UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIERÍA INDUSTRIAL VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

DISEÑO DE UN SISTEMA SEMIAUTÓMATICO DE CLASIFICACIÓN DE WHOLE SIDE IMAGES PARA LA GRADACIÓN DEL CÁNCER DE PRÓSTATA A PARTIR DE ANOTACIONES NO EXPERTAS Y REDES NEURONALES CONVOLUCIONALES

AUTORA: María Beser Robles

TUTORA: Valery Naranjo Ornedo

COTUTOR: Julio José Silva Rodriguez

Curso Académico: 2019-20

Agradecimientos

A Valery, por darme la oportunidad en este proyecto.

A Julio, por ser el pilar fundamental de este trabajo,
y por estar siempre dispuesto a ayudar.

A Josema por cuidarme y por estar
cuando nadie más está.

A los amigos que he encontrado durante esta etapa
tan bonita, por enseñarme a disfrutar de los
pequeños ratitos.

A mis amigas por saber sacarme una sonrisa y crecer
a mi lado.

A mi familia por apoyarme, motivarme y creer
siempre en mí.

Resumen

El presente TFG pretende abordar el desarrollo de un algoritmo de clasificación capaz de catalogar de manera automática muestras histológicas de tejido prostático. El desarrollo de un sistema automático fiable necesita un gran número de anotaciones de expertos. Dado la elevada carga de trabajo de esta tarea y la escasez de recursos de patólogos expertos (en nuestro caso uropatólogos), surge la idea de desarrollar un modelo que obtenga un resultado más preciso sin necesidad de anotación de personal experto.

Así pues, el objetivo final es realizar una clasificación de los grados de cáncer según la escala Gleason en imágenes histopatológicas de próstata utilizando un modelo híbrido basado en redes neuronales convolucionales y anotaciones no expertas.

Con el fin de llevar a cabo este objetivo, se anotan por el no experto los cortes histológicos de tejido de próstata a partir de la literatura aprendida sobre la escala de Gleason, se utiliza una arquitectura VGG16 para desarrollar un modelo que permita realizar una clasificación de las muestras anotadas por un experto, y finalmente se combinan la anotación no experta y el modelo en la etapa de clasificación de la red neuronal convolucional para realizar una clasificación más precisa de las muestras anotadas.

Una vez entrenados ambos modelos, estos se evalúan utilizando el estadístico de kappa de cohen. Obteniendo como resultados un valor de 0,571 para el modelo que utiliza entradas anotadas únicamente por el experto, y un valor estadístico de kappa de cohen de 0,685l, al combinar las entradas anotadas por el experto con las anotadas por el no experto mediante el modelo híbrido. También se ha comparado el acuerdo entre los dos anotadores obteniendo un valor estadístico de kappa de cohen de 0,440.

Tras analizar los resultados, concluimos que la capacidad del no experto para detectar los patrones cancerosos según su grado de Gleason es limitada, sobre todo para el grado 4, pero aún así los resultados obtenidos tras la combinación de ambas anotaciones a la CNN ayudan a mejorar la precisión del método considerablemente. Por lo que, la aplicación del modelo híbrido es una posible solución para reducir la carga de trabajo de patólogos sin que la precisión del sistema se vea tan afectada como en literatura previa.

Resum

El present TFG pretén abordar el desenrotllament d'un algoritme de classificació capaç de catalogar de manera automàtica mostres histològiques de teixit prostàtic. El desenrotllament d'un sistema automàtic fiable necessita un gran nombre d'anotacions d'experts. Dau l'elevada càrrega de treball d'esta tasca i l'escassetat de recursos de patòlegs experts (en el nostre cas uropatólogos), sorgix la idea de desenrotllar un model que obtinga un resultat més precís sense necessitat d'anotació de personal expert.

Així doncs, l'objectiu final és realitzar una classificació dels graus de càncer segons l'escala Gleason en imatges histopatológicas de pròstata utilitzant un model híbrid basat en xarxes neuronals convolucionales i anotacions no expertes.

A fi de dur a terme este objectiu, s'anoten pel no expert els talls histològics de teixit de pròstata a partir de la literatura apresa sobre l'escala de Gleason, s'utilitza una arquitectura VGG16 per a desenrotllar un model que permeta realitzar una classificació de les mostres anotades per un expert, i finalment es combinen l'anotació no experta i el model en l'etapa de classificació de la xarxa neuronal convolucional per a realitzar una classificació més precisa de les mostres anotades.

Una vegada entrenats ambdós models per l'anotador no expert, estos s'avaluen utilitzant l' estadístic de kappa de cohen. Obtenint com resultats un valor de 0,571 per al model que utilitza entrades anotades únicament per l'expert, i un valor estadístic de kappa de cohen de 0,685l, al combinar les entrades anotades per l'expert amb les anotades pel no expert per mitjà del model híbrid. També s'ha comparat l'acord entre els dos anotadors obtenint un valor estadístic de cohen kappa de 0,440.

Després d'analitzar els resultats, concloem que la capacitat del no expert per a detectar els patrons cancerosos segons el seu grau de Gleason és limitada, sobretot per al grau 4, però encara així els resultats obtinguts després de la combinació d'ambdós anotacions a la CNN ajuden a millorar la precisió del mètode considerablement. Pel que, l'aplicació del model híbrid és una possible solució per a reduir la càrrega de treball de patòlegs sense que la precisió del sistema es veja tan afectada com en literatura prèvia.

Abstract

The present TFG aims to address the development of a classification algorithm capable of automatically cataloguing histological samples of prostatic tissue. The development of a reliable automatic system requires a large number of expert annotations. Given the high workload of this task and the scarcity of resources of expert pathologists (in our case uropathologists), the idea arises of developing a model that obtains a more precise result without the need for expert annotation.

Thus, the final objective is to perform a classification of cancer grades according to the Gleason scale in prostate histopathological images using a hybrid model based on convolutional neural networks and non-expert annotations.

In order to carry out this objective, histological sections of prostate tissue are annotated by the non-expert from the literature learned about the Gleason scale, a VGG16 architecture is used to develop a model that allows a classification of the annotated samples by an expert, and finally non-expert annotation and the model are combined in the classification stage of the convolutional neural network to make a more precise classification of the annotated samples.

Once both models have been trained by the non-expert annotator, they are evaluated using the kappa cohen statistic. The results are a value of 0.571 for the model that uses entries annotated only by the expert, and a statistical value of cohen kappa of 0.685l, when combining the entries annotated by the expert with those annotated by the non-expert using the hybrid model. The agreement between the two note-takers has also been compared, obtaining a statistical value of cohen kappa of 0.440.

After analyzing the results, we conclude that the ability of the non-expert to detect cancer patterns according to their Gleason grade is limited, especially for grade 4, but still the results obtained after combining both CNN annotations help to improve the accuracy of the method considerably. Therefore, the application of the hybrid model is a possible solution to reduce the workload of pathologists without affecting the accuracy of the system as much as in previous literature.

Índice general

Parte I

Memoria

Capítulo 1

Introducción

1.1 Descripción del problema

1.1.1 Cáncer de próstata

La Organización Mundial de la Salud (OMS) define el cáncer como un proceso de crecimiento de células incontrolado que puede afectar prácticamente a cualquier parte del organismo [1]. El cáncer aparece como resultado de una división anormal de células mutadas que no pueden repararse ni morir. La enfermedad del cáncer se constituye como una de las principales causas de morbilidad y mortalidad en el mundo, y se prevé que el número de casos nuevos aumente aproximadamente en un 70 % durante las próximas décadas [2].

Su desarrollo suele estar relacionado con una serie de alteraciones en la actividad de los reguladores del ciclo celular, normalmente debido a mutaciones en los genes que codifican estas proteínas. Estas alteraciones son el resultado de los factores genéticos y los agentes externos que se listan a continuación [3]:

- Carcinógenos físicos, como radiaciones ultravioletas o ionizantes.
- Carcinógenos químicos, como el amianto o el humo de tabaco.
- Carcinógenos biológicos, como virus o bacterias.

Los factores de riesgo varían en función del país, pero cabe destacar que los principales a nivel mundial son: el consumo de tabaco, el alcohol, la mala alimentación y la inactividad física. En general, los factores de riesgo se dividen en factores evitables como la exposición a ciertas sustancias o algunos comportamientos, y factores no evitables como la edad o los antecedentes familiares.

