Tabla 3.2.

Estas son las 19 características que se han empleado en el presente trabajo para describir las texturas de las imágenes a partir de matrices de coocurrencia.

El primer paso es calcular, a partir de cada imagen de hematoxilina y de eosina, dos matrices de coocurrencia, una con los parámetros (d,θ)=(1,0º) y otra con (d,θ)=(1,45º). Estas matrices tendrán tamaño 256×256, ya que es el número de niveles de gris que tienen las imágenes.

Características	Ecuaciones	
Promedio acumulado	$SumP = \sum_{i=2}^{2N_g} ip_{x+y}(i)$	(3.17)
Varianza acumulada	$SumV = \sum_{i=2}^{2N_g} (i - sumE)^2 p_{x+y}(i)$	(3.18)
Entropía acumulada	$SumE = -\sum_{i=2}^{2N_g} p_{x+y}(i) \log(p_{x+y}(i))$	(3.19)
Entropía diferencial	$DifE = -\sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$	(3.20)
Varianza diferencial	$DifV =$ varianza de p_{x-y}	(3.21)
Coeficiente de correlación máximo	$MCC =$ (segundo mayor eigenvalor de Q)$^{1/2}$ $Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$	(3.22)

Tabla 3.1: Expresiones matemáticas de las características de [18].

A partir de cada una de estas matrices, se calculan las 19 características mencionadas haciendo uso de una función con código libre desarrollada por Patrik Brynolfsson en febrero de 2016.

Siguendo este procedimiento, se obtienen finalmente 76 características por cada imagen de entrada: 38 para la componente de Hematoxilina de la imagen (19 para (d,θ)=(1,0º) y 19 para (d,θ)=(1,45º)) y otras 38 para la componente de Eosina de la imagen (19 para (d,θ)=(1,0º) y 19 para (d,θ)=(1,45º)).

3.2.2.2. **Filtros de Gabor.**

La *transformada de Gabor* se puede definir como una sinusoide compleja modulada por una función gaussiana. Sin embargo, es habitual descomponer la señal en sub-bandas de frecuencia y procesarla en estos segmentos, ya que se presupone que la energía de la señal posee distribuciones específicas a lo largo del dominio de la frecuencia. Esta descomposición de la señal consiste simplemente en aplicar varios filtros en cascada; esta técnica se conoce como *banco de filtros*. Son capaces de eliminar la complejidad que tienen los métodos estadísticos para caracterizar texturas en una imagen. En el caso de un banco de filtros de Gabor, cada filtro utiliza frecuencias espaciales y orientaciones de la sinusoide y de propagación de la gaussiana en las direcciones x e y (σ_x y σ_y) concretas [20].

Características	Ecuaciones	
Autocorrelación	$AutoCorr = \sum_i \sum_j (ij)p(i,j)$	(3.23)
Cluster prominence	$ClusterP = \sum_i \sum_j (i+j-\mu_x-\mu_y)^4 p(i,j)$	(3.24)
Cluster Shade	$ClusterS = \sum_i \sum_j (i+j-\mu_x-\mu_y)^3 p(i,j)$	(3.25)
Disimilitud	$Dis = \sum_i \sum_j \lvert i-j \rvert p(i,j)$	(3.26)
Probabilidad máxima	$MaxP = \max_{i,j} p(i,j)$	(3.27)

Tabla 3.2: Ecuaciones matemáticas de las características propuestas por [19].

La *función elemental de Gabor* se representa como:

$$h(x,y) = g(x',y') \exp\left(j2\pi(U_x + V_y)\right) \tag{3.28}$$

siendo U_x e V_y las frecuencias espaciales en cada dirección.

Una función Gabor en 2-D:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(\frac{1}{2}\left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right]\right) \tag{3.29}$$

donde: x'=xcos θ+ysin θ e y'=-xsin θ+ycos θ.

Teniendo en cuenta las ecuaciones anteriores, se puede reescribir la función Gabor elemental como:

$$g(x,y) = \frac{1}{2\pi} \exp\left[\frac{x^2+y^2}{\sigma_x+\sigma_y}\right] \exp\left[j2\pi u_0(x\cos\theta + y\sin\theta)\right] \tag{3.30}$$

donde σ_x y σ_y son la propagación de la gaussiana en las direcciones x e y, respectivamente.

Si se asume que σ_x=σ_y, u_0 es la frecuencia central de la sinusoide y su orientación será:

$$U_0 = \sqrt{U^2 + V^2} \tag{3.31}$$

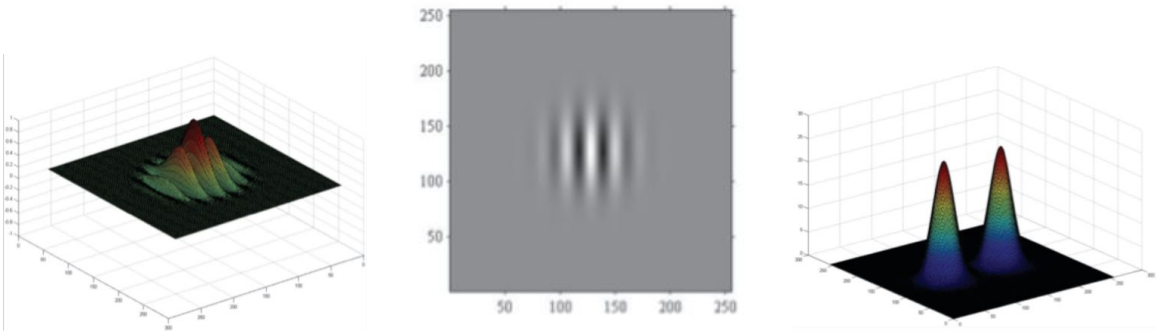
En el dominio de la frecuencia, se puede escribir el filtro de Gabor como:

$$H(u,v) = \exp\left\{-2\pi^2\sigma^2\left[([u-U]')^2 + ([v-V]')^2\right]\right\} \tag{3.32}$$

con u'=ucos θ+vsin θ y v'=usin θ+vcos θ.

Es importante seleccionar correctamente los parámetros de los filtros para poder caracterizar las texturas presentes en la imagen.

En la Figura 3.8 hay una representación de un filtro Gabor 2-D, con su respuesta en los dominios espacial y frecuencial.



(a) (b) (c)

Figura 3.8: Ejemplo de un filtro Gabor 2-D. (a) sinusoide modulada por una gaussiana, (b) respuesta en el dominio espacial y (c) respuesta en el dominio de la frecuencia espacial [21].

El resultado del filtrado de una imagen es una imagen filtrada. Por tanto, no se puede emplear la salida del banco de filtros de Gabor como características para el análisis de las texturas de la imagen. En este trabajo se han propuesto cuatro características para caracterizar las texturas a partir de las imágenes filtradas, recogidas en la Tabla 3.3.

Es necesario explicar con más profundidad la Ecuación 3.36: p(k) se corresponde con cada una de las entradas del histograma normalizado de la imagen filtrada, con k=1,...,I-1, siendo I el número de niveles de gris de la imagen.

Como ya se ha mencionado, para la implementación de un banco de filtros de Gabor hay que determinar sus parámetros principales. En este caso:

- **Orientación** en grados. Se ha empleado θ=[0 30 60 90 120 150]º.
- **Wavelength** de la sinusoide en pixels/ciclo. Se ha empleado σ=[2 4 6 8 10 12] pixels/ciclo.

Con estos parámetros se construye un banco de filtros con concretamente 36 filtros diferentes, uno para cada posible combinación entre los parámetros anteriores.

Una vez que se tienen los parámetros de todos los filtros determinados, se procede a filtrar las imágenes con la función de MATLAB *imgaborfilt*, a partir de la cual se obtienen tanto la magnitud como la fase de las imágenes filtradas.

Para hacer al descriptor invariante con respecto a la rotación, se promedian los resultados de los 36 filtros con la misma wavelength para todas las orientaciones, de manera que finalmente se obtienen 12 imágenes filtradas (6 relativas a la magnitud y 6 a la fase) [22].

Características	Ecuaciones
Desviación media absoluta de la media	$DMA = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\lvert x_{i,j} - \bar{x}\rvert$ (3.33)
Media	$\bar{x} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} x_{i,j}}{MN}$ (3.34)
Desviación típica	$\sigma = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j} - \bar{x})^2}{MN}}$ (3.35)
Energía	$E = \sum_{k=0}^{I-1} p(k)^2$ (3.36)

Tabla 3.3: Características obtenidas a partir de un filtro de Gabor empleadas para la determinación de texturas en la imagen.

En la Figura 3.9 se muestra un ejemplo de la aplicación de este algoritmo.

A partir de cada una de estas imágenes se calculan las 4 características recogidas en la Tabla 3.3. Por tanto, para cada imagen de entrada se obtienen 96 características: 4 características*6 wavelengths*2 tipos de imagen (H y E)*2 salidas del filtro (magnitud y fase).

3.2.2.3. **Granulometría.**

La *morfología matemática* se ha convertido en las últimas décadas en una herramienta empleada de manera recurrente en el campo del procesamiento digital de imagen. La morfología se refiere al estudio la estructura o forma de objetos en imágenes. Considera que la imagen está formada por conjuntos (regiones) de píxeles a los que, por tanto, se les puede aplicar herramientas de la Teoría de Conjuntos. En general, se puede definir como un procesado no lineal basado en operaciones de máximos y mínimos. En los filtros lineales de imagen, h[m,n] caracteriza completamente al sistema; en el caso de los operadores morfológicos, lo hace el *elemento estructurante (EE)*, que determina la operación concreta a realizar.

Los operadores básicos de la morfología matemática son:

- **Dilatación**. Se desplaza por la imagen el EE traspuesto y se calcula el supremo de los píxeles de la imagen bajo el mismo.

$$Y = \delta_B(X) = sup\{X_b, b \in B\} = \{B_x, x \in X\} \tag{3.37}$$

$$\delta_B(X) = X \oplus B \tag{3.38}$$

  En imágenes de grises, produce los siguientes efectos:

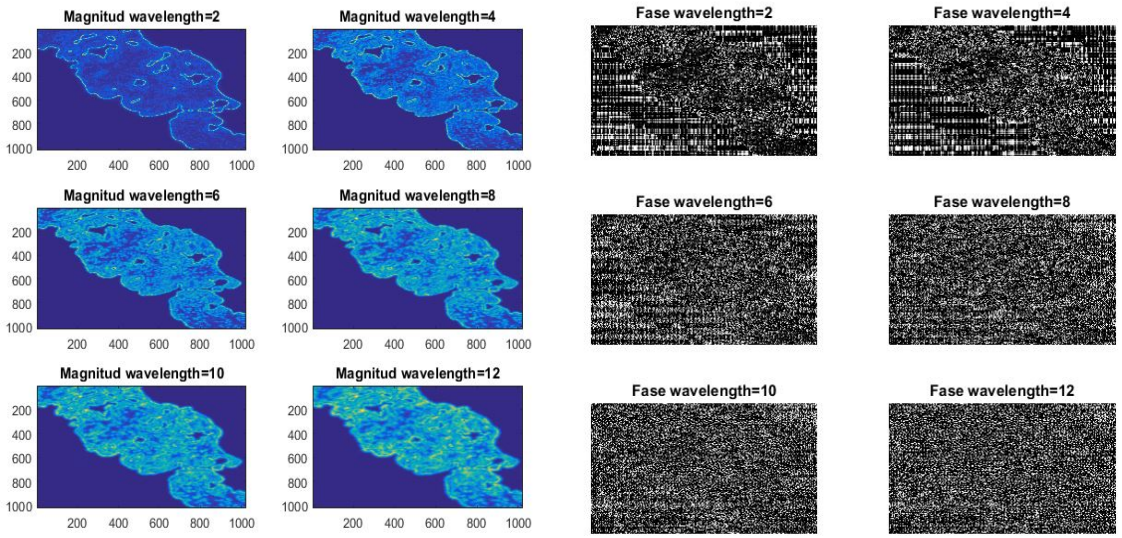


Figura 3.9: Imágenes filtradas promediadas con respecto a la wavelength. (a) magnitud, (b) fase.

- Elimina las zonas oscuras más estrechas que el EE.
- Estrecha las zonas oscuras más anchas que el EE.
- Ensancha las zonas claras.
- Siempre está por encima de la señal original.

- **Erosión**. Se desplaza por la imagen el EE y se calcula el ínfimo de los píxeles de la imagen bajo el mismo.

$$Y = \varepsilon_B(X) = inf\{X_b, b \in B^*\} = \{z, B_z \in X\} \tag{3.39}$$

$$\varepsilon_B(X) = X \ominus B \tag{3.40}$$

En imágenes de grises, produce los siguientes efectos:

- Elimina las zonas claras más estrechas que el EE.
- Estrecha las zonas claras más anchas que el EE.
- Siempre está por debajo de la señal original.