Cabe destacar que el cáncer es susceptible de tratamiento bien mediante técnicas de radioterapia o de quimioterapia, o bien empleando cirugía, principalmente si la enfermedad se detecta en un estado temprano. El problema subyace en que, a menudo, el cáncer se diagnostica en fases

avanzadas, lo cual conlleva un tratamiento tardío que se asocia con un aumento de la mortalidad y un empeoramiento de la calidad de vida.

En el presente proyecto se investiga sobre el cáncer de próstata, este es el tercer tumor más frecuente en varones españoles y la tercera causa de muerte por cáncer. Según la OMS existen 1.276.106 nuevos casos al año en el mundo, en 2018 se diagnosticaron 1.3 millones de pacientes y se estima que el número de casos nuevos aumente anualmente hasta el 40.2 % en 2030 [4]. El cáncer de próstata se diagnostica más frecuentemente en hombres de edad avanzada, siendo la edad media en la que se diagnostica a un paciente con dicha enfermedad de 66 años.

Esta prevalencia viene determinada por la cantidad de años que un paciente puede sobrevivir con cáncer de próstata y, dado que la tasa de supervivencia a 5 años es del 99 %, la prevalencia es muy alta. Además, actualmente el 92 % de los casos se detecta cuando la enfermedad está ubicada en la glándula prostática o en los órganos adyacentes, es decir, cuando la enfermedad se encuentra localmente avanzada [5].

La anatomía de la próstata es relevante para comprender los diferentes métodos diagnósticos y sus dificultades. La próstata es un órgano interno del aparato reproductor masculino que se encuentra situado en la pelvis, detrás del pubis, delante del recto y por debajo de la vejiga (véase Figura 1.1). Abarca la primera porción de la uretra y la atraviesa en toda su longitud, debido a ello, gran parte de los procesos patológicos ocasionados en esta glándula provocan alteraciones en la orina o en su evacuación [6].

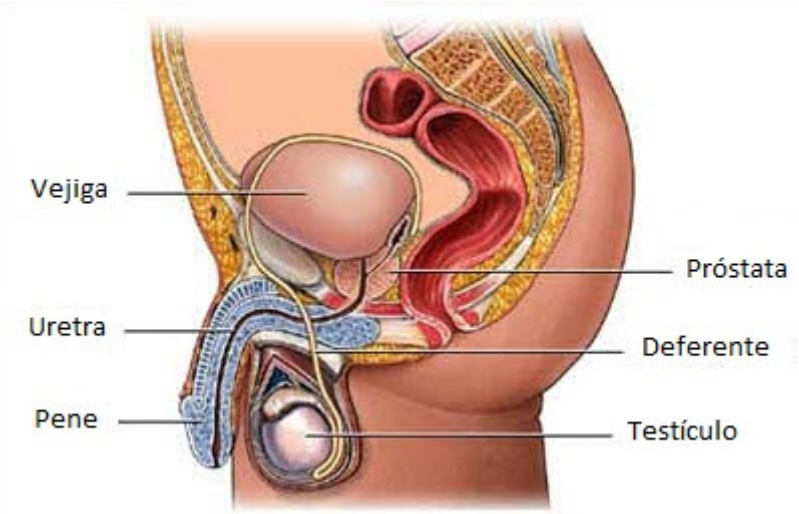

Figura 1.1: Anatomia de la próstata

El tamaño normal de este órgano es de 4 cm de largo por 3 cm de ancho, pero se debe tener en cuenta que aumenta con la edad. Su función es producir el líquido prostático que junto con el líquido vesicular forman el líquido seminal que aporta el medio necesario para mantener, proteger y ayudar a los espermatozoides o gametos masculinos.

El organismo del ser humano está formado por tejidos, que a su vez están compuestos por células que siguen un conjunto ordenado de sucesos denominado el ciclo celular. Como se ha explicado anteriormente, el cáncer se origina cuando las células sanas sufren un cambio y proliferan sin

control. De esta forma, cuando las células tumorales están ubicadas en el órgano prostático se habla de cáncer de próstata.

Un diagnóstico a tiempo es vital en esta enfermedad ya que, de lo contrario, el pronóstico de la enfermedad empeora considerablemente, así como la calidad de vida de las personas que la padecen.

1.1.2 Diagnóstico del cáncer de próstata

Hoy en día existen una serie de pruebas diagnósticas preliminares para detectar la sospecha de un cáncer. Estas incluyen:

- Análisis de PSA en sangre. Busca la presencia del antígeno prostático específico (PSA). Cuanto mayor es el valor de la concentración de PSA, mayor es el riesgo de progresión. Aun así, es un marcador inespecífico, puede estar alterado por más razones y sus resultados deben ser interpretados cautelosamente.
- Examen rectal digital (DRE). Exploración médica que consiste en la introducción de un dedo a través del esfínter anal para realizar la palpación digital en busca de anomalías en el tamaño de la glándula prostática.
- Análisis de PCA3. Busca la existencia del gen PCA3 de forma más específica que el análisis de PSA en sangre, ya que, en este caso, el gen únicamente se encuentra sobreexpresado en casos de cáncer de próstata. No se encuentra aún implementado en los hospitales.
- Ecografía transrectal. Procedimiento en el que se introduce una sonda en el recto produciendo ecos que rebotan en los tejidos. A partir de estos ecos se forman imágenes computarizadas de la próstata. Además, se utiliza para guiar las agujas durante la biopsia.

Pero la única manera de formular un diagnóstico definitivo para confirmar si el cáncer está presente en la próstata es la realización de una biopsia y el análisis visual de esta por el patólogo. La biopsia es una técnica invasiva y, por ello, se realizan previamente las pruebas inespecíficas de detección de cáncer de próstata nombradas anteriormente con el fin de evitar la biopsia. Esta técnica consiste en la extirpación de una pequeña zona de tejido para poder examinarla y analizar indicios de cáncer. Mediante la biopsia se obtiene un diagnóstico definitivo cuando las pruebas de detección proporcionan resultados anormales. Este procedimiento conlleva ciertos riesgos, debido a que se realiza vía transrectal, como dificultad para orinar, hemorragia e infección. Normalmente, se lleva a cabo un abordaje transrectal con una pistola para biopsias y una guía ecográfica.

Tras la extracción del tejido, el patólogo realiza la preparación de la biopsia, mediante la fijación y tinción con H&E de los cortes histológicos y estos son analizados bajo el microscopio. En particular el experto estudia el grado de diferenciación de las glándulas del tejido, que es el principal marcador de cáncer de próstata. A menor diferenciación el cáncer presenta un mayor avance en el tejido. Para poder cuantificar el grado de diferenciación se emplea la escala de Gleason (la cual se explica en 4.1), que se centra en el patrón arquitectónico de los tejidos, determinando la agresividad del cáncer. Esta diferenciación únicamente es observable en tejido laminado a partir del grado 3 de Gleason [7].

Del grado de Gleason diagnosticado depende el pronóstico y el tratamiento a aplicar en el paciente, por lo que debe ser lo más precisa posible. Debido al gran tamaño de las biopsias magnificadas

al microscopio este análisis consume grandes cantidades de tiempo al experto médico y presenta variabilidad intra e inter patólogo. Es por ello por lo que en los últimos años ha surgido la necesidad de desarrollar sistemas de ayuda al diagnóstico para facilitar la clasificación y reducir la carga de trabajo que esta tarea presenta, basados en el análisis mediante técnicas de visión artificial de las biopsias digitalizadas en *Whole Slide Images (WSIs)*.

1.1.3 Imágenes histopatológicas

Como se ha detallado anteriormente, el diagnóstico definitivo para el cáncer de próstata reside en el análisis de las muestras del tejido extraído en una biopsia. Dichas muestras, tras un proceso de preparación, se colocan en un portaobjetos para obtener las imágenes histopatológicas que posteriormente analizan los expertos, a fin de determinar la presencia o ausencia de cáncer en el paciente.

Para poder realizar el análisis histológico del tejido extraído mediante la biopsia es necesario un procesado de la muestra conocido como técnica histológica. Dicho procesado consiste en un conjunto de operaciones a las que se somete el material biológico para que sea posible su estudio bajo el microscopio por los especialistas en anatomía patológica.

El primer paso es la fijación, mediante perfusión, cuyo objetivo es conservar las estructuras celulares y moleculares del tejido. El siguiente paso es la inclusión, cuya finalidad es proporcionar consistencia al tejido para permitir obtener cortes delgados que puedan ser observados en el microscopio. Normalmente se utiliza parafina. Posteriormente, se hace uso del micrótomo para realizar cortes en secciones lo suficientemente delgadas como para permitir el paso de luz.

Finalmente, se elimina la parafina y se aplica una tinción al corte que permita observar la morfología tisular. El tipo de tinte escogido varía en función de las estructuras que se quieran diferenciar. La tinción empleada para el análisis de las muestras de próstata es la hematoxilina y eosina (H&E). La tinción H&E es la usada comúnmente para medicina diagnóstica. La hematoxilina debido a su naturaleza básica tiñe en tonos azul y púrpura las estructuras ácidas (basófilas) como los núcleos. En cambio, la eosina, por ser ácida, tiñe en tonos rosa los componentes básicos (acidófilos) como el estroma [8]. Gracias a esta tinción se pueden observar resaltadas las estructuras de interés. Un ejemplo de glándulas prostáticas benignas vistas con HyE se presenta en la Figura 1.2

Figura 1.2: Imagen histológica prostática donde se resaltan sus elementos principales [9]

1.1.4 Motivación

Como se ha detallado anteriormente, el diagnóstico del cáncer de próstata es realizado por expertos patólogos mediante el análisis visual de biopsias siguiendo la escala de Gleason. Para reducir la carga de trabajo que este procedimiento conlleva a los patólogos, en los últimos años se han desarrollado sistemas de ayuda al diagnóstico basados en el análisis automático de imágenes histológicas digitalizadas (*whole slide images*, WSIs). Este análisis está basado en algoritmos de visión por computador basados en técnicas de aprendizaje profundo.

Sin embargo, estos sistemas presentan ciertas limitaciones. El mayor problema en el análisis de imágenes patológicas mediante el aprendizaje automático es que solo hay disponible un pequeño número de imágenes etiquetadas, la clave del éxito del aprendizaje profundo es la utilización de grandes cantidades de datos, a partir de los cuales el modelo pueda aprender, por lo que el número de imágenes con las que cuenta el modelo es un factor limitante [10]. Otra de las limitaciones que se pueden observar en la utilización de métodos de aprendizaje profundo son los limitados resultados obtenidos para la gradación de Gleason en muestras de próstata, como se presenta en el estado del arte.

Por ello, en este TFG se busca desarrollar un sistema semi-automático basado en la combinación de sistemas de visión por computador y anotadores no expertos, que mejore la precisión de los sistemas de ayuda al diagnóstico actuales sin la necesidad de requerir grandes cantidades de datos anotados por patólogos expertos, utilizando nuevas muestras anotadas por no expertos. Este sistema, sería de gran ayuda para la formación de residentes, así como para el desarrollo de sistemas automáticos más precisos sin necesidad de anotaciones del personal experto.

1.2 Proyecto SICAP

En el presente TFG se busca desarrollar un sistema de clasificación automática de glándulas prostáticas mediante una rama de la inteligencia artificial denominada aprendizaje profundo *deep learning*, este trabajo se enmarca en un proyecto de mayor envergadura denominado SICAP (Sistema de Interpretación de Imágenes Histopatológicas para la Detección del Cáncer de Próstata).

SICAP es un proyecto nacional subvencionado por el Ministerio de Economía, Industria y Competitividad. El grupo CVBLab (Computer Vision and Behavoir Analysis Lab), perteneciente al I3B (Instituto de Investigación e Innovación en Bioingeniería) de la Universitat Politècnica de València (UPV) colabora en este proyecto junto con la Universidad de Granada y el Servicio de Anatomía Patológica del Hospital Clínico Universitario de Valencia. En este proyecto trabajan conjuntamente médicos, ingenieros y matemáticos como un equipo coordinado y multidisciplinar cuya finalidad es la de proporcionar un software (con vistas a la integración en la práctica clínica) para el diagnóstico automático del cáncer de próstata [11].

El objetivo global del proyecto es diseñar y desarrollar un sistema de ayuda al diagnóstico que permita clasificar de forma automática las muestras biopsiadas, según la escala Gleason. De esta forma, sería posible ayudar a los patólogos a mejorar en términos de tiempo y eficacia, así como a reducir el nivel de discordancia que existe entre ellos cuando intentan clasificar una determinada muestra. Así como aumentar la calidad y la exactitud diagnóstica, y mejorar en términos de coste-efectividad. Para ello, se implementan novedosas técnicas de segmentación, robustos métodos de extracción y selección de características y estrategias de *machine learning* y *deep learning* para la clasificación.

El presente TFG sigue las lineas de investigación del proyecto relacionadas con la optimización de los datos anotados por expertos patólogos y se introduce el estudio de la inclusión de anotaciones no expertas en el sistema. La temática abarcada en este trabajo es una extensión de las asignaturas cursadas por la autora del presente TFG en el grado de ingeniería biomédica en concreto, histología, imágenes biomédicas y morfología celular.

1.3 Estado del arte

1.3.1 Grados de Gleason

Los carcinomas de próstata se clasifican según el sistema de puntuación de Gleason, establecido por primera vez por Donald Gleason en 1966. El sistema de puntuación de Gleason es reconocido por la Organización Mundial de la Salud (OMS) y ha sido modificado y revisado en 2005 y 2014 por la Sociedad Internacional de Patología Urológica (ISUP). A pesar de varios cambios en el diagnóstico clínico del cáncer de próstata, el sistema de puntuación histológica de Gleason sigue siendo la herramienta de pronóstico más poderosa [12].

La evaluación se basa exclusivamente en el patrón arquitectónico del tumor, es decir, los patrones de Gleason. A diferentes patrones histológicos se les asignan números del 1 (bien diferenciados) al 5 (poco diferenciados). La puntuación final de Gleason se informa como la suma de los dos patrones más predominantes presentes en la muestra histológica, y en la práctica clínica actual la puntuación de Gleason más bajo asignado es Gleason 6 (3 + 3) [13].

La anotación final de la puntuación de Gleason de las imágenes de biopsia de tejido depende de la evaluación del patólogo respectivo, que por lo tanto constituye un factor importante para el diagnóstico y las decisiones terapéuticas. La evaluación histológica del tejido canceroso, basada en una evaluación visual sobre microscopía de patrones morfológicos y celulares no triviales, es un trabajo tedioso de anotación humana, propenso a errores y, a menudo, tiene una reproducibilidad limitada. El cáncer de próstata es una enfermedad muy heterogénea que se manifiesta en una variedad de patrones histológicos muy diferentes entre pacientes, en particular, los patrones de Gleason de riesgo intermedio 3 y 4 pueden ser muy difíciles de asignar sin ambigüedades [14].

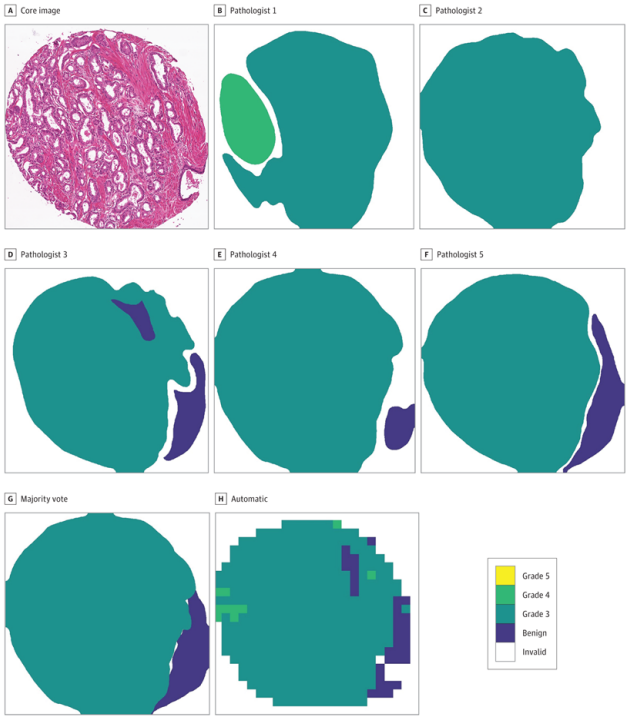

Figura 1.3: Misma muestra de tejido canceroso anotada por distintos patólogos, para mostrar la gran variabilidad interobservador. Obtenida de [13]

1.3.2 Sistemas de ayuda al diagnóstico del cáncer de próstata

Para poder superar las limitaciones mencionadas anteriormente, las líneas de investigación actuales han optado por dar un enfoque computacional automatizado, introduciendo los sistemas de exploración de diapositivas completas (WSI), digitalizando portaobjetos de vidrio completos y analizando estas mediante algoritmos computacionales. Las imágenes digitales de diapositivas completas permiten la aplicación de técnicas de análisis de imágenes para ayudar a los patólogos en el examen y cuantificación de las diapositivas. De esta manera podemos obtener resultados reproducibles (reduciendo la variabilidad interobservador en el diagnóstico) y lograr un alto rendimiento. Debido al gran tamaño que tienen estas imágenes cada WSI se divide comúnmente en regiones parciales de aproximadamente 512×512 píxeles (parches), y cada parche se analiza de forma independiente.