A partir de la combinación de los operadores anteriores se consiguen otros, que simplifican las señales en forma de filtro paso bajo no lineal. Permiten seleccionar objetos por tamaños:

- **Apertura**. Erosión seguida de dilatación con el mismo EE. Elimina ramificaiones estrechas de zonas claras y las zonas claras en las que no cabe el EE.

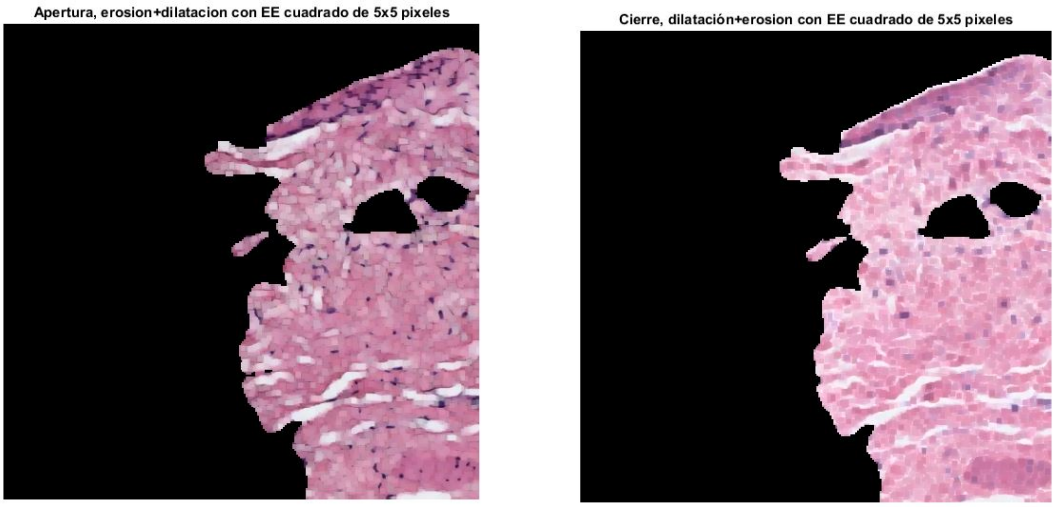
$$\gamma_B(X) = \delta_B(\varepsilon_B(X)) \tag{3.41}$$

- **Cierre**. Dilatación seguida de erosión con el mismo EE. Elimina ramificaiones estrechas de zonas oscuras y las zonas oscuras en las que no cabe el EE.

(a) (b) (c)

Figura 3.10: Ejemplo de dilatación (b) y erosión (c) de una imagen.

$$\varphi_B(X) = \varepsilon_B(\delta_B(X)) \tag{3.42}$$



(a) (b)

Figura 3.11: Ejemplo de apertura (a) y cierre (b) de una imagen.

Una de las técnicas más interesantes basada en morfología matemática es la *granulometría*. La granulometría y algunas métricas obtenidas a partir de ella se pueden aplicar para numerosas tareas, como segmentación, extracción de características, caracterización de texturas o estimación del tamaño, entre otras.

Consiste en la aplicación de una serie de operaciones de apertura o cierre morfológicos con elementos estructurantes de tamaño creciente. Con ella, se obtienen las distribuciones de tamaño de los diferentes objetos en una imagen. La información que aportan estas distribuciones es diferente en función del tipo de textura a la que se aplique: para texturas ordenadas, aporta información de la forma y el tamaño y para texturas desordenadas, extrae el grado de granularidad.

Cuando se computan aperturas en una imagen con un EE que crece de tamaño (λ), se obtiene una *pirámide de aperturas morfológicas*, que se conoce como *perfil granulométrico*. Esto puede ser formalizado como:

$$\Pi_\gamma(f) = \{\Pi_{\gamma\lambda} \colon \Pi_{\gamma\lambda} = \gamma_\lambda(f), \forall\lambda \in [0, ..., n_{max}]\} \tag{3.43}$$

donde n_{max} es el tamaño máximo del EE.

Dualmente, se pueden construir *pirámides de cierres morfológicos* computando cierres con un EE que crece de tamaño (λ). En este caso, lo que se obtiene es el perfil antigranulométrico. Se define formalmente como:

$$\Pi_\varphi(f) = \{\Pi_{\varphi\lambda} \colon \Pi_{\varphi\lambda} = \varphi_\lambda(f), \forall\lambda \in [0, ..., n_{max}]\} \tag{3.44}$$

Haciendo uso de las pirámides de aperturas morfológicas explicadas, se puede definir un descriptor de forma: la *curva de granulometría* (o espectro del patrón), que asigna cada tamaño n a alguna medida de las estructuras brillantes de la imagen con ese tamaño. Es una función de densidad de probabilidad (histograma) en la que un impulso grande a un tamaño determinado indica la presencia de muchas estructuras en la imagen a ese tamaño.

$$PS_\Gamma(f, n) = \frac{m(\Pi_{\gamma n}(f)) - m(\Pi_{\gamma n+1}(f))}{m(f)} \tag{3.45}$$

con $n \geq 0$.

Análogamente, con las pirámides de cierres morfológicos se puede definir la curva de antigranulometría, que caracteriza el tamaño de las estructuas oscuras de la imagen.

$$PS_\Phi(f, -n) = \frac{m(\Pi_{\varphi n}(f)) - m(\Pi_{\varphi n+1}(f))}{m(f)} \tag{3.46}$$

Para implementar esta técnica en el trabajo, se determinaron:

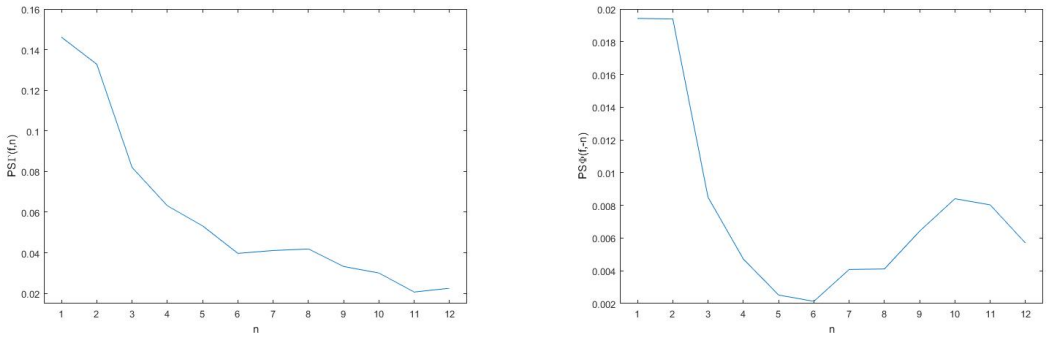
- Los parámetros que regulan y definen el elemento estructurante que se va a utilizar para construir las pirámides de cierre morfológico:
  - Tamaño de inicio= 0.
  - Tamaño final= 24.
  - *Step*: incremento de tamaño entre un cierre y el siguiente= 2.
  - Tipo de EE: disco. Por tanto, los tamaños anteriores se refieren al radio de este disco.
- Los parámetros que regulan y definen el EE que se va a utilizar para construir las pirámides de apertura morfológica:
  - Tamaño de inicio= 0.
  - Tamaño final= 48.
  - *Step*: incremento de tamaño entre una apertura y la siguiente= 4.

- Tipo de EE: disco. Por tanto, los tamaños anteriores se refieren al radio de este disco.

Con estos parámetros, se van calculando aperturas o cierres consecutivos, incrementando el tamaño del EE cada vez según el *step* que se haya determinado.

Es importante especificar que las pirámides de cierre morfológico se aplican a las imágenes de hematoxilina, ya que el objetivo es detectar los núcleos celulares, que son los elementos oscuros de estas imágenes. Por su lado, las pirmámides de aperturas morfológicas se aplican a las imágenes de eosina; se quieren detectar los lúmenes de las glándulas, que son los elementos claros de estas imágenes.

Se obtienen los patrones espectrales (PS) analizando el cambio de intensidad que provoca la técnica en las imágenes. Por tanto, PSop es el espectro de la apertura y PScl el del cierre. Estos espectros tienen 12 puntos, uno por cada apertura/cierre que se ha realizado. En la Figura 3.12 se puede observar un ejemplo a partir de una de las imágenes de la base de datos.



(a) (b)

Figura 3.12: Ejemplo de los espectros de apertura (a) y cierre (b) obtenidos aplicando la granulometría a una imagen de la base de datos.

Estos espectros son empleados directamente como características en este trabajo. Por tanto, para cada imagen de entrada se obtienen 24 características, 12 referidas a la imagen de hematoxilina y 12 a la de eosina.

3.2.3 Balanceo y división de las muestras en train/test.

Para poder hacer un estudio estadístico robusto con las características extraídas de las imágenes es necesario equilibrar el número de muestras sanas y de grado 3 que, como se puede apreciar en la Tabla 3.4, son muy dispares (exceptuando las muestras de BBDD_multi, que se creó a partir de las otras dos y se procuró que ya estuviera balanceada).

Tamaño de muestras	Grado 3	Sanas
512×512	766	7783
1024×1024	228	2366
Multiresolución	456	456

Tabla 3.4: Resumen del contenido de las bases de datos.

La clase limitante son las muestras patológicas en ambas bases de datos. El objetivo es, por tanto, seleccionar de entre las muestras sanas la misma cantidad que de patológicas. Esta selección se ha realizado de manera aleatoria.

Una vez la base de datos tiene la estructura deseada, hay que reservar una cantidad de muestras para que el clasificador empleado en el último paso del proceso no aprenda de ellas y puedan ser utilizadas para testear la calidad del método desarrollado. Es decir, se utilizan como si fueran muestras nuevas que se quieren clasificar.

Se ha decidido reservar el 20 % de todas las muestras como test, manteniendo el balanceo entre sanas y patológicas.

Por tanto, tras aplicar esto, las bases de datos se vuelven a dividir en dos grupos diferentes:

- **Train**. Contiene el 80 % de las muestras. Estas son las imágenes que van a seguir el proceso completo y van a servir para entrenar el clasificador final.
  - Con tamaño 1024×1024, finalmente se tienen 364 muestras.
  - Con tamaño 512×512 se tienen 1226 muestras.
  - En la base de datos con ambos tamaños de imagen se tienen 730 muestras.
- **Test**. Contiene el 20 % de las muestras. Son las imágenes que se van a emplear para medir la calidad del algoritmo.
  - Con tamaño 1024×1024, se reservan 92 imágenes.
  - Con tamaño 512×512, se reservan 306 imágenes.
  - En la base de datos con ambos tamaños de imagen se reservan 182 imágenes.

3.2.4 Selección de características.

Una vez acabado el proceso de extracción de características, es preciso realizar un análisis estadístico para poder discernir qué características sirven para la clasificación y cuáles aportan información redundante y por tanto, no mejoran el rendimiento de la clasificación.

Este proceso de selección se va a realizar solo sobre las características de train, que serán aquellas de las que el clasificador va a aprender. Se barajan concretamente 196 características por cada patch.

Además de estas 196 características, es imprescindible tener su *groundtruth*, es decir, sus etiquetas: 0 si el patch es sano y 1 si es patológico. Esta información sirve, por un lado, para realizar el análisis estadístico de manera que se determine si las características sirven para diferenciar ambas clases y, por otro lado, para la etapa de clasificación, ya que se van a emplear algoritmos de clasificación supervisada que necesitan de las etiquetas de los datos para el proceso de aprendizaje.

El análisis estadístico se divide en dos etapas diferentes: el estudio de la capacidad discriminatoria de las características y el análisis de la independencia entre pares de variables.

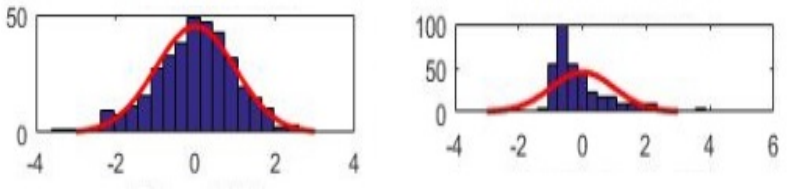
Antes de comenzar con el análisis en sí mismo, es preciso normalizar las características. En este trabajo se han normalizado calculando su *Zscore*:

$$Z = \frac{X - mean(X)}{STD(X)} \tag{3.47}$$

- X es la matriz de características.
- MU=μ=mean=media.
- SIGMA=σ=STD=desviación típica.

Tras esto, comienza el proceso de selección:

- **Estudio de la capacidad discriminatoria de las características** para determinar cuáles van a resultar útiles para separar las dos clases. Se realizan una serie de test estadísticos que rechazan o aceptan en cada caso una hipótesis nula (H_0) diferente. El resultado de cada test tiene asociado un p-valor[1], el cual se comparará con α[2]: si el p-valor es menor que α, se rechaza la hipótesis nula con un 99.99 % de confianza. Por el contrario, cuando este valor sea mayor que α, se acepta con ese mismo nivel de confianza.
  - El primer paso es **estudiar la normalidad de las características**. Con este objetivo se realiza el *test de Kolmogorov-Smirnov (KS test)*, que compara cada muestra con una distribución de probabilidad de referencia, en este caso una distribución normal N(0,1). El resultado de este test es importante a la hora de realizar los siguientes pasos, ya que según sigan o no una distribución normal, la capacidad discriminatoria de las características se analizará de una u otra forma. En la Figura 3.13 hay un ejemplo de una característica que sigue una distribución normal (a) y de otra que no la sigue (b).