Históricamente, los enfoques computacionales desarrollados para este propósito se basan en características de imagen explicitas, definidas a priori y emplean técnicas convencionales de regresión o clasificación para realizar la selección de características y la asociación con los parámetros clínicos, como por ejemplo, algoritmos para detectar neuroblastoma [15], cuantificación de la infiltración linfocítica en el tejido de biopsia de mama [16], y la clasificación de los astrocitomas en el tejido cerebral [17], por nombrar algunos.

En el contexto del cáncer de próstata, los investigadores han utilizado una variedad de características para analizar tejidos que van desde características de imagen de bajo nivel (color, textura, *wavelets*) [18], características de coincidencia de segundo orden [19] y atributos morfométricos Farjam y col. [20] emplearon la morfología de las glándulas para identificar la malignidad de los tejidos de biopsia, mientras que Diamond et al. [21] utilizaron características morfológicas y de textura para identificar regiones de tejido de 100 por 100 píxeles como estroma, epitelio o tejido canceroso (un problema de tres clases). Tabesh y col. [9] desarrolló un sistema CAD que emplea textura, color y morfometría en microarrays de tejidos para distinguir entre regiones cancerosas y no cancerosas, así como entre cánceres de próstata de alto y bajo grado de Gleason (ambos casos usaron clasificación binaria). Utilizando estas características fue posible definir un esquema de clasificación para cada subregión, de forma que el sistema de visión artificial integra estas reglas de clasificación y genera mapas digitales de composición de tejidos a partir de la clasificación de subregiones, que pueden ser cuantificables.

En los últimos años, el aprendizaje profundo se ha convertido en una alternativa a las técnicas basadas en ingeniería de características antes mencionadas. Los sistemas de aprendizaje profundo se basan en redes neuronales que pueden extraer características complejas de forma automática y relacionadas con las tareas directamente de los datos. En particular, dentro del campo de la visión por computador, los enfoques basados en *deep learning* utilizan redes neuronales convolucionales (CNNs), las cuales han tenido éxito al abordar una amplia gama de tareas de análisis de imágenes biomédicas.

En el campo de la histopatología, Ciresan y col. fueron los primeros en aplicar redes neuronales convolucionales a la tarea de contar mitosis para la clasificación primaria del cáncer de mama. Además, mostraron la aplicabilidad de las redes neuronales convolucionales controladas por parches a las tareas de segmentación. Más tarde, Su et al. utilizó otra técnica de aprendizaje profundo, llamada auto-codificadores de eliminación de ruido apilados para realizar la detección y segmentación de células en cáncer de pulmón y tumores cerebrales [19].

Estudios recientes han demostrado que los sistemas de aprendizaje profundo pueden detectar malignidades con precisión en imágenes histopatológicas, en concreto en el cáncer de próstata. El análisis de las imágenes de patología digital del cáncer de próstata incluye la detección de tejido canceroso, la predicción del estado de mutación SPOP [22] y la recurrencia del cáncer, así como la caracterización de la heterogeneidad del tejido a través del aprendizaje no supervisado.

Un primer enfoque basado en el aprendizaje profundo para la predicción de la puntuación de Gleason es el estudio de Källén et al. [23]. Sin embargo, la evaluación de su método se limitó a los portaobjetos de tejido con clasificación homogénea de Gleason, a pesar de que, típicamente, los portaobjetos de tejido contienen regiones de patrón de Gleason heterogéneas. En un trabajo más reciente, Zhou et al. enfocado en puntuaciones intermedias de Gleason [24]. Su algoritmo se probó en WSIs de The Cancer Genome Atlas (TCGA), logrando una precisión general del 75 % en la diferenciación de Gleason 3 + 4 de las diapositivas de Gleason 4 + 3. Finalmente, del Toro et al. también usó WSIs de cáncer de próstata de TCGA y entrenó un clasificador binario para discriminar imágenes de puntuación de Gleason bajo (7 o más bajo) versus alto (8 o más alto).

Estudios más recientes sobre sistemas de aprendizaje profundo para la clasificación automática de Gleason de microarrays de tejido de cáncer de próstata se basan en el re-entrenamiento de redes neuronales convolucionales tales como VGG, MobileNet o ResNet-50. Los resultados obtenidos respecto al estadístico kappa cuadrático de Cohen son alentadores. Un estudio para dos anotadores independientes realizado por Arvaniti et al. [25] presenta un kappa cuadrático de 0,49 y 0,55 en la clasificación de las imágenes entre no canceroso, grado 3, grado 4 y grado 5 en imágenes anotadas por dos expertos patólogos. Logrando la estratificación de pacientes en grupos pronósticamente distintos.

La mayoría de los estudios encontrados en la literatura utilizan los datos de un solo experto para entrenar y evaluar sus modelos, una práctica que ignora la evidencia de alta variabilidad interobservador. No obstante un estudio realizado por Nir G, et al. [13] presenta una comparación para evaluar el rendimiento de un clasificador automático del cáncer de próstata en WSIs en función del número de anotadores, obteniendo un kappa cuadrático de 0,3-0,58 al entrenarse el modelo con un solo anotador y de 0,6 al entrenarse con datos de múltiples anotaciones.

Podemos ver que una gran parte de los estudios han seguido una base de datos basada en parches, en la que se incluyen los parches de todos los pacientes en los distintos grupos de entrenamiento y prueba, en esta aparece un problema y es que, es probable que los parches extraídos de un paciente incluyan mucha información exclusiva de este que puede hacer que el modelo no sea entrenado correctamente. Debido a que el objetivo de la clasificación automática mediante aprendizaje profundo es si un paciente puede ser evaluado correctamente en función del aprendizaje del modelo con otros pacientes, parece que una base de datos basada en los pacientes es la única forma de evaluar el modelo.

Aunque el aprendizaje profundo es un campo de investigación activo, la aplicación de este a la histopatología es relativamente nueva. La mayoría del trabajo ya publicado se ha centrado en la detección de figuras mitóticas o la identificación y segmentación de células individuales. Los estudios realizados hasta el momento en el campo de la clasificación del cáncer de próstata muestran resultados prometedores con respecto a la aplicabilidad de soluciones basadas en el aprendizaje profundo, hacia una clasificación más objetiva y reproducible. Además, diversos

estudios ofrecen buenos resultados al introducir la utilización de datos multiexpertos para mejorar la reproducibilidad del modelo.

Capítulo 2

Objetivos

El objetivo final del presente trabajo es la clasificación automática de glándulas prostáticas a partir de imágenes histopatológicas digitalizadas, desarrollando un sistema híbrido que combine anotaciones no expertas y redes neuronales convolucionales. Este estudio forma parte de un proyecto mayor que tiene como finalidad ayudar a los profesionales en anatomía patológica a formular un diagnóstico preciso y precoz sobre el cáncer de próstata.