(a) (b)

Figura 3.13: Comparación de características con una distribución normal N(0,1). En (a) la característica sigue esta distribución y en (b) no la sigue.

  - Cuando se determine que una característica sigue una distribución normal, el siguiente paso es **compararar la media asociada a la clase 1 y la media asociada a la clase 2**. Esta comparación se realiza con el *test t de Student (t-test)* con H_0 = "las medias son iguales para ambas clases". El test calcula un estadístico t; cuando H_0 sea cierta, t seguirá una distribución t de Student.
  - En el caso de que se determine que una característica no sigue una distribución normal, lo que **se comparan son las medianas de ambas clases**. Para ello se realiza el *test Wilcoxon rank sum* con H_0 = "las medianas son iguales para ambas clases". Este test determina si la diferencia entre las dos variables se debe o no al azar con el nivel de confianza establecido.

[1] Probabilidad correspondiente a la variable de ser posible bajo la hipótesis nula.

[2] 1-*trust*, siendo trust el nivel de significancia seleccionado. Indica la confianza con la que se rechaza o se acepta una hipótesis. En el caso de este trabajo, trust=0.9999

Cuando la hipótesis de igualdad de medias/varianzas sea aceptada, estas características son eliminadas, ya que no tienen capacidad de discriminación suficiente entre clases.

- **Análisis de la independencia entre pares de variables**. El objetivo es descartar aquellas características que no aportan información nueva para el aprendizaje de los modelos de clasificación. La independencia entre pares de variables se analiza por medio de la correlación, la cual puede ser positiva o negativa. En este caso concreto, se trata la correlación en valor absoluto, pues lo que interesa es que las variables no estén correlacionadas sea cual sea el signo de la tendencia. Se compara el coeficiente de correlación (R) con th_cor[3] y el p-valor asociado con α. Cuando: R>th_cor y p-valor<α, se considerará que la característica está muy correlacionada y será eliminada.

Tras esta etapa, la cantidad de características extraídas en el paso anterior se ha reducido a aquellas que realmente tienen el poder de discriminar las imágenes sanas de las de grado 3. Según la base de datos, esta cantidad es diferente debido al tamaño de las imágenes, ya que contienen una mayor o menor cantidad de información y la distribución y correlación entre las características es diferente:

- Para BBDD_512, son 37 las características seleccionadas.
- Para BBDD_1024, son 20 las características seleccionadas.
- Para BBDD_multi, son 33 las características seleccionadas para el entrenamiento.

3.2.5 Clasificación.

3.2.5.1. **Teoría de la clasificación automática de datos.**

Como se introdujo en la Sección 1.2, los alforitmos de clasificación se dividen en dos grupos en función de cómo extraen la información de los datos que usan para la clasificación: los supervisados y los no supervisados.

Debido a que se pretende medir la calidad del sistema de clasificación desarrollado en base al número de aciertos/fallos, a que se dispone de las etiquetas de los datos y a que se conoce a priori el número de clases en las que se quieren separar los datos, se ha decidido emplear un **clasificador supervisado**.

Existen numerosas familias de clasificadores supervisados:

- **Árboles de decisión**. Se basan en un método de aprendizaje inductivo[4] supervisado no paramétrico. Son estructuras resultantes de la partición recursiva del espacio de representación a partir de un conjunto de prototipos. Los árboles pueden ser entendidos como un conjunto de reglas; cada nodo del árbol es una carácterística, y cada rama representa un posible valor. Los algoritmos para la construcción de árboles de decisión suelen trabajar de manera *top-down*, escogiendo en cada paso la variable que mejor divide el conjunto de elementos. La condición de separación de los datos se elige de manera que maximice la medida de "pureza" del árbol seleccionada; esta pureza se mide con la varianza si los datos son numéricos y la entropía o el índice de Gini si los datos son categóricos.

[3] Umbral de correlación. Cuando se afirme que dos variables están correlacionadas, lo estarán al menos en un 90 % (valor seleccionado para este trabajo). Por tanto, th_cor=0.9.

[4] Partiendo de la observación y el análisis de una característica se llega a la formulación de una regla que la explique.

Lo más común es emplear el criterio de Gini:

$$GINI(t) = \sum_{k=1}^{J} p(k|t)(1 - p(k|t)) \tag{3.48}$$

Donde $p(k|t)$ es la probabilidad de que un caso asociado a un nodo t sea de la clase j.

En la Figura 3.14 se muestra un ejemplo de clasificación de datos utilizando este método.

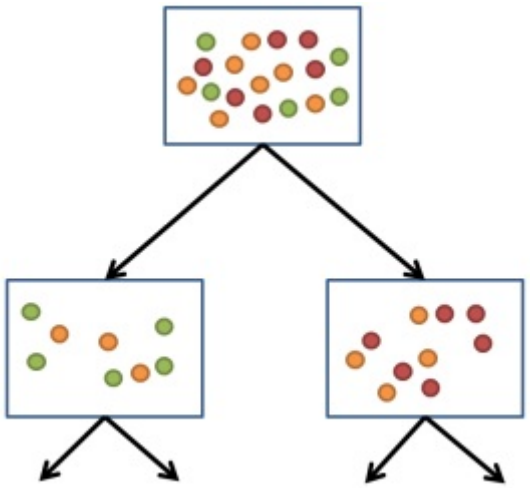


Figura 3.14: Ejemplo de una partición de los datos usando un árbol de decisión.

Dentro de los árboles de decisión aparecen subtipos en función de su complejidad. Esta complejidad viene marcada por el número de particiones (*splits*) que se le pide hacer al algoritmo.

Resultan muy útiles porque son robustos frente a *outliers*[5], aceptan datos tanto numéricos como categóricos, las reglas de decisión son fáciles de interpretar y, además, una vez generado el árbol la clasificación es muy rápida. Sin embargo, suelen tener algunos inconvenientes: tienden al *overfitting*[6] de los datos, su ratio de aciertos no suele ser de los mejores cuando se comparan con otros métodos y además el árbol cambia si cambian los datos de entrenamiento.

- **Análisis discriminante**. Se emplean *modelos generativos*, que calculan la probabilidad de que la muestra pertenezca a una u otra clase mediante el teorema de Bayes:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \tag{3.49}$$

$$\text{Donde: } p(x) = \sum_{c' \in C} p(x|c')p(c') \tag{3.50}$$

El análisis es distinto según los supuestos que se hagan sobre la probabilidad condicionada:

[5] Valores atípicos. Observaciones que son numéricamente distantes del resto de los datos.

[6] Sobreajuste: los parámetros del modelo acaban por representar perfectamente los datos de entrenamiento, pero han perdido la capacidad de generalización a la hora de clasificar nuevas muestras.

  - **Análisis discriminante lineal (LDA)**. Se asume: que los datos de cada clase siguen una distribución gaussiana y que las matrices de covarianza son idénticas para los datos de las distintas clases. Como consecuencia de estas asunciones, la frontera de clasificación resulta lineal.
  - **Análisis discriminante cuadrático (QDA)**. También asume que los datos siguen distribuciones gaussianas, pero no que las matrices de covarianza tengan que ser iguales. Por tanto, las fronteras de clasificación resultantes son cuadráticas. Obtiene mayor capacidad de ajuste que LDA, pero al tener más parámetros que ajustar es, por un lado, más complejo y, por otro, más probable que se de el *overfitting*.

- **Regresión logística**. Calcula la probabilidad de que la muestra pertenezca a una u otra clase de manera directa.

$$g(x; w) = \log\left(\frac{p(y=1|x)}{p(y=0)|x)}\right) \tag{3.51}$$

  Es un clasificador binario, solo discrimina entre dos clases.

  Como se pretende obtener una frontera lineal, se asume que la Ecuación 3.51 es una función lineal:

$$\log\left(\frac{p(y=1|x)}{p(y=0)|x)}\right) = a_0^{k,l} + \sum_{j=1}^{p} a_j^{k,l} x_j \tag{3.52}$$

  La ecuación de partida de este modelo es:

$$Pr(y=1|x) = \frac{\exp b_0 + \sum_{i=1}^{n} b_i x_i}{1 + \exp b_0 + \sum_{i=1}^{n} b_i x_i} \tag{3.53}$$

  donde: $Pr(y=1|x)$ es la probabilidad de que y tome el valor 1 en presencia de las covariables x, X es un conjunto de n covariables, b_0 es la constante del modelo (término independiente) y b_i son los coeficientes de las covariables.

  La Ecuación 3.53 es lo que se denomina *distribución logística*.

- **Máquina de vectores soporte (SVM)**. Dado un conjunto de puntos, su *envolvente convexa* es el conjunto convexo mínimo que los contiene (Figura 3.15 (a)).

  El *hiperplano de margen máximo* es aquel que proporciona la máxima separación entre dos clases linealmente separables. Los *vectores soporte* son las instancias más próximas de cada clase al hiperplano de margen máximo; al menos hay uno por clase, aunque posiblemente haya más. Un ejemplo de ambos se muestra en la Figura 3.15 (b). El conjunto de Vectores Soporte define de forma única el hiperplano de margen máximo, siendo el resto de instancias irrelevantes.

  La ecuación del hiperplano es:

$$x = b + \sum_{i,vectorsoporte} \alpha_i y_i a(i) a \tag{3.54}$$

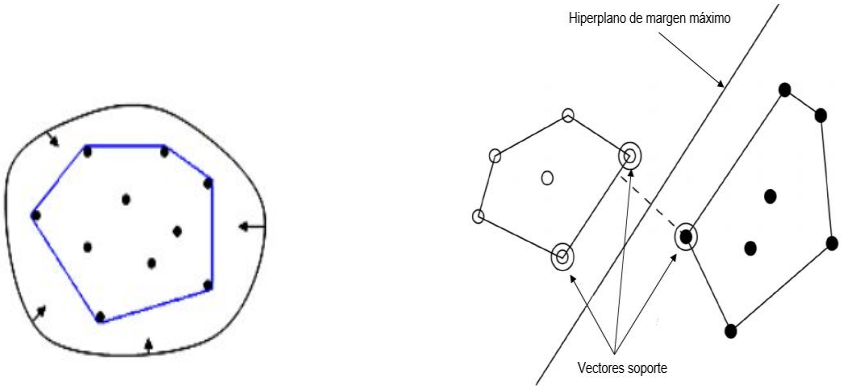


Figura 3.15: Elementos del algoritmo SVM: componente convexa (a) e hiperplano de margen máximo y vectores soporte (b).

siendo: b y α_i parámetros numéricos a determinar, a la instancia a clasificar, a(i) el vector soporte i-ésimo e y_i la clase de a(i).

Cuando las clases no son linealmente separables, se realiza una transformación no lineal del espacio de entrada, pero se mantienen estables los vectores soporte, por lo que no se realizaa de forma explícita la transformación:

$$x = b + \sum_{i,vectorsoporte} \alpha_i y_i (a(i)a)^n \tag{3.55}$$

Es similar a un producto de n factores, realizándose el producto escalar en el espacio original. En la Figura 3.16 aparece un ejemplo de una transformación de los datos de entrada.

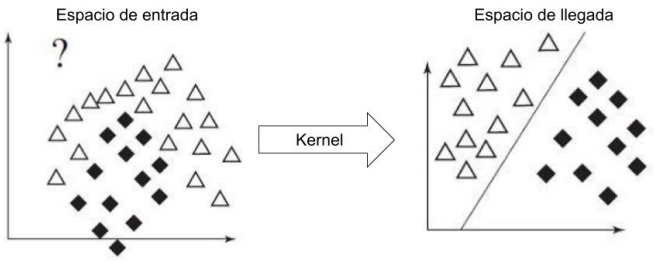
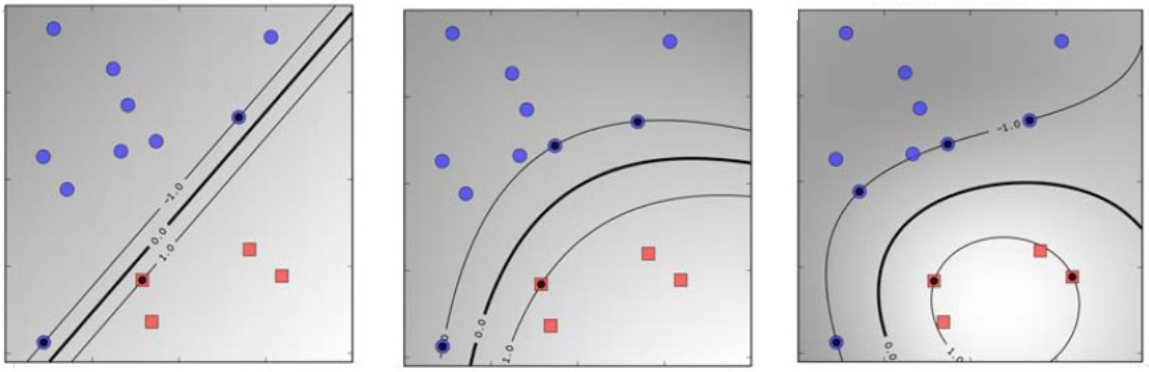


Figura 3.16: Transformación no lineal a través del kernel para conseguir que la división entre clases sea lineal.