En este trabajo se pretende desarrollar un algoritmo que sea capaz de discriminar entre muestras sanas y patológicas de grado 3, 4 y 5, atendiendo a las estructuras glandulares del tejido y de acuerdo con la escala Gleason. Además de la clasificación, se pretende hacer hincapié en la necesidad de mejorar la precisión de un sistema entrenado con pocas muestras anotadas por un patólogo experto, utilizando anotaciones de un no experto. Para lograr dicho objetivo se debe abordar una serie de objetivos secundarios que se listan a continuación:

1. Revisar la literatura científica actual para conocer el estado del arte relativo a las técnicas de segmentación y clasificación implementadas en imágenes histopatológicas de próstata, principalmente la clasificación de Gleason.
2. Anotación de la base de datos de imágenes *whole slide images* de próstata que permite analizar imágenes digitales de alta resolución de muestras completas de biopsias. Esta anotación se lleva a cabo a partir de la información aprendida por la literatura, creando así un *ground truth* anotado por un no experto.
3. Creación de un *ground truth* partiendo de la anotación de un experto, a partir de las mismas imágenes anotadas por el no experto.
4. Partición de las dos bases de datos en diferentes subconjuntos en función de paciente y de grado de malignidad.
5. Desarrollo y optimización de un modelo basado en redes neuronales convolucionales para predecir el grado de Gleason en las imágenes de tejido, a partir sólo de las anotaciones expertas. Hallando los hiperparámetros óptimos para llevar a cabo la fase de entrenamiento.

6. Combinación del modelo basado en redes neuronales convolucionales (utilizando anotaciones expertas) con las anotaciones no expertas para mejorar la precisión del sistema.
7. Llevar a cabo una comparación entre el *ground truth* y los resultados obtenidos por el anotador no experto, el modelo basado en redes neuronales convolucionales, y el modelo híbrido.
8. Extraer conclusiones mediante el análisis de los resultados obtenidos y comprobar que se ha cumplido el objetivo principal. Identificar problemas y proponer posibles mejoras y líneas para investigaciones futuras.

En el presente trabajo se introduce como novedad respecto a literatura previa para la gradación semi-automática de Gleason la mejora en la precisión de la clasificación del sistema creado con anotaciones realizadas por un experto, a partir de anotaciones realizadas por un no experto. Esto conlleva una considerable mejora ante uno de los problemas más importantes en la creación de modelos automáticos, que es la anotación de grandes bases de datos a partir de anotadores expertos.

Capítulo 3

Materiales

3.1 Base de datos de imágenes histopatológicas

El material de partida para este proyecto corresponde a los cortes histológicos obtenidos a partir de las biopsias realizadas a diversos pacientes en el Hospital Clínico de Valencia. Las muestras de biopsia están disponibles, en formato digital, en una base de datos privada, donde únicamente tiene acceso el personal autorizado de las distintas instituciones que participan en el proyecto. Es necesaria la privacidad de la base de datos para garantizar la confidencialidad del paciente, los cuales firman un consentimiento informado con la finalidad de proteger sus derechos y su seguridad. En este TFG se utilizan imágenes de 17 pacientes diferentes.

Una vez anotados los cortes histológicos tanto por el patólogo como por el no experto, se tienen dos bases de datos de imágenes de tamaño del orden de [50000, 90000] x [90000, 190000] píxeles, lo que se traduce en un espacio de almacenamiento que puede ocupar desde los 300 hasta los 1.500 MB (1,5 GB). Por ello, al digitalizar las imágenes para crear la base de datos, se realiza el muestreo con el máximo nivel de zoom, 40x, con el fin de conseguir una resolución espacial lo más alta posible. Sin embargo, trabajar a máximo nivel de zoom presenta inconvenientes relacionados con la ausencia de estructuras glandulares completas.

Es por ello por lo que los parches se remuestrean a imágenes de tamaño aceptable computacionalmente para su análisis digital, ya que con este aumento óptico es posible ver las estructuras glandulares en su totalidad e incluso encontrar un cierto número por cada imagen. Aunque la resolución no sea óptima, es necesario distinguir todos los tipos de estructuras tisulares presentes en una imagen y la posterior clasificación del tejido en una clase u otra [26].

A partir de las imágenes de partida se extraen parches pequeños (512×512× 3 píxeles) con los que poder trabajar a una resolución adecuada y con un número de glándulas por imagen razonable, estos parches de imágenes ocupan 258 MB.

Tras la anotación de las imágenes por parte del experto y el parcheado de estas como se ha explicado en el apartado anterior, el número de pacientes y el número de parches empleados en este trabajo se divide como se observa en la Tabla 3.1

	Nº pacientes	Parches tejido benigno	Parches tejido patológico grado 3	Parches tejido patológico grado 4	Parches tejido patológico grado 5
Anotación experto	17	3546	293	255	308

Tabla 3.1: Número de pacientes e imágenes de cada tejido anotadas por el experto

3.2 Software y hardware

3.2.1 Software

Para la anotación de las muestras histológicas se utiliza la aplicación “MicroDraw". MicroDraw es una aplicación desarrollada sobre una librería de código abierto llamada OpenSeadragon que permite una visualización interactiva multi-resolución sin pérdida de calidad. La plataforma MicroDraw está escrita en JavaScript, usando HTML5, CSS3 y jQuery, y ha sido modificada por los ingenieros del CVBLab, junto con los patólogos, con la finalidad de que puedan realizar anotaciones en las imágenes según los diferentes grados de la escala Gleason. Para poder utilizar la aplicación de MicroDraw, primero es necesario hacer uso de una librería de procesado de imagen llamada “vips” que permite convertir las imágenes completas, proporcionadas por el escáner en formato *.tiff, en otras con diferentes niveles de resolución en formato *.dzi. Esto permite que MicroDraw pueda ir cargando la imagen *.dzi que requiera cada nivel de aumento del que precisa el patólogo, manteniendo en todo momento la calidad original de la imagen. Esta aplicación es utilizada tanto por el patólogo experto como por el anotador no experto, para que las anotaciones se encuentren en el mismo formato.

Para el desarrollo y la implementación de los distintos algoritmos de preprocesado de las imágenes, así como para el cálculo de los resultados de este TFG se ha empleado el programa MATLAB® v.R2019a, de The MathWorks, Inc. (Natick, Massachusetts, Estados Unidos). Este software combina un entorno de escritorio perfeccionado para el análisis iterativo y los procesos de diseño con un lenguaje de programación que expresa las matemáticas de matrices y vectores directamente. MATLAB permite el desarrollo de algoritmos, la adquisición de datos y otras tareas relacionadas con el modelado, la simulación y la construcción de interfaces gráficas de usuario [27].

Para la etapa de entrenamiento y predicción de redes convolucionales (CNN) basadas en *deep learning*, se emplea el lenguaje de programación Python3.8.1 mediante el uso de PyCharm 2020.1 como entorno de desarrollo integrado (IDE). Además, se hace uso de la librería NVIDIA Cuda R Deep Neural Network y de la librería OPENCV. Finalmente, a lo largo de este proyecto se ha empleado la biblioteca de Keras como *framework* usando TensorFlow como *backend* para implementar, entrenar y probar diferentes tipos neuronales con diferentes parámetros.

3.2.2 Hardware

Es importante conocer las características internas del hardware en el que se ha llevado a cabo el proyecto, ya que de esto depende tanto el funcionamiento como la velocidad y el rendimiento que se puede alcanzar al compilar y ejecutar los distintos algoritmos.

Este proyecto ha sido desarrollado y ejecutado en un equipo compuesto por un procesador HP intel® Core™ i5 8250U @1,6 GHz. En cuanto a la memoria RAM, la capacidad es de 6 MB, con un disco duro SSD 256 GB y, respecto al tipo de sistema operativo, se trata de un Windows 10 Home 64 bits. Para el entrenamiento de los algoritmos de *deep learning*, se ha utilizado una tarjeta gráfica Tital V donada por NVIDIA Corporation.

Capítulo 4

Métodos

4.1 Anotación de los grados de Gleason

En esta sección, se describen los patrones y propiedades histológicas en el tejido en las cuales el anotador no experto se ha basado para realizar las anotaciones.

En la glándula prostática se identifican diferentes tipos celulares: células basales, intermedias, secretoras y neuroendocrinas. La gran mayoría corresponde a adenocarcinomas (glándulas de tamaño intermedio o pequeño con tendencia a la agrupación irregular que crecen entre glándulas grandes benignas). A medida que se pierde la diferenciación, el tamaño de las glándulas se va reduciendo, pudiendo fusionarse entre sí, adoptando forma de cordón, donde no es posible identificar componente glandular alguno. Esta evolución tumoral fue utilizada por Gleason para establecer sus patrones de clasificación de los grados de cáncer de próstata, por lo que puede considerarse un modelo evolutivo [28].