Esta transformación se denomina *kernel*; en función del tipo de kernel se distinguen los diferentes subtipos dentro de esta familia. El kernel marca la forma del hiperplano (Figura 3.17).

Constituyen un clasificador resistente al sobreajuste y con una precisión generalmente alta; sin embargo, son costosos de entrenar y difíciles de interpretar, ya que las transformaciones matemáticas para conseguir que las clases sean linealmente separables son complejas.

- **K-vecinos más próximos (KNN)**. Es un modelo basado en memoria: almacena las observaciones para utilizarlas como prototipos. Además, es no paramétrico y basado en



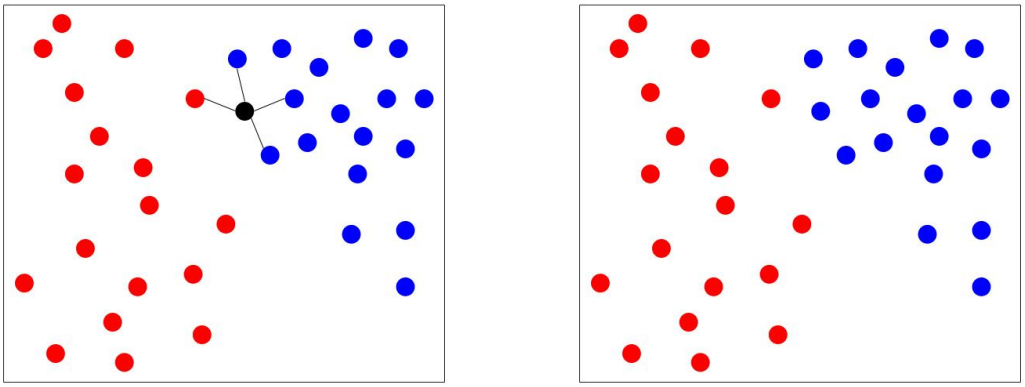
(a) (b) (c)

Figura 3.17: Ejemplos de hiperplanos de margen máximo: kernel lineal (a), kernel cuadrático (b) y kernel gaussiano con $\sigma = 1$ (c) [23].

distancias. Asume que el espacio muestral es un espacio métrico definido por {X,d}, donde X es el conjunto de puntos y d una distacia o métrica.

Lo que acaba por definir su funcionamiento es: el tipo de métrica empleado (d) y el número de vecinos (k) que se quieran considerar a la hora de asignar una clase a la nueva muestra. Aumentando el valor de k se reduce el error de clasificación siempre que se disponga de un número suficiente de muestras de entrenamiento.

Busca la clase de los vecinos más cercanos en función de d y le asigna la clase a la que pertenezcan la mayoría de los vecinos. En la Figura 3.18 hay un ejemplo del funcionamiento del método con $k = 4$.



(a) (b)

Figura 3.18: Ejemplo de clasificación de una nueva muestra (punto negro) mediante KNN con k=4 y 2 clases (azul y rojo). Conectados con el punto los vecinos que van a votar su clase (a). Clasificación en la clase azul por mayoría de votos (b).

Las fronteras que describe quedan determinadas por el conjunto de puntos, de forma que resultan lineales a trozos.

Algunas de las métricas que se pueden emplear son: distancia euclídea, distancia de Minkowski, distancia coseno, distancia cityblock o de Mahalanobis, entre otras.

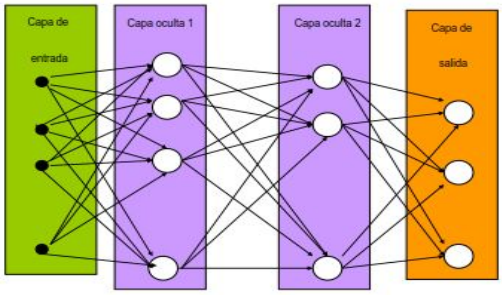


Figura 3.19: Ejemplo de la arquitectura de una red neuronal.

Además, existe una variante en la que la clase que se asigna a una nueva muestra no es la de la mayoría, sino que se le asignan pesos a cada voto en función de si están más o menos cerca del punto a clasificar. En este caso, el método se conoce como *weighted KNN*.

- **Redes neuronales artificiales (RNA)**. La teoría y los modelos de redes neuronales artificiales se han desarrollado tomando como modelo la estructura y el funcionamiento del sistema nervioso. Son sistemas de procesamiento de la información adaptativos, distribuidos y paralelos, que desarrollan su funcionalidad en respuesta a la información disponible para la red. Son tolerantes a fallos, ya que aprenden a reconocer patrones con ruido, distorsionados o incompletos y pueden realizar su función (con cierta degradación) aunque se destruya parte de la red (cierto grado de redundancia). Además, permiten descubrir interrelaciones que pueden permanecer ocultas cuando se dispone de un gran número de datos.

  Cada neurona de la red es una unidad de procesamiento de información: recibe información a través de las conexiones con neuronas de capas anteriores, procesa la información y emite el resultado a las neuronas de la capa siguiente siempre que este supere un valor umbral. Cuando acaba el proceso de entrenamiento de la red, a cada conexión entre neuronas se le ha asignado un determinado peso; cada neurona dará mas importancia a la información que le llegue por una conexión de peso mayor. En la Figura 3.19 se puede observar un ejemplo de una RNA.

- También se puede emplear lo que se conoce como **métodos combinados**. Estos utilizan varios de los algoritmos anteriores para obtener un rendimiento predictivo que mejore el que podría obtenerse por medio de cualquiera de los algoritmos individuales que lo constituyen.

  Existen formas diferentes de realizar esta combinación [24]:

  - **Bagging**. Crea combinaciones de modelos a partir de una familia inicial, disminuyendo la varianza y evitando el sobreajuste.

    Dado un conjunto de datos D de tamaño n, genera m nuevos conjuntos, D_i, de tamaño n', tomando al azar elementos de D de manera uniforme. A partir de estos m conjuntos se construyen m nuevos modelos de aprendizaje; la respuesta final del algoritmo será la votación de las m respuestas.

  - **Boosting**. Es un método iterativo que manipula los pesos de los datos para generar modelos distintos.

La idea es que en cada iteración se incremente el peso de los objetos mal clasificados por el predictor en esa iteración, por lo que en la construcción del próximo predictor estos objetos serán más importantes y será más probable clasificarlos bien.

Dentro de este método, destaca el algoritmo *AdaBoost*, que va modificando iterativamente los pesos en función del error que ha cometido el modelo en el conjunto de entrenamiento hasta las iteraciones fijadas. El modelo final se consigue por votación ponderada de cada modelo usando los pesos de todos los modelos.

RUSBoost es una modificación de AdaBoost que está pensado para datos de entrenamiento desequilibrados entre clases. Remuestrea aleatoriamente la clase mayoritaria en cada iteración para crear un subset balanceado antes de entrenar los clasificadores [25].

- **Subespacios aleatorios**. En este método, cada modelo se entrena considerando un subconjunto de los atributos, pero teniendo en cuenta todas las observaciones.

  El parámetro que marca el funcionamiento del método es el tamaño N de estos subconjuntos. D es el número de atributos de los datos y L el número de clasificadores individuales del conjunto. Para cada clasificador individual l se toma un número reducido de atributos d_l para entrenarlo.

  Para clasificar un nuevo objeto, se combinan las salidas de los L clasificadores individuales por medio del voto de la mayoría o por combinación ponderada.

3.2.5.2. Selección y entrenamiento del clasificador.

Una vez se han extraído y seleccionado las características que van a aportar la información necesaria para realizar una buena clasificación final y se ha hecho un estudio de las diferentes familias de clasificadores supervisados, se va a proceder detallar algunos detalles prácticos de la etapa de clasificación.

Para llevar a cabo el entrenamiento de los modelos predictivos, se ha utilizado la App de MATLAB *"Classification Learner"*, que permite evaluar el comportamiento de diferentes tipos de clasificadores supervisados en función de los datos de entrada que se le proporcionen. El objetivo del uso de esta plataforma es generar el mejor clasificador para los datos de entrada concretos de los que se dispone.

El primer paso es elegir un **método de validación** con el objetivo de estimar el rendimiento frente a nuevos datos comparados con los empleados para el entrenamiento y así poder seleccionar cuál de todos los modelos es el mejor y evitar el overfitting. La herramienta Classification Learner ofrece dos opciones a la hora de escoger el método de validación que se pasan a detallar a continuación:

- **Holdout validation**. Selecciona un porcentaje de los datos para usarlos como test. La aplicación entrena los modelos con los datos que no han sido seleccionados y evalúa su rendimiento con el porcentaje selccionado. El modelo resultante, por tanto, solamente se ha entrenado con una porción de los datos.
- **Cross-Validation**. La totalidad de las muestras o instancias se dividen en k particiones o conjuntos disjuntos; para cada uno de estos conjuntos, se entrena un modelo predictivo diferente haciendo uso de las muestras pertenecientes a k-1 particiones. El modelo predictivo

en cuestión se evalúa con las muestras de la partición restante. Se puede observar esta división en diferentes conjuntos en la Figura 3.20. Calcula la media del error cometido con los datos utilizados como test de todos los conjuntos. Por tanto, da una buena estimación de la precisión predictiva que vayan a tener los modelos cuando se empleen todos los datos. Esta es la validación que se ha empleado para este trabajo, con $k = 5$ divisiones de los datos.

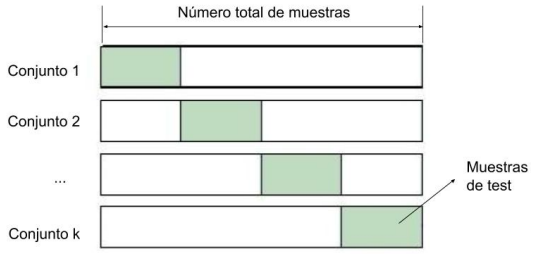


Figura 3.20: Ejemplo de división del conjunto de datos para realizar cross-validation.

Una vez seleccionado el tipo de validación y proporcionados los datos y sus etiquetas, se pueden seleccionar qué clasificadores se quieren evaluar. En el presente trabajo se ha decidido compararlos todos los que ofrece la aplicación para seleccionar finalmente el que mejor rendimiento ofrezca:

- Árboles de decisión. Para todos ellos, la medida de la pureza del árbol se estima con el índice de Gini.
  - Complex tree: splits=100.
  - Medium tree: splits=20.
  - Simple tree: splits=4.
- Análisis discriminante: linear o quadratic.
- Regresión logística.
- SVM. Según el kernel se distingue:
  - Linear.
  - Quadratic.
  - Cubic.
  - Gaussian. Según el tamaño: fine ($k = 1{,}5$), medium ($k = 6{,}1$) o coarse ($k = 24$).
- KNN:
  - Fine: $k = 1$ y distancia euclídea.
  - Medium: $k = 10$ y distancia euclídea.
  - Coarse: $k = 100$ y distancia euclídea.
  - Cosine: $k = 10$ y distancia cosine.
  - Cubic: $k = 10$ y distancia de Minkowski cúbica.

- Weighted: $k = 10$, distancia euclídea y pesos=cuadrado inverso de la distancia.

- Métodos combinados:

  - Boosted trees: emplea AdaBoost para crear un método combinado basado en árboles de decisión. Concretamente: 30 clasificadores, cada uno con splits=20.
  - RUSBoosted trees: emplea RUSBoost para crear un método combinado basado en árboles de decisión. Concretamente: 30 clasificadores, cada uno con splits=20.
  - Bagged trees: emplea Bagging como método de combinación de árboles de decisión. Concretamente, 30 clasificadores.
  - Subspace discriminant: como método de combinación emplea subespacios aleatorios (concretamente $N = 19$) y 30 clasificadores que trabajan con anlaísis discriminante.
  - Subspace KNN: como método de combinación emplea subespacios aleatorios (concretamente $N = 19$) y 30 clasificadores KNN.

3.2.5.3. **Evaluación de modelos predictivos.**

Cuando todos los modelos han sido entrenados, es preciso evaluar su funcionamiento de manera que se seleccione finalmente el que mejores resultados haya obtenido con los datos de entrenamiento. Para realizar esta evaluación, la aplicación obtiene diferentes medidas de la calidad de cada algoritmo de clasificación.

- De manera gráfica, proporciona:

  - **Matriz de confusión**. Las columnas representan las clases verdaderas y las filas las predicciones del clasificador. A través de estas matrices se conoce el número de verdaderos negativos (posición (1,1)), falsos positivos (posición (1,2)), falsos negativos (posición (2,1)) y verdaderos positivos (posición (2,2)).
  - **Curva ROC**.Es un gráfico en el que se sitúa el ratio de verdaderos positivos en el eje de ordenadas y el ratio de falsos positivos en el eje de abcisas. A través de esta curva se puede calcular el parámetro AUC (Area Under Curve), que es una medida de la calidad general del clasificador. Idealmente, el ratio de verdaderos positivos sería 1 y el de falsos positivos 0, por lo que la curva describiría un ángulo de 90º y, por tanto, $AUC = 1$; cuanto más cercano a 1 sea este valor, mejor será el clasificador.

- De manera numérica, la herramienta proporciona el **accuracy** (precisión). Este valor es el que se prima a la hora de seleccionar el mejor clasificador: será aquel que tenga una mayor precisión. Es el número de muestras correctamente clasificadas dividido entre el número total de muestras:

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.56}$$

Como ya se ha mencionado, analizando los datos anteriores para cada algoritmo de clasificación se selecciona aquel que haya resultado mejor. Este será el clasificador final que vaya a emplearse para la clasificación, por lo que se almacena el modelo y se entrena con los datos de train.

El último paso de este trabajo es comprobar cuál es el rendimiento del algoritmo desarrollado para la clasificación muestras nuevas (que no hayan sido empleadas para el entrenamiento del clasificador). Estas muestras nuevas son las muestras que se habían reservado como test. Es preciso preparar estos datos para que tengan la misma estructura que los datos empleados para el entrenamiento del modelo. Por tanto, se normaliza la matriz mediante el Zscore. En este caso, la media y desviación típica que se emplean son las de los datos de train, ya que se van a clasificar como si fueran nuevas muestras independientes.Además de estar normalizadas, es preciso eliminar de la matriz aquellas características que han sido rechazadas tras el análisis estadístico, ya que el clasificador no ha sido entrenado con ellas.

Por último, estas características son proporcionadas al clasificador entrenado sin indicarle sus etiquetas. Cuando el algoritmo realiza la clasificación, se obtiene la clase que ha predicho para cada muestra. Sobre estos resultados de clasificación se lleva a cabo una cuantificación de la bondad del modelo predictivo empleando diferentes figuras de mérito comparando las etiquetas predichas con el groundtruth que se tenía almacenado de las observaciones de test. Estas figuras de mérito pueden ser:

- **Sensibilidad**. Indica la capacidad del clasificador para detectar la presencia de la patología:

$$Sensibilidad = \frac{VP}{VP + FN} \tag{3.57}$$

- **Especificidad**. Indica la capacidad del clasificador para no detectar la presencia de la patología cuando no está presente:

$$Especificidad = \frac{VN}{VN + FP} \tag{3.58}$$

- **Matriz de confusión**.
- **Ratio de acierto**. Número de muestras correctamente clasificadas dividido entre el número total de muestras clasificadas.
- **Ratio de error**. Número de muestras clasificadas incorrectamente dividido entre el número total de muestras clasificadas.
- **Ratio de verdaderos positivos**. Número de muestras patológicas clasificadas correctamente entre el número total de muestras patológicas.
- **Ratio de falsos positivos**. Número de muestras patológicas clasificadas incorrectamente entre el número total de muestras patológicas.

Nomenclatura:

- Verdadero positivo (VP). El clasificador indica que la muestra es patológica cuando verdaderamente lo es.
- Verdadero negativo (VN). El clasificador indica que la muestra es sana cuando verdaderemente lo es.
- Falso positivo (FP). El clasificador indica que la muestra es patológica cuando no lo es en realidad.

- Falso negativo (FN). El clasificador indica que la muestra es sana cuando no lo es en realidad.

Capítulo 4

Resultados

En este capítulo se van a mostrar y discutir los resultados obtenidos en las diferentes etapas del método propuesto .

4.1 Resultados de la selección de características.

Como ya se ha explicado en la Sección 3.2.4, el proceso de selección de características sirve para descartar de entre todas las características extraídas de las imágenes aquellas que no aportan la suficiente información para el proceso de clasificación.

Este proceso consiste en un análisis estadístico que está dividido en dos pasos. En el primero se estudia la capacidad discriminatoria de las características atendiendo a sus medias/medianas (según sigan una distribución normal o no); en el segundo, se estudia la correlación entre características para evitar que haya varias que aporten la misma información y, por tanto, sean redundantes.

4.1.1 Características extraídas de BBDD_512.

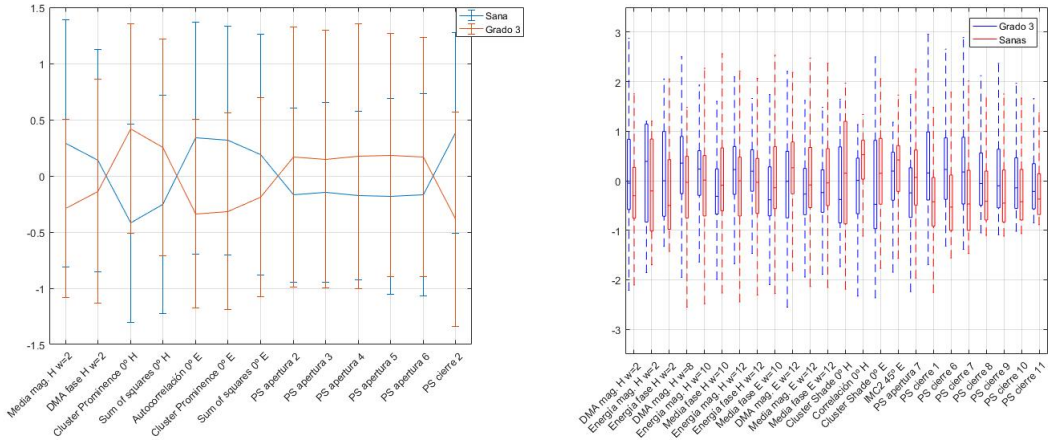
De las 196 características extraídas para cada imagen de esta base de datos, han sido seleccionadas 37 tras el análisis estadístico.

Estudiando la capacidad de discriminación de las características, 59 resultan eliminadas. De estas, 26 siguen una distribución normal y, por tanto, son sus medias lo que se compara. Las otras 33 características se eliminan por la comparación entre sus medianas.

En el siguiente paso, que se corresponde con el cálculo de la correlación dos a dos entre características, se estudian aquellas que no hayan sido eliminadas en el paso anterior. Tras el estudio de la correlación, son 100 las características cuyos coeficientes de correlación superan el umbral seleccionado (th_cor =0.9), por lo que son descartadas.

En la Figura 4.1 (a) se muestran las medias de las características seleccionadas que siguen una distribución normal para poder comprobar visualmente su capacidad de discriminación. En la

Figura 4.1 (b) se puede observar la comparación de medianas de las características seleccionadas que no siguen una distribución normal; se aprecia que efectivamente sirven para diferenciar ambas clases. Por último, en la Figura 4.2 se ha calculado la correlación de las características seleccionadas para comprobar que sus coeficientes de correlación son bajos y, por tanto, cada una de ellas aporta información nueva al clasificador.



(a) (b)

Figura 4.1: Comparación de medias (a) y medianas (b) de cada clase de las características seleccionadas para BBDD_512.

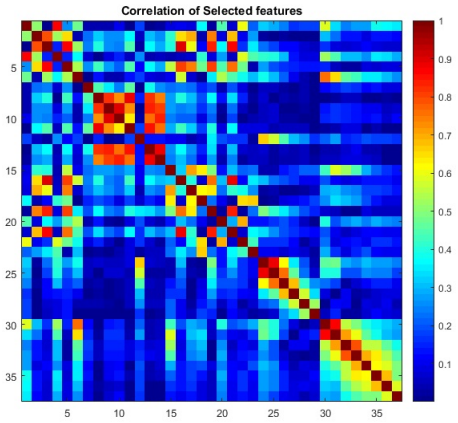


Figura 4.2: Correlación por pares de las características seleccionadas para BBDD_512.

Las características finalmente seleccionadas se encuentran recogidas en la Tabla 4.1.

4.1.2 Características extraídas de BBDD_1024.

El análisis estadístico de las características extraídas de esta base de datos concluye que solo 20 de ellas aportan la información necesaria para ser tenidas en cuenta en la etapa de clasificación posterior.

Concretamente, tras el estudio de la capacidad de discriminación de las características se eliminan 94 de ellas. Dentro de este grupo, 76 siguen una distribución normal y, por tanto, se han

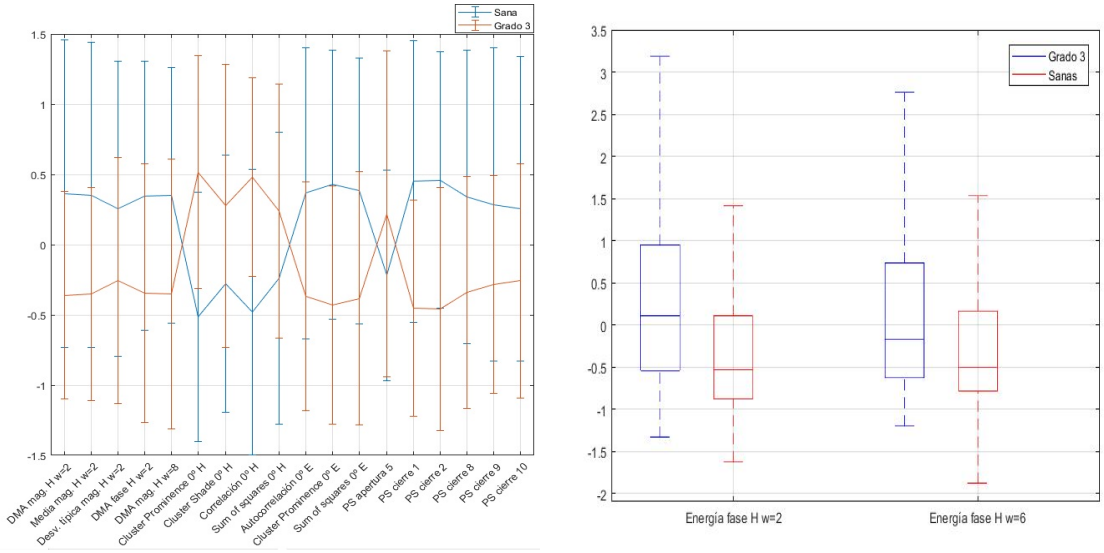
Filtros de Gabor	Matrices de coocurrencia	Granulometría
DMA de la magnitud para las imágenes de Hematoxilina con w=2	Cluster Shade 0º H	Apertura EE=8
Energía de la magnitud para las imágenes de Hematoxilina con w=2	Correlación 0º H	Apertura EE=12
Energía de la fase para las imágenes de Hematoxilina con w=2	Cluster Prominence 0º H	Apertura EE=16
Media de la magnitud para las imágenes de Hematoxilina con w=2	Sum of squares 0º H	Apertura EE=20
DMA de la fase para las imágenes de Hematoxilina con w=2	Cluster Shade 0º E	Apertura EE=24
DMA de la magnitud para las imágenes de Hematoxilina con w=8	Sum of squares 0º E	Apertura EE=28
Energía de la magnitud para las imágenes de Hematoxilina con w=10	Autocorrelación 0º E	Cierre EE=2
Media de la fase para las imágenes de Hematoxilina con w=10	Cluster Prominence 0º E	Cierre EE=4
Media de la fase para las imágenes de Eosina con w=10	IMC2 45º E	Cierre EE=12
Energía de la magnitud para las imágenes de Hematoxilina con w=12		Cierre EE=14
Energía de la fase para las imágenes de Hematoxilina con w=12		Cierre EE=16
DMA de la magnitud para las imágenes de Eosina con w=12		Cierre EE=18
Media de la magnitud para las imágenes de Eosina con w=12		Cierre EE=20
Media de la fase para las imágenes de Eosina con w=12		Cierre EE=22

Tabla 4.1: Características seleccionadas para BBDD_512.

comparado sus medias a la hora de medir su capacidad discriminatoria; el resto de características (18), se eliminan tras la comparación de sus medianas.

En el último paso (el cálculo de la correlación por pares), se estudian aquellas que no hayan sido eliminadas en el paso anterior. Al estudiarse la correlación se determina que 82 de estas características superan el umbral de correlación establecido y son eliminadas.

En la Figura 4.3 (a) se ilustra la comapración entre las medias de cada clase de las 18 características seleccionadas que siguen una distribución normal. En la Figura 4.3 (b) se muestra la comparación de las medianas de las 2 características seleccionadas que no siguen una distribución normal. Se puede apreciar visualmente en ambas imágenes que tanto las medias como las medianas de cada clase se encuentran separadas, por lo que sirven para discriminar entre las dos clases posibles.