La gradación de Gleason se basa pues en la evaluación progresiva de la pérdida de patrón glandular, y la creciente invasión del estroma peritumoral. Se definen 5 categorías o patrones de diferenciación que van del 1 (bien diferenciado) al 5 (pobremente diferenciado). Teniendo en cuenta la heterogeneidad morfológica del adenocarcinoma de próstata primero se da una clasificación al patrón más dominante, y después al segundo patrón más prevalente. Los grados se suman para obtener la puntuación de Gleason, que va de 2 a 10. Estableciendo un sistema de 9 niveles que ofrecen una idea sobre el pronóstico y la progresión del tumor [29].

Este sistema fue adoptado como sistema principal de clasificación de manera gradual en todo el mundo en la década de 1990, reemplazando a una multitud de sistemas de calificación competitivos, debido a su capacidad de capturar la compleja heterogeneidad arquitectónica de los cánceres de próstata en un solo dibujo de Gleason [30].

El sistema de gradación de Gleason ha ido evolucionando con los años, estudiando el modo de agrupar algunos de estos grupos que presentaban un pronóstico similar para aumentar su impacto pronóstico, reducir la variación interobservador y mejorar la concordancia entre la biopsia con

aguja de próstata y la clasificación de prostatectomía radical, simplificando así los protocolos terapéuticos disponibles. Esto resultó en el actual sistema de calificación de cinco niveles, con una descripción mucho más detallada de los patrones arquitectónicos individuales [31].

Este sistema de grados de grupos pronósticos (GG) está basado en las observaciones originales de Gleason y es el resultado de la evolución gradual de nuestra comprensión de la naturaleza biológica y de los distintos patrones morfológicos del cáncer de próstata. Consta de 5 grados (GG): grado 1 (GG1: Gleason <6); grado 2 (GG2: Gleason 3 + 4= 7); grado 3 (GG3: Gleason 4 + 3= 7); grado 4 (GG4: Gleason 8), y grado 5 (GG5: Gleason 9-10), siendo los grados 3, 4 y 5 (Tabla 4.1) .

Grado	**Puntuación de Gleason**	**Características:**
1	<6	Glándulas individuales discretas y bien diferenciadas
2	3+4=7	Glándulas predominantemente bien formadas con un menor componente de glándulas pobremente formadas, fusionadas y/o cribriformes
3	4+3=7	Glándulas predominantemente pobremente diferenciadas, fusionadas y/o cribriformes con menor componente de glándulas bien formadas
4	4+4=8	Glándulas pobremente diferenciadas, fusionadas y/o cribriformes
	3+5=8	Glándulas predominantemente bien diferenciadas con un componente menor de tejido no glandular
	5+3=8	Predominio de tejido no glandular con un menor componente de glándulas bien formadas
5	9-10	Ausencia de tejido glandular (o con necrosis) con o sin glándulas pobremente diferenciadas, fusionadas y/o cribriformes

Tabla 4.1: Tabla puntuación de grados de Gleason actual (grados ISUP), obtenida de [30]

Esta evolución del sistema de Gleason nos ofrece múltiples ventajas como la diferenciación entre grados 2 y 3, lo que puede ser un factor clave para distinguir entre un cáncer de riesgo intermedio favorable o desfavorable. Esta evolución del sistema de Gleason presenta un problema y es que no se realiza distinciones entre los grados 4+4, 3+5 y 5+3, que han sido explicados anteriormente, ya que todos se engloban en el grado 4, por ello se recomienda informar del grado agregado para cada tejido de biopsia etiquetado individualmente para estratificar a los pacientes con diferentes pronósticos clínicos. A continuación, se explican los patrones 3, 4 y 5 de Gleason cuyo diagnóstico presenta un cáncer más agresivo (véase 4.1)

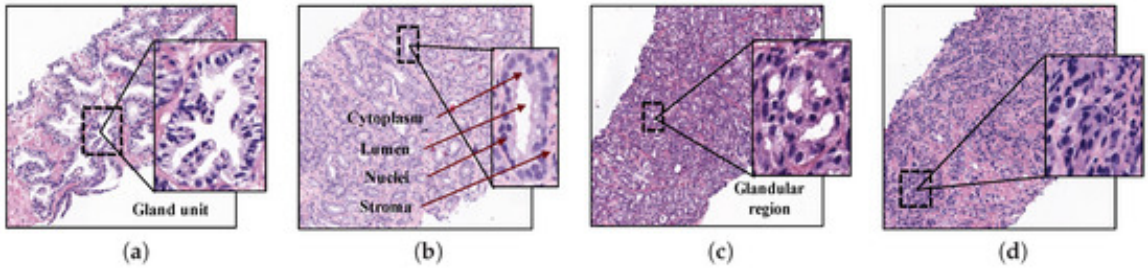

Figura 4.1: Muestras de tejidos de próstata histopatológicos con diferentes patrones según la escala de Gleason. (a) No canceroso; (b) Grado 3; (c) Grado 4; (d) Grado 5. Obtenida de [32]

El patrón de grado 3 de Gleason se define por glándulas bien diferenciadas, separadas entre sí por estroma y tiene dimensiones de lumen y glándulas pequeñas y circulares. Se pueden distinguir algunos subpatrones, según la densidad de las glándulas neoplásicas (disperso, intermedio y denso) y las variaciones arquitectónicas (atrófico, derivación, fibroplasia mucinosa...) Además, la densidad celular del núcleo epitelial es más baja en el tejido patológico. Este patrón puede comportarse localmente de forma agresiva, pero rara vez en el momento de la prostatectomía.

El patrón de grado 4 de Gleason se caracteriza por tener regiones glandulares compuestas por la fusión de glándulas poco definidas, pero no unidades glandulares. Comprende glándulas fusionadas (glándulas pequeñas o grandes que ya no están completamente separadas por el estroma) que suelen aparecer combinadas con glándulas malformadas o pobremente formadas (glándulas discretas de pequeño tamaño sin o con raros lúmenes), cuya presencia en altas cantidades indica un mayor riesgo de presentar recurrencia del carcinoma.

También es característica de este grado la arquitectura glomeruloide (glándulas dilatadas con salientes hacia el lumen, que por su forma podrían ser una arquitectura predecesora a la cribiforme) y cribiforme (área sin estroma y con mucho lumen perforado), la presencia de esta última aporta un mayor riesgo de progresión de la enfermedad en comparación con el resto de las arquitecturas. También aparece una arquitectura de adenocarcinoma papilar o ductal que presentan largas columnas de células recubriendo la glándula con núcleos alargados.

El patrón de grado 5 de Gleason representa áreas de carcinoma detectables a baja potencia, que esencialmente carecen de características glandulares como la formación de lumen o contornos glandulares y se caracteriza por presentar muchos núcleos dispersos por el estroma. Se pueden distinguir algunos patrones principales que incluyen células individuales pobremente diferenciadas, láminas tumorales, nidos sólidos, cuerdas y matrices lineales, así como comedonecrosis. Está relacionado con carcinoma anaplásico (células que se multiplican rápidamente) [33].

4.2 Predicción automática de los grados de Gleason mediante Redes Neuronales Convolucionales

Los sistemas de aprendizaje profundo se basan en redes neuronales de varias capas que pueden extraer características cada vez más complejas y relacionadas con las tareas directamente de los datos. Los desarrollos recientes en el diseño y la capacitación de la arquitectura de redes neuronales han permitido a los investigadores resolver tareas de aprendizaje que antes no se podían resolver en el campo de la visión por ordenador. Como resultado, los enfoques basados en el aprendizaje profundo han tenido mucho éxito al abordar una amplia gama de tareas de análisis de imágenes biomédicas [34].

En este estudio nos centramos en un conjunto de datos de *whole slide images* de tejido de cáncer de próstata anotados por un experto patólogo y un no experto, para demostrar que una red neuronal convolucional puede ser entrenada con éxito como un anotador de puntuación de Gleason.

Nuestro conjunto de datos comprende diecisiete WSI de tejido, cada uno con 5000 × 5000 píxeles en promedio. Los parches de tejido prostático fueron anotados por un primer patólogo delineando cuidadosamente las regiones cancerosas y de forma independiente los mismos parches fueron anotados por un anotador no experto asignando un patrón de Gleason de 3, 4 o 5 a cada región en ambos casos, de acuerdo con los criterios de la Unión Internacional contra el Cáncer (UICC) y de la OMS / ISUP. Las regiones cancerosas se etiquetaron con los patrones de Gleason correspondientes utilizando la aplicación web microDraw.