(a) (b)

Figura 4.3: Comparación de medias (a) y medianas (b) de cada clase de las características seleccionadas para BBDD_1024.

Además, en la Figura 4.4 se muestra el coeficiente de correlación por pares de todas las características seleccionadas para comprobar que sean bajos y, por tanto, que cada una de ellas aporta información nueva al clasificador.

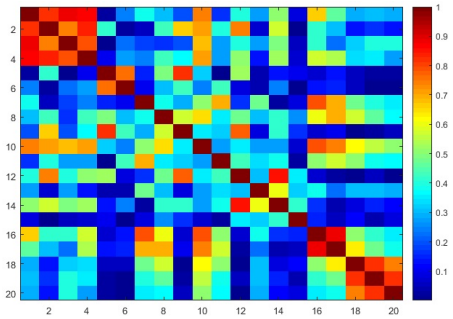


Figura 4.4: Correlación por pares de las características seleccionadas para BBDD_1024.

Las características que se han seleccionado finalmente aparecen en la Tabla 4.2.

4.1.3 Características extraídas del experimento multiresolución.

En este caso, han resultado relevantes para la clasificación 33 características.

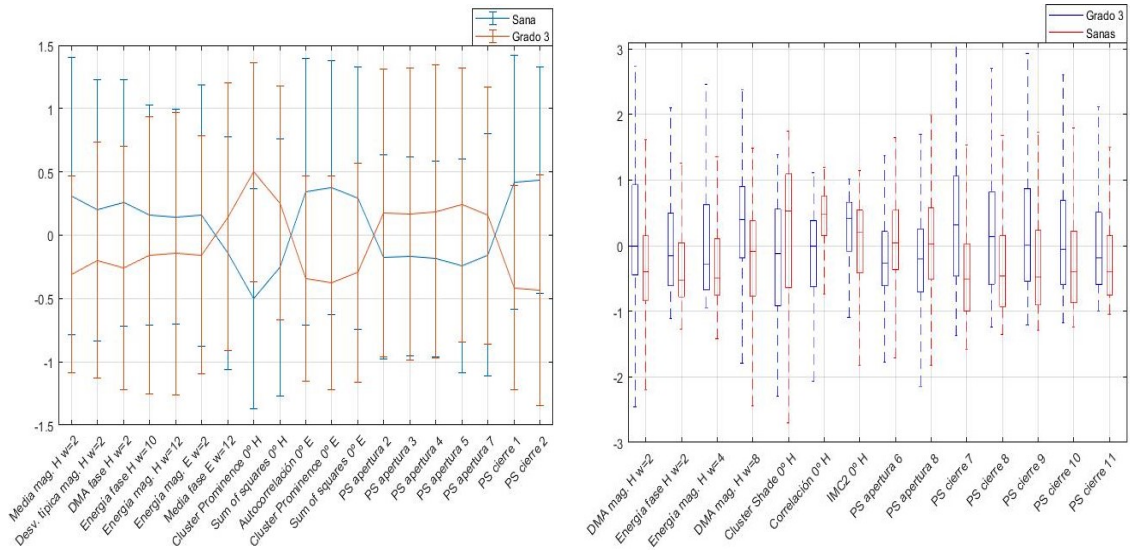
Por un lado, atendiendo a la capacidad de discriminación entre clases, son 67 las características que no se consideran lo suficiente discriminativas y, por tanto, son eliminadas en este paso. Para ello, se han comparado las medias de las clases de 46 características, ya que seguían una distribución normal, y las medianas de 21, ya que no seguían esta distribución.

Por otro lado, para las características que no han sido descartadas en el paso anterior se calcula la correlación por pares. En esta etapa, 96 características han resultado estar correlacionadas y, por tanto, se han descartado.

Filtros de Gabor	Matrices de coocurrencia	Granulometría
DMA de la magnitud para las imágenes de Hematoxilina con w=2	Cluster prominence 0º H	Apertura EE=20
Media de la magnitud para las imágenes de Hematoxilina con w=2	Cluster shade 0º H	Cierre EE=2
Desviación típica de la magnitud para las imágenes de Hematoxilina con w=2	Correlación 0º H	Cierre EE=4
Energía de la fase para las imágenes de Hematoxilina con w=2	Sum of squares 0º H	Cierre EE=16
Energía de la fase para las imágenes de Hematoxilina con w=6	Autocorrelación 0º E	Cierre EE=18
DMA de la fase para las imágenes de Hematoxilina con w=8	Cluster prominence 0º E	Cierre EE=20
	Sum of squares 0º E	

Tabla 4.2: Características seleccionadas para BBDD_1024.

En la Figura 4.5 (a) se pueden observar las medias de las 19 características seleccionadas que siguen una distribución normal para poder comprobar visualmente su capacidad de discriminación. En la Figura 4.5 (b) aparece la comparación de medianas de las 14 características seleccionadas que no siguen una distribución normal; se aprecia que efectivamente sirven para diferenciar ambas clases, ya que no se solapan los valores.



(a) (b)

Figura 4.5: Comparación de medias (a) y medianas (b) de cada clase de las características seleccionadas en el experimento multiresolución.

En la Figura 4.6 se ha calculado la correlación de las características seleccionadas para comprobar que sus coeficientes de correlación son bajos y, por tanto, cada una de ellas aporta información nueva al clasificador.

Las características finalmente seleccionadas se encuentran recogidas en la Tabla 4.3.

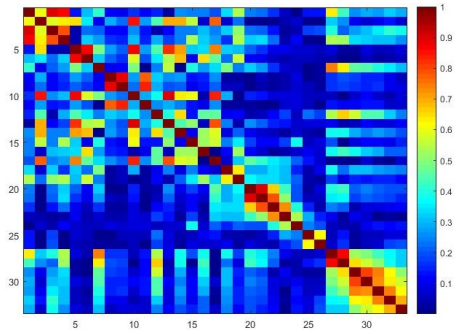


Figura 4.6: Correlación por pares de las características seleccionadas en el experimento multiresolución.

4.2 Resultados de la clasificación.

En esta sección se van a exponer los resultados obtenidos en la etapa de clasificación. Concretamente, qué clasificadores han sido seleccionados de entre todos los entrenados para cada base de datos y sus rendimientos a la hora de clasificar nuevas muestras, en este caso las características del conjunto de test.

4.2.1 Selección del clasificador.

Como se ha mencionado en la Sección 3.2.5.2, se ha empleado la herramienta Classification Learner de MATLAB para elegir el algoritmo de clasificación que es capaz de separar mejor los datos de entrada atendiendo a su clase.

El proceso de selección del clasificador consiste en entrenar con los datos de *training* los 23 clasificadores disponibles, calcular distintas métricas de la calidad de su funcionamiento (*accuracy*, curva ROC, AUC...) y, por último, quedarse con el clasificador que obtenga mejores resultados referidos a los parámetros anteriores.

En la Tabla 4.18 se muestra la precisión de cada clasificador para poder discutir este resultado.

- Los mejores 4 clasificadores para las características extraídas de las imágenes de tamaño 512×512 píxeles son, por orden descendente: cubic SVM (84.1 %), medium gaussian SVM (81.9 %), quadratic SVM (81.6 %) y boosted trees (81.2 %).

  En las Tablas 4.5, 4.6, 4.7 y 4.8 se muestran las matrices de confusión de cada uno de estos clasificadores. Comparando estos valores para los diferentes clasificadores, se aprecia que efectivamente el número de muestras bien clasificadas (VN+VP) va decreciendo conforme lo hace la calidad del clasificador para estos datos concretos.

  Como ya se ha indicado, otra medida de la calidad de la clasificación es la curva ROC. En la Figura 4.7 se muestran las curvas de los cuatro clasificadores indicados, indicando sus ratios de verdaderos y falsos positivos y el valor AUC obtenido. Comparando los valores de AUC, se aprecia que tanto para cubic SVM como para quadratic SVM se ha obtenido 0.91; sin embargo, analizando los valores de los ratios se ve que el clasificador quadratic SVM comete más errores (el ratio de verdaderos positivos es más bajo y el de falsos positivos más alto que los obtenidos por cubic SVM). Frente al resto de clasificadores, cubic SVM obtiene un AUC mayor.

Filtros de Gabor	Matrices de coocurrencia	Granulometría
DMA de la magnitud para las imágenes de Hematoxilina con w=2	Cluster prominence 0º H	Apertura EE=8
Media de la magnitud para las imágenes de Hematoxilina con w=2	Cluster shade 0º H	Apertura EE=12
Desviación típica de la magnitud para las imágenes de Hematoxilina con w=2	Correlación 0º H	Apertura EE=16
DMA de la fase para las imágenes de Hematoxilina con w=2	IMC2 0º H	Apertura EE=20
Energía de la fase para las imágenes de Hematoxilina con w=2	Sum of squares 0º H	Apertura EE=24
Energía de la magnitud para las imágenes de Eosina con w=2	Sum of squares 0º E	Apertura EE=28
Energía de la magnitud para las imágenes de Hematoxilina con w=4	Autocorrelación 0º E	Apertura EE=32
DMA de la magnitud para las imágenes de Hematoxilina con w=8	Cluster prominence 0º E	Cierre EE=2
Energía de la fase para las imágenes de Hematoxilina con w=10		Cierre EE=4
Energía de la magnitud para las imágenes de Hematoxilina con w=12		Cierre EE=14
Media de la fase para las imágenes de Eosina con w=12		Cierre EE=16
		Cierre EE=18
		Cierre EE=20
		Cierre EE=22

Tabla 4.3: Características seleccionadas para el experimento multiresolución.

Por tanto, tras comparar los resultados obtenidos, el clasificador seleccionado para esta base de datos es **cubic SVM**.

- Analizando ahora los resultados para las características extraídas de las imágenes de tamaño 1024×1024 píxeles, se observa que los 4 mejores clasificadores son, por orden descendente: quadratic SVM (86.3 %), subspace KNN (86.2 %), cubic SVM (83.8 %) y medium gaussian SVM (81.6 %).

  En las Tablas 4.9, 4.10, 4.11 y 4.12 se recogen las matrices de confusión de cada uno de estos clasificadores. Comparando sus valores, para los dos mejores clasificadores el número de muestras mal clasificadas es el mismo, pero su distribución no: quadratic SVM ha cometido el mismo error para muestras sanas y para patológicas y subspace KNN ha cometido más FN que FP. Los otros dos clasificadores han cometido más errores en la clasificación.

  En cuanto al valor AUC extraído de las curvas ROC de los clasificadores, mostradas en la Figura 4.8, se ve que quadratic SVM y cubic SVM han obtenido el valor más alto: 0.93; sin embargo, comparando los ratios de verdaderos y falsos positivos se aprecia que

Clasificador	BBDD_512	BBDD_1024	Multiresolución
Complex tree	73.8 %	75.0 %	74.4 %
Medium tree	74.1 %	75.5 %	74.8 %
Simple tree	70.8 %	72.5 %	73.0 %
Linear discriminant	79.9 %	80.2 %	82.2 %
Quadratic discriminant	80.3 %	79.4 %	80.1 %
Logistic regression	80.1 %	79.1 %	82.1 %
Linear SVM	80.1 %	80.2 %	81.4 %
Quadratic SVM	81.6 %	**86.3 %**	84.8 %
Cubic SVM	**84.1 %**	83.8 %	**85.6 %**
Fine gaussian SVM	70.1 %	78.8 %	70.0 %
Medium gaussian SVM	81.9 %	81.6 %	83.3 %
Coarse gaussian SVM	79.4 %	79.4 %	81.5 %
Fine KNN	76.9 %	79.7 %	79.3 %
Medium KNN	79.4 %	78.3 %	81.6 %
Coarse KNN	76.9 %	78.3 %	77.0 %
Cosine KNN	80.3 %	77.5 %	80.7 %
Cubic KNN	79.0 %	77.7 %	80.8 %
Weighted KNN	80.8 %	81.5 %	82.3 %
Boosted trees	81.2 %	78.6 %	81.2 %
Bagged trees	80.9 %	79.1 %	82.3 %
Subspace discriminant	79.1 %	80.5 %	80.8 %
Subspace KNN	80.8 %	86.2 %	82.5 %
RUSBoosted trees	74.9 %	76.4 %	74.8 %

Tabla 4.4: Precisión de los clasificadores para los datos de train.

quadratic SVM ha obtenido resultados mucho mejores (ratio de falsos positivos más bajo y de verdaderos positivos más alto). Comparando con el resto de clasificadores, el AUC de quadratic SVM es mayor.

Por todo lo mencionado, para la base de datos de 1024×1024, el clasificador seleccionado es **quadratic SVM**.

- Los mejores 4 clasificadores para las características extraídas de las imágenes tenidas en cuenta para el experimento multiresolución son, por orden descendente: cubic SVM (85.6 %), quadratic SVM (84.8 %), medium gaussian SVM (83.3 %) y subspace KNN (82.5 %).