Debido al gran tamaño de las imágenes WSIs anotadas, estas se dividen en subregiones y se extraen parches del mismo tamaño. A partir de los conjuntos de parches se realiza una partición de la base de datos que nos servirá para evaluar el modelo desarrollado. Este modelo consiste en una red neuronal convolucional que utilizamos para predecir el grado de agresividad del cáncer, distinguiendo entre no canceroso, grado 3, grado 4 y grado 5.

4.2.1 Creación de parches

Partiendo de cada uno de las imágenes, se extraen parches que fueron etiquetados de acuerdo con la anotación en su región central de 250 × 250. Se excluyeron del estudio los parches que contenían artefactos, tejido no prostático o aquellas con menos del 20 % de tejido en la imagen. El porcentaje de tejido se obtuvo mediante la máscara binaria obtenida mediante el método de Otsu.

Posteriormente se realizó una umbralización de las máscaras de los parches. A cada parche se le asignó una etiqueta entre no canceroso, grado 3, grado 4 o grado 5 según las anotaciones mayoritarias realizadas en el mismo. En caso de que un parche no presentara anotaciones, se le asignó la etiqueta de no canceroso. Con ello se obtuvo una base de datos con la clasificación de los parches dependiendo de la etiqueta que se le ha asignado en la anotación, estableciendo las clases: NC, G3, G4, G5.

4.2.2 Balanceo y división de las muestras en train/validation/test

Para poder hacer un estudio estadístico robusto con las características extraídas de las imágenes es necesario equilibrar el número de muestras en cada uno de los grupos (entrenamiento, validación y test), para poder realizar el entrenamiento de la red neuronal y el posterior testeo de esta.

La clase limitante son el número de muestras de cada clase (NC, G3, G4, G5). El objetivo es, por tanto, clasificar cada parche en un grupo de forma que en cada uno de los grupos quede la misma cantidad de todas las clases. Para ello se parte de la etiqueta de los parches y se utilizan un 60 % de las muestras como entrenamiento, otro 20 % para la validación del entrenamiento y se reserva un 20 % de muestras nuevas que pertenecerán al grupo de test (véase 4.2), utilizando el método hold out.

En este método, se entrena el modelo con los datos seleccionados para entrenamiento, ajustando sus parámetros con los datos de validación, y finalmente evaluando su rendimiento con el conjunto de datos de prueba que hemos dejado aparte. Este ultimo paso se realiza para que, a la hora de evaluar el modelo, este no aprenda de las muestras que ya ha visto y se pueda evaluar la calidad real del método desarrollado. Por lo tanto, estas muestras se usan como muestras nuevas que se quieren clasificar [35]. La división de los datos se realiza tratando de mantener en la medida de lo posible el balance de las clases entre cada conjunto.

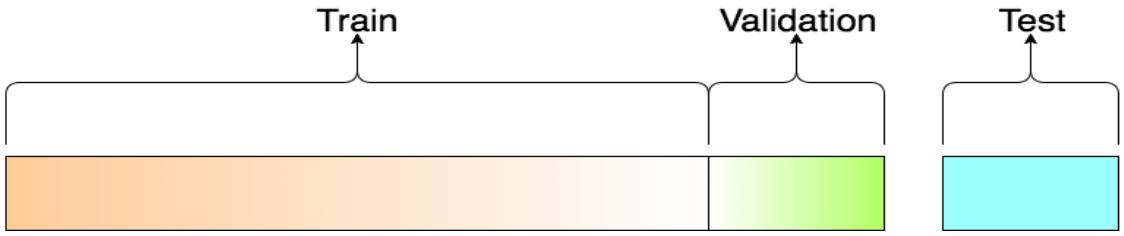

Figura 4.2: Visualización de las divisiones

4.2.3 Deep learning

El *Deep learning* (aprendizaje profundo) es una nueva técnica dentro del aprendizaje automático (*machine learning*) basado en arquitecturas de redes neuronales. Está relacionado con algoritmos inspirados en la estructura y función del cerebro. En el campo de las imágenes, uno de los principales algoritmos que ha contribuido en el desarrollo y perfeccionamiento del campo de visión artificial son las redes neuronales convolucionales (CNN), que están diseñadas para el reconocimiento y clasificación de imágenes [36], [37].

La CNN es un tipo de Red Neuronal Artificial con aprendizaje supervisado que procesa sus capas imitando al cortex visual del ojo humano para identificar distintas características en las entradas que en definitiva hacen que pueda identificar objetos y clasificarlos. El aprendizaje supervisado consiste en que cada una de las entradas que forman la base de datos con la que se entrena el algoritmo de aprendizaje automático está correctamente etiquetada con la categoría que queremos obtener tras la clasificación, por lo que todas las imágenes de entrada deben estar etiquetadas previamente [38].

La visión humana se basa principalmente en la detección de bordes, y está organizada de forma que, dado un patrón, se activan distintos grupos de neuronas conectados entre sí, los cuales hacen posible la comprensión del patrón y la asociación con su significado. Para ello, la CNN contiene varias capas ocultas especializadas y con una jerarquía: esto significa que las primeras capas detectan propiedades o formas básicas y se van especializando hasta llegar a capas más profundas capaces de reconocer formas complejas como un rostro o una silueta, como vemos en la figura 4.3.

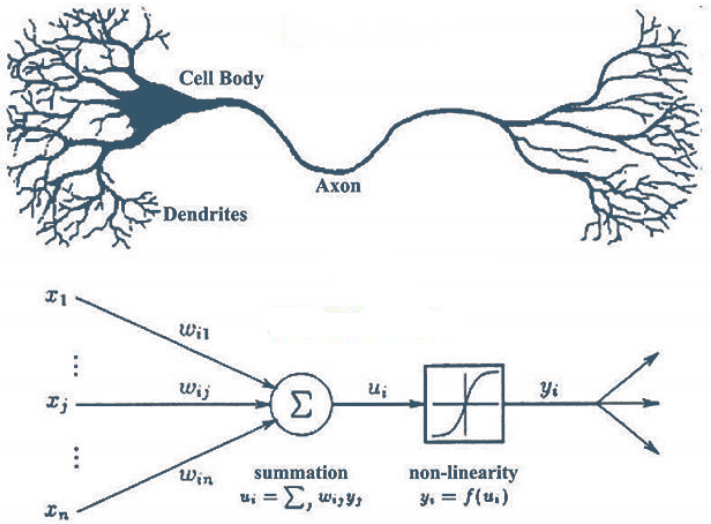

Figura 4.3: Comparación entre neurona y red neuronal

El problema de aprendizaje de una red neuronal se formula en términos de minimización de la función de pérdidas asociada para obtener los pesos óptimos para los filtros de las capas convolucionales. El valor de esta función depende totalmente de los parámetros de la red neuronal: los pesos sinápticos y los bias asociados. Estos a su vez dependen de la arquitectura del modelo, de los hiperparámetros escogidos y de la base de datos empleada.

Estas redes buscan aprender desde cero de forma automática estas conexiones y su significado correspondiente. Para el aprendizaje, toman un grupo de imágenes etiquetadas como entrada, asignándole importancias (pesos) a ciertos elementos en la imagen para así poder diferenciar unos de otros. Este es un proceso costoso, por lo cual se necesitan una enorme cantidad de datos etiquetados, y potentes unidades computacionales.

En este trabajo se utilizan las CNN para clasificar el grado de cáncer de una imagen procedente de una biopsia de cáncer de próstata, por ello podemos utilizar las CNN para que reconozcan patrones en las imágenes. Se toma como entrada del algoritmo las imágenes de la base de datos, reescaladas a una matriz de dimensiones 224 × 224 de ancho y alto para disminuir la memoria necesaria y 3 canales de color, obteniendo una matriz tridimensional (224×224×3) que contiene los valores numéricos de los píxeles.

Las CNNs se conforman de dos etapas: extracción de características y clasificación. En la etapa de extracción de características, como su nombre lo indica, se extraen las características que poseen los datos de entrada o entrenamiento (usualmente imágenes), formando los mapas de características (*feature maps*), los cuales van aumentando en número y disminuyendo en tamaño

a medida que haya más capas ocultas en esta etapa. Además, esta consta de dos capas, la capa convolucional y la capa de *subsampling* o *pooling*, las dos capas van en cascada y, por lo general, no se suelen separar. Una vez los mapas de características han sido extraídos, una red *fully connected* como clasificadora, se encarga de separar y clasificar cada característica [39].