  En las Tablas 4.13, 4.14, 4.15 y 4.16 aparecen las matrices de confusión de los clasificadores anteriores. Comparando sus valores, se pone en evidencia que efectivamente el número de muestras bien clasificadas (VN+VP) va decreciendo conforme lo hace la calidad del clasificador.

  En cuanto a la información que aportan las curvas ROC, mostradas en la Figura 4.9: el valor de AUC más alto lo ha obtenido el clasificador cubic SVM.

	Cubic SVM		
True class	0	506	107
	1	88	525
		0	1
		Predicted class	

Tabla 4.5: Matriz de confusión del clasificador Cubic SVM para la BBDD_512.

	Medium Gaussian SVM		
True class	0	479	134
	1	88	525
		0	1
		Predicted class	

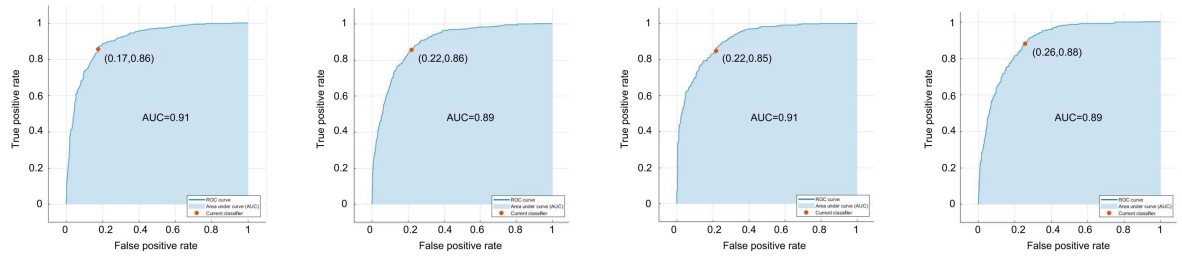
Tabla 4.6: Matriz de confusión del clasificador Medium Gaussian SVM para la BBDD_512.

	Quadratic SVM		
True class	0	480	133
	1	93	520
		0	1
		Predicted class	

Tabla 4.7: Matriz de confusión del clasificador Quadratic SVM para la BBDD_512.

	Boosted trees		
True class	0	456	157
	1	73	540
		0	1
		Predicted class	

Tabla 4.8: Matriz de confusión del clasificador Boosted trees para la BBDD_512.



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

Figura 4.7: Curvas ROC de los cuatro mejores clasificadores de BBDD_512: cubic SVM (a), median gaussian SVM (b), quadractic SVM (c) y boosted trees (d).

Es por todo esto por lo que el clasificador finalmente elegido para realizar la clasificación de esta base de datos es **Cubic SVM**.

Para las tres bases de datos se han seleccionado clasificadores de la familia SVM. Este resultado era de esperar debido a que de entre todas las familias propuestas, estos son los clasificadores que mejor manejan la alta dimensionalidad de los datos.

4.2.2 Validación del modelo.

Una vez seleccionados los clasificadores que mejor agrupan los datos de entrenamiento según la base de datos, se valida el entrenamieto con esos mismos datos empleando *cross validation* con $k = 5$. Los indicadores que se han empleado para medir la calidad de la fase de entrenamiento y sus valores están recogidos en la Tabla 4.17.

Resultan valores similares a los obtenidos en el paso anterior; el resto de parámetros sirven para reforzar la decisión tomada. En todos los casos son resultados de calidad bastante elevada, por lo que la validación de los modelos ha resultado positiva y son estos los que se van a emplear para predecir nuevas muestras.

	Quadratic SVM		
True class	0	157	25
	1	25	157
		0	1
		Predicted class	

Tabla 4.9: Matriz de confusión del clasificador Quadractic SVM para BBDD_1024.

	Subspace KNN		
True class	0	158	24
	1	26	156
		0	1
		Predicted class	

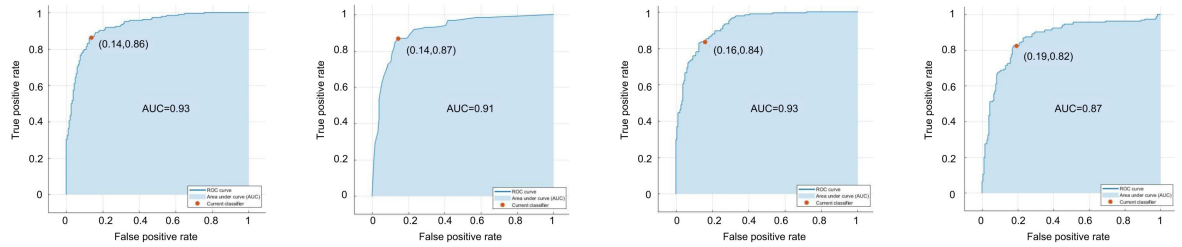
Tabla 4.10: Matriz de confusión del clasificador Subspace KNN para BBDD_1024.

	Cubic SVM		
True class	0	152	30
	1	29	153
		0	1
		Predicted class	

Tabla 4.11: Matriz de confusión del clasificador Cubic SVM para BBDD_1024.

	Medium gaussian SVM		
True class	0	150	32
	1	35	147
		0	1
		Predicted class	

Tabla 4.12: Matriz de confusión del clasificador Medium Gaussian SVM para BBDD_1024.



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

Figura 4.8: Curvas ROC de los cuatro mejores clasificadores de BBDD_1024: quadratic SVM (a), subspace KNN (b), cubic SVM (c) y medium gaussian SVM (d).

4.2.3 Predicción de etiquetas.

Una vez seleccionados los clasificadores que generan los modelos de predicción óptimos, se realiza la clasificación en sí con los datos reservados como test. Los resultados para cada una de las bases de datos están recogidos en la Tabla 4.18. Estos datos resumen la calidad del proceso completo, ya que una vez que se ha desarrollado todo el algoritmo se lleva a cabo este último paso para comprobar cómo funcionan los modelos en la tarea de predecir muestras completamente nuevas.

En vista de estos resultados, se puede afirmar que se ha conseguido evitar el sobreajuste de los clasificadores con los datos de entrenamiento, ya que los resultados obtenidos con nuevas muestras se asemejan a los conseguidos con los datos de train. Además, en todos los casos los valores de sensibilidad y especificidad son relativamente elevados; cabe añadir que estos valores son bastante similares, por lo que los clasificadores actúan de manera homogénea, sin diferenciarse la calidad en función de la clase.

Comparando todos los parámetros entre los clasificadores se concluye que los mejores resultados han sido obtenidos con la base de datos de tamaño 1024×1024 píxeles. Este resultado no era el esperado a priori, ya que es la base de datos con menor número de muestras y, en general, cuantas

	Cubic SVM		
True class	0	307	58
	1	47	318
		0	1
		Predicted class	

Tabla 4.13: Matriz de confusión del clasificador Cubic SVM para el experimento multiresolución.

	Quadratic SVM		
True class	0	298	67
	1	44	321
		0	1
		Predicted class	

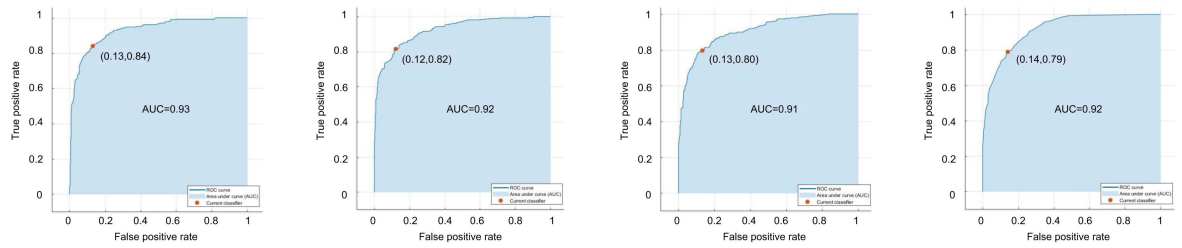
Tabla 4.14: Matriz de confusión del clasificador Quadratic SVM para el experimento multiresolución.

	Medium gaussian SVM		
True class	0	291	74
	1	48	317
		0	1
		Predicted class	

Tabla 4.15: Matriz de confusión del clasificador Medium gaussian SVM para el experimento multiresolución.

	Subspace KNN		
True class	0	288	77
	1	51	314
		0	1
		Predicted class	

Tabla 4.16: Matriz de confusión del clasificador Subspace KNN para el experimento multiresolución.



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

Figura 4.9: Curvas ROC de los cuatro mejores clasificadores para los datos del experimento multiresolución: cubic SVM (a), quadratic KNN (b), medium gaussian SVM (c) y subspace KNN (d).

más muestras se le puedan facilitar a un clasificador mejores serán sus resultados. Sin embargo, se ha considerado que estas imágenes, al ser de mayor tamaño, permiten obtener información más global del tejido; esto, sumado al tipo de descriptores de imagen que se han empleado, se traduce en que la información extraída sea de mayor utilidad. Por ese mismo motivo, el número de características que han resultado relevantes es mucho menor que en los otros casos, lo que permite emplear el clasificador más sencillo de los tres.

En la tabla 4.19 se muestran los resultados obtenidos en los distintos estudios del estado del arte. Es importante remarcar que estos resultados no son directamente comparables entre sí ni con el método desarrollado en el presente trabajo, ya que, al no existir bases de datos públicas, cada experimento se ha realizado con imágenes diferentes. Además hay que destacar que en el presente TFG el objetivo es dar solución al problema sano/grado 3 mientras que en los trabajos del estado del arte con los que se puede llevar a cabo la comparativa tienen como objetivo distinguir entre sano/patológico [11] o entre los distintos grados de la escala Gleason [12].

	BBDD_512	BBDD_1024	Multiresolución
Accuracy	84.26 %	81.87 %	83.15 %
Sensibilidad	0.8238	0.8406	0.8328
Especificidad	0.8613	0.7967	0.8301
Ratio de error	0.1574	0.1813	0.1684
Ratio de verdaderos positivos	0.8608	0.7864	0.8306
Ratio de verdaderos negativos	0.8260	0.8197	0.8324

Tabla 4.17: Indicadores de la calidad de los clasificadores para los datos de entrenamiento según la base de datos.

	BBDD_512	BBDD_1024	Multiresolución
Accuracy	82.67 %	84.78 %	82.96 %
Sensibilidad	0.8169	0.8478	0.8351
Especificidad	0.8366	0.8478	0.8241
Ratio de error	0.1732	0.1521	0.1703
Ratio de verdaderos positivos	0.8333	0.8478	0.8260
Ratio de verdaderos negativos	0.8205	0.8478	0.8333

Tabla 4.18: Resultados de la predicción.

Observando la Tabla 4.18 se puede llegar a la conclusión de que esta metodología, que no pasa por la segmentación glandular, proporciona mejores resultados cuando las glándulas comienzan a no ser distinguibles en el tejido; esto sucede en tejido patológico con más de grado 3 en la escala de Gleason.

Autores	Clases	Clasificador	Resultados
[7]	Sano, grado 3 y grado 4	SVM	Accuracy=85.6 %
[8]	Sano y cancerígeno	SVM	Accuracy=79.0 %
[10]	Grado 3 y grado 4	Random forest	Accuracy=83.0 %
[11]	Sano y cancerígeno	Keypoint matching	Accuracy=88.12 %
[12]	Sano, grado 3, grado 4 y grado 5	K-Nearest Neighbor	Accuracy=98.6 %
Método propuesto	Sano y grado 3	Quadratic SVM	Accuracy=84.78 %

Tabla 4.19: Resultados de los trabajos del estado del arte.

Capítulo 5

Conclusiones y líneas futuras

5.1 Conclusiones.

En el Capítulo 1 se ha tratado de poner de manifiesto la dificultad que tiene el diagnóstico convencional de cáncer de próstata por un patólogo mediante el análisis visual de las muestras de biopsia. A lo largo del trabajo se ha propuesto un algoritmo para tratar de ayudar en este diagnóstico mediante distintas herramientas de procesamiento digital de imagen. Los resultados obtenidos muestran que es posible distinguir entre tejido sano y cancerígeno de grado 3 a través de la imagen digitalizada de biopsia de próstata. Aunque estos resultados estén lejos de ser ideales, sirven para afirmar que es válido abordar el problema del diagnóstico de cáncer de próstata mediante herramientas informáticas.

El primer paso ha sido conocer la enfermedad desde un punto de vista histológico para ser capaces de determinar qué aspectos del tejido eran susceptibles a cambio por la enfermedad y, por tanto, iban a aportar la información necesaria para la clasificación. También en esta línea se ha realizado un análisis del estado del arte para conocer qué técnicas han sido empleadas hasta el momento y qué calidad de resultados ha sido obtenida para, por un lado, abordar el problema mediante una aproximación adecuada y, por otro lado, proponer nuevas herramientas y evaluar qué pueden aportar a la resolución del problema principal.

Tras esto, se han creado y normalizado dos bases de datos diferentes y una combinación de ambas para analizar el problema desde un punto de vista de estudio multiresolución. Estas bases de datos contienen imágenes histológicas procedentes de biopsias de próstata de distinto tamaño para poder evaluar cuál ellas acaba por proporcionar mejores resultados. Concretamente, las bases de datos son: BBDD_1024, con imágenes de tamaño 1024×1024 píxeles, BBDD_512, con imágenes de tamaño 512×512 píxeles y BBDD_multi, que contiene imágenes de las dos bases de datos anteriores.

Se han evaluado diversas aproximaciones tanto de selección de características como de clasificación para determinar cuál de todas ellas proporciona mayor poder discriminatorio; además, se ha valorado de manera cuantitativa esta capacidad de discriminación. Las diferentes técnicas de ex-

tracción de características empleadas han sido las matrices de coocurrencia, los filtros de gabor y la granulometría. Para la clasificación se han entrenado 23 clasificadores distintos pertenecientes a diferentes familias de clasificadores supervisados. Con la combinación de los mejores resultados en ambas etapas se han obtenido los siguientes resultados de precisión: 82.67 % para BBDD_512, 84.78 % para BBDD_1024 y 82.96 % para BBDD_multi. Analizando los resultados se concluye que la mejor combinación de todos los parámetros del proceso es la que se ha realizado con las imágenes de 1024×1024 píxeles.

Como ya se ha mencionado, los resultados obtenidos están lejos de ser ideales. Se ha determinado que esto puede deberse a que se han analizado las imágenes completas, sin segmentar las glándulas; es precisamente en las glándulas donde en grados iniciales de la enfermedad se pueden ver signos de ella. Como los cambios histológicos entre próstatas sanas y de grado 3 son a nivel glandular (el estroma sigue intacto), existe mucho parecido entre las imágenes cuando se analizan globalmente. Por tanto, la aproximación de segmentación glandular parece la más adecuada a la hora de distinguir entre grados bajos en la escala de Glason (hasta grado 3). Sin embargo, conforme avanza la enfermedad, las unidades glandulares comienzan a ser indistinguibles del resto del tejido, por lo que parece lógico pensar que la metodología propuesta de análisis de la imagen completa será capaz de distinguir entre tejido sano y de grados 3, 4 o 5 en la escala de Gleason.

Por tanto, se ha determinado que es probable que este algoritmo obtenga mejores resultados a la hora de clasificar entre grados superiores, ya sea entre sanas y de grado 4 o entre grado 3 y grado 4.

Cabe añadir que una ventaja importante de la metología propuesta en este trabajo es el tiempo de cómputo, al no tener que realizar la etapa de segmentación de la imagen. Esta etapa, además de consumir tiempo y recursos del procesador, es una parte crítica del proceso: si se segmenta erróneamente, la clasificación también será errónea.

Como conclusión global del trabajo se puede afirmar que se ha logrado alcanzar el objetivo principal planteado: la creación de un algoritmo capaz de clasificar imágenes histológicas de próstata en sanas y cancerígenas de grado 3.

5.2 Líneas futuras.

En base a las conclusiones extraídas de todo el proceso del trabajo se van a proponer algunas mejoras que partan de los puntos débiles del algoritmo para mejorarlo y conseguir resultados de mayor calidad.

Por un lado, resultaría interesante realizar pruebas con el algoritmo para tratar de distinguir entre grado 3 y grado 4. Como ya se ha mencionado, parece más interesante esta clasificación en base a la metodología empleada y se esperan mejores resultados. Además, haría más completo el algoritmo de manera que el diagnóstico de la enfermedad con este método resultara más competente para la ayuda al profesional médico.

En cuanto a la etapa de clasificación, se podría probar la actuación de clasificadores no supervisados; estos son capaces de mapear las características extrayendo relaciones intrínsecas de los datos que quizá un clasificador supervisado no es capaz de contemplar, pudiendo suponer también una mejora en los resultados. También se propone abordar el problema de clasificación basándose en

algoritmos de *deep learning* (aprendizaje profundo). Por otro lado, los esfuerzos en este trabajo se han dedicado principalmente a las etapas de extracción y selección de características; la etapa de clasificación se ha realizado con una herramienta propia de MATLAB. También se podría diseñar un clasificador supervisado con los parámetros que más convengan a los datos de los que se dispone.

Por último, los resultados de la clasificación, una vez se implementara el algoritmo en un sistema de ayuda al diagnóstico completo, podrían mostrarse con un porcentaje de confianza en lugar de hacer una clasificación binaria, de manera que el patólogo solo analizara visualmente aquellas zonas en las que el algoritmo obtenga bajos niveles de confianza.

Parte II

Presupuesto

Presupuesto

Este capítulo contiene la valoración económica del trabajo realizado. Se detallan las diferentes partes de las que consta el presupuesto y se realizará un breve estudio de la viabilidad económica en base a los resultados obtenidos para determinar si es eficiente emplear este tipo de métodos para alcanzar el objetivo.

En primer lugar, se van a desglosar los precios de manera que se analicen por separado la mano de obra y la maquinaria y los materiales empleados.

Dentro de la mano de obra se recogen los recursos humanos que han sido requeridos para la realización del trabajo en función del tiempo que ha sido necesario dedicar. Concretamente, las contribuciones al trabajo han venido de la mano de: Dra. Valery Naranjo Ornedo (tutora del trabajo), Adrián Colomer Granero (cotutor del trabajo) y María Jesús García Gozález (alumna y autora).

Denominación	Uds.	Cantidad	Precio unitario (€)	Total (€)
Catedrática	h.	15	42,00	630,00
Doctor	h.	45	20,50	922,50
Alumna GIB	h.	400	12,50	5.000,00
			Total	6.552,50

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

En cuanto a los recursos materiales necesarios en el trabajo se realiza una primera distinción: maquinaria y materiales.

La maquinaria se refiere a aquellos materiales que han sido necesarios para el desarrollo del trabajo explícitamente, a las herramientas que han sido necesarias en todo el proceso. Se distingue entre hardware y software.

En el caso de este trabajo, las herramientas software utilizadas han sido: MATLAB, concretamente *Image Processing Toolbox* y *Statistics and Machine Learning Toolbox*, y para la creación de la presente memoria se ha empleado Overleaf, una herramienta de escritura, edición y publicación en línea gratuita.

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
Acer Aspire E1 serie PC	u	1	600,00	4	6	150,00
					Total	1.750,00

Tabla 5.2: Cuadro de precios unitarios de la maquinaria.

Por otro lado, los materiales que han sido requeridos para el presente trabajo son las imágenes, que provienen de pruebas de biopsia realizadas en el Hospital Clínico Universitario de Valencia. El precio de las imágenes se divide en el precio relativo a la adquisición de la muestra del paciente y en el coste asociado a cada corte histológico que se digitaliza. Es importante destacar que este coste no se ha asumido explícitamente en el presente trabajo.

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 unitarios de los materiales.

Por último, se va a calcular el presupuesto total del trabajo. Para ello, a los cuadros de presupuestos parciales anteriores (Tablas 5.1, 5.2 y 5.3) hay que sumar el porcentaje de gastos generales (13 %), el beneficio industrial (6 %) y el IVA (21 %). Tras esto, se obtiene el presupuesto total que supone la realización del presente trabajo.

CAPÍTULO	IMPORTE (€)
Coste de la mano de obra	6.552,50
Coste de la maquinaria	1.750,00
Coste de los materiales	15.350,00
PRESUPUESTO DE EJECUCIÓN DE MATERIAL	23.652,50
13 % de gastos generales	3.074,83
6 % de beneficio industrial	1.419,15
PRESUPUESTO DE EJECUCIÓN POR CONTRATA	28.146,48
21 % de IVA	5.910,76
PRESUPUESTO TOTAL	**34.057,24**

Tabla 5.4: Presupuesto total.

Bibliografía

[1] Sociedad Española de Oncología Médica, "Las cifras del cáncer en España", inf. téc., 2018 (vid. pág. 3).

[2] Asociación Española Contra el Cáncer (AECC), *Cáncer de próstata*, https://www.aecc.es/es/todo-sobre-cancer/tipos-cancer/cancer-prostata (vid. pág. 5).

[3] Dr. Nicolás Vivar Díaz, *Manual de procedimientos en anatomía patológica*. 2010 (vid. pág. 6).

[4] Mayo Clinic, *Biopsia transrectal de la próstata*, https://www.mayoclinic.org/es-es/tests-procedures/prostate-biopsy/about/pac-20384734 (vid. pág. 6).

[5] D. F. Gleason, "Histologic grading of prostate cancer: A perspective", *Human Pathology*, vol. 23, n.º 3, págs. 273 -279, 1992, The Pathobiology of Prostate Cancer-Part 1, ISSN: 0046-8177. DOI: https://doi.org/10.1016/0046-8177(92)90108-F (vid. págs. 6, 7).

[6] D. Marín y E. Romero, "Sistemas de microscopía virtual: análisis y perspectivas", vol. 31, págs. 144 -155, mar. de 2011, ISSN: 0120-4157 (vid. pág. 8).

[7] 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. 9, 56).

[8] K. Nguyen, A. Sarkar y A. K. Jain, "Structure and Context in Prostatic Gland Segmentation and Classification", en *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, N. Ayache, H. Delingette, P. Golland y K. Mori, eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, págs. 115-123, ISBN: 978-3-642-33415-3 (vid. págs. 9, 56).

[9] 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. pág. 9).

[10] J. Ren, E. Sadimin, D. J. Foran y X. Qi, "Computer aided analysis of prostate histopathology images to support a refined Gleason grading system", vol. 10133, 2017, págs. 10133 -10133 -8. DOI: 10.1117/12.2253887 (vid. págs. 9, 56).

[11] O. Oğuz, A. E. Çetin y R. Çetin Atalay, "Classification of Hematoxylin and Eosin Images Using Local Binary Patterns and 1-D SIFT Algorithm", *Proceedings*, vol. 2, n.º 94, 2018, ISSN: 2504-3900. DOI: 10.3390/proceedings2020094 (vid. págs. 9, 55, 56).

[12] M. T. Farooq, A. Shaukat, U. Akram, O. Waqas y M. Ahmad, "Automatic gleason grading of prostate cancer using Gabor filter and local binary patterns", en *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*, 2017, págs. 642-645. DOI: 10.1109/TSP.2017.8076065 (vid. págs. 9, 55, 56).

[13] *MicroDraw*, http://microdraw.pasteur.fr/ (vid. pág. 14).

[14] The MathWorks, Inc., *MATLAB*, url=https://es.mathworks.com/products/matlab.html (vid. pág. 16).

[15] 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. págs. 17, 19).

[16] T. Macura, *Colour deconvolution code translated from ImageJ Java Pluging into MATLAB MEX c code*, 2006 (vid. pág. 19).

[17] O. B. Sassi, L. Sellami, M. B. Slima, K. Chtourou y A. B. Hamida, "Improved spatial gray level dependence matrices for texture analysis", *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 4, n.º 6, págs. 209-219, 2012. DOI: 10.5121/ijcsit.2012.4615 (vid. pág. 20).

[18] R. M. Haralick, K. Shanmugam e I. Dinstein, "Textural Features for Image Classification", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, n.º 6, págs. 610-621, 1973, ISSN: 0018-9472. DOI: 10.1109/TSMC.1973.4309314 (vid. págs. 20, 23).

[19] L. K. Soh y C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, n.º 2, págs. 780-795, 1999, ISSN: 0196-2892. DOI: 10.1109/36.752194 (vid. págs. 22, 24).

[20] Rahmadwati, G. Naghdy, M. Ros, C. Todd y E. Norahmawati, "Cervical Cancer Classification Using Gabor Filters", en *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*, 2011, págs. 48-52. DOI: 10.1109/HISB.2011.15 (vid. pág. 23).

[21] O. S. Al-Kadi, "A Gabor filter texture analysis approach for histopathological brain tumor subtype discrimination", *ISESCO JOURNAL of Science and Technology*, vol. 12, n.º 22, págs. 25-32, 2017 (vid. pág. 25).

[22] J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaise, A. Marx y F. G. Zöllner, "Multi-class texture analysis in colorectal cancer histology", *NATURE, Scientific reports*, vol. 6, n.º 27988, págs. 1-11, 2016. DOI: DOI:10.1038/srep27988 (vid. pág. 25).

[23] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf y G. Rätsch, "Support Vector Machines and Kernels for Computational Biology", *PLOS Computational Biology*, vol. 4, n.º 10, págs. 1-10, oct. de 2008. DOI: 10.1371/journal.pcbi.1000173 (vid. pág. 37).

[24] M. R. Berthold, C. Borgelt, F. Höppner y F. Klawonn, *Guide to Intelligent Data Analysis*. Springer-Verlag London, 2010. DOI: DOI:10.1007/978-1-84882-260-3 (vid. pág. 38).

[25] R. Soleymani, E. Granger y G. Fumera, "Loss factors for learning Boosting ensembles from imbalanced data", en *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, págs. 204-209. DOI: 10.1109/ICPR.2016.7899634 (vid. pág. 39).