Las capas ocultas que forman las CNN y realizan las transformaciones necesarias para realizar la clasificación, son: capas convolucionales, *pooling*, de activación y *fully-connected*.

- Capas convolucionales:

  El objetivo inicial de las capas convolucionales es resaltar los bordes y curvas de la imagen. Para ello se efectúan una serie de productos y sumas entre la matriz de partida y una matriz kernel y mediante el submuestreo se reduce la dimensión de la matriz de entrada dividiéndola en subregiones para generalizar las características.

  Combinando capas convolucionales, se busca obtener un mayor nivel de abstracción, detectando objetos. Los filtros se van moviendo a través de la imagen, realizando una operación de convolución y transformando esa zona en un píxel con el valor resultado de la operación de la convolución a la salida de la capa. Los filtros aplicados tienen unos pesos que afectan a la operación realizada. El tamaño de paso define cuántas columnas o filas nuevas afecta el filtro en cada convolución.

  Utilizando capas convolucionales una detrás de otra, de forma que la salida de una es la entrada de la siguiente (véase 4.4) , se consigue una mayor precisión en la detección de patrones, bordes y relación entre ellos.

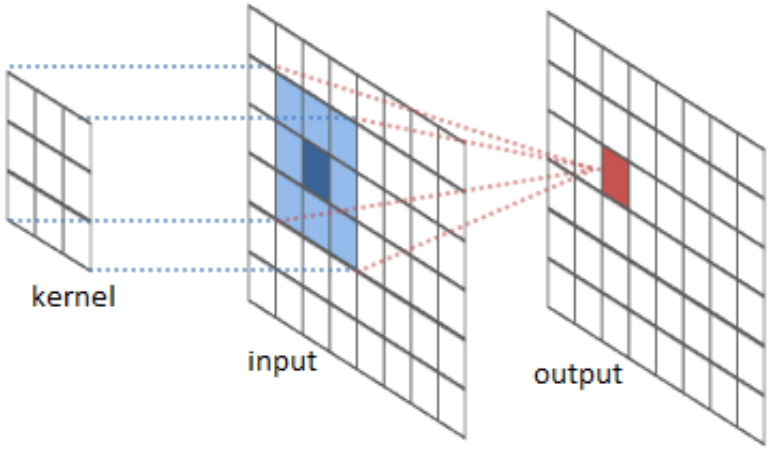

Figura 4.4: Capa convolucional

  El tamaño del mapa de salida es definido por la siguiente ecuación, donde M es la cantidad de mapas de una determinada capa de igual tamaño $(M_x, M_y), (K_x, K_y)$ es el tamaño del kernel utilizado, (S_x, S_y) son los factores de desplazamiento que indican cuantos píxeles en x y en y se salta u omite entre cada convolución, teniendo en cuenta que el kernel siempre debe quedar dentro del mapa de características o imagen. n Es el índice de la capa presente.

$$M_x^n = \frac{M_x^{n-1} - K_x^n}{S_x^n + 1} + 1 \tag{4.1}$$

$$M_y^n = \frac{M_y^{n-1} - K_y^n}{S_y^n} + 1 \tag{4.2}$$

- Capas de activación:

  A continuación de cada capa convolucional suele seguirle una capa de activación. Esta aporta términos no lineales al proceso, complementando las operaciones lineales de convolución. Una de las más utilizadas es la ReLU (unidad de rectificación lineal). Esta realiza una operación de saturación para valores menores que 0 en la salida de la capa convolucional. Su objetivo principal es eliminar todos los valores negativos de la convolución. Todos los valores positivos siguen siendo los mismos, pero todos los valores negativos se cambian a cero [40].

$$f_A(x) = max(0, x) \tag{4.3}$$

- Capa de *pooling*:

  Este tipo de capa busca disminuir la dimensionalidad de la salida de una capa convolucional resumiendo las respuestas de las capas cercanas. Este agrupamiento es beneficioso ya que agrupa zonas de la imagen, reduciendo el coste computacional y previniendo el sobreajuste. Uno de los métodos de agrupamiento, utilizado en el presente trabajo, es el de *max-pooling*. En este, a partir de una ventana de dimensión K _ K, con cierto paso, se cogen tramos de la imagen de la capa previa, pasando la región a un píxel de valor el máximo de la zona a la salida de la capa de agrupamiento. Estas operaciones se realizan en una región cuadrada, la cual no se puede solapar [41].

  Esta capa nos permite reducir el tamaño espacial de los datos y ayuda a obtener una representación invariante a traslaciones de entrada (véase 4.5).

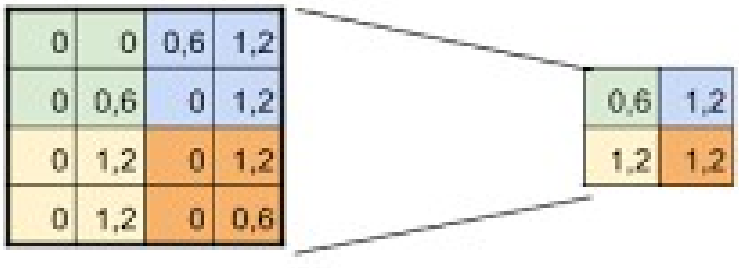

Figura 4.5: Ejemplo de max pooling

- Capas fully-connected:

  El objetivo de esta capa es utilizar los resultados obtenidos del proceso de convolución / agrupación para clasificar la imagen en una etiqueta. La salida de convolución / agrupación nos ofrece un mapa de activación, que se aplana en un solo vector de valores, cada uno de los cuales representa una probabilidad de que cierta característica pertenezca a una etiqueta [37].

  Esto se consigue aplicando una convolución unidimensional con pesos al resultado de la capa previa, reduciendo las activaciones a un número N de neuronas presentes en la capa. Para aumentar la capacidad de aprendizaje y relación de patrones se pueden enlazar unas capas completamente conectadas con otras en serie. Tras la última capa totalmente conectada, la cual debe tener tantas neuronas como etiquetas posibles, se aplica una función conocida como *softmax*, la cual convierte los valores de salida para acotarlos en el intervalo de probabilidades [0,1], bajo la premisa de que las probabilidades de todas las neuronas sumen 1. Finalmente, existe una capa de clasificación que obtiene como salida la etiqueta con mayor probabilidad. De esta manera, la salida de cada neurona o nodo es computada por medio de la siguiente ecuación:

$$y_j = f(\sum_i w_{ij} x_i + b_j) \tag{4.4}$$

  Para mejorar los resultados del algoritmo y evitar el sobreajuste del modelo, una de las técnicas utilizadas es la de *dropout* [42] en capas *fully-connected*. En cada iteración durante la etapa de entrenamiento, esta técnica desactiva aleatoriamente cierto porcentaje de las neuronas en la capa oculta, acorde a una probabilidad de descarte previamente definida. Lo que se consigue con esto es que ninguna neurona memorice parte de la entrada, que es precisamente lo que sucede cuando tenemos sobreajuste.

En este trabajo, se parte de un modelo ya entrenado en la conocida base de datos de referencia ImageNet (VGG-16), el cual ajustamos a partir de los parámetros que este ya ha aprendido.

VGG16 es una arquitectura de red neuronal de convolución (CNN) que se utilizó para ganar la competencia ILSVR en 2014. Se considera una de las mejores arquitecturas de modelos de visión hasta la fecha. Lo más característico de VGG16 es que, en lugar de tener una gran cantidad de hiperparámetros, se centraron en tener capas de convolución de filtro 3x3 con zancada 1 y siempre usaron el mismo relleno y capa *maxpool* del filtro 2x2 de zancada 2. Sigue esta disposición de convolución y capas *max-pooling* consistentemente en toda la arquitectura. Al final tiene 2 FC (capas *fully conected*) seguido de un *softmax* para salida. El número 16 se refiere a que tiene 16 capas que tienen pesos. Esta red es una red bastante grande y tiene aproximadamente 138 millones de parámetros [23]. En la figura 4.6 podemos ver la arquitectura de la red.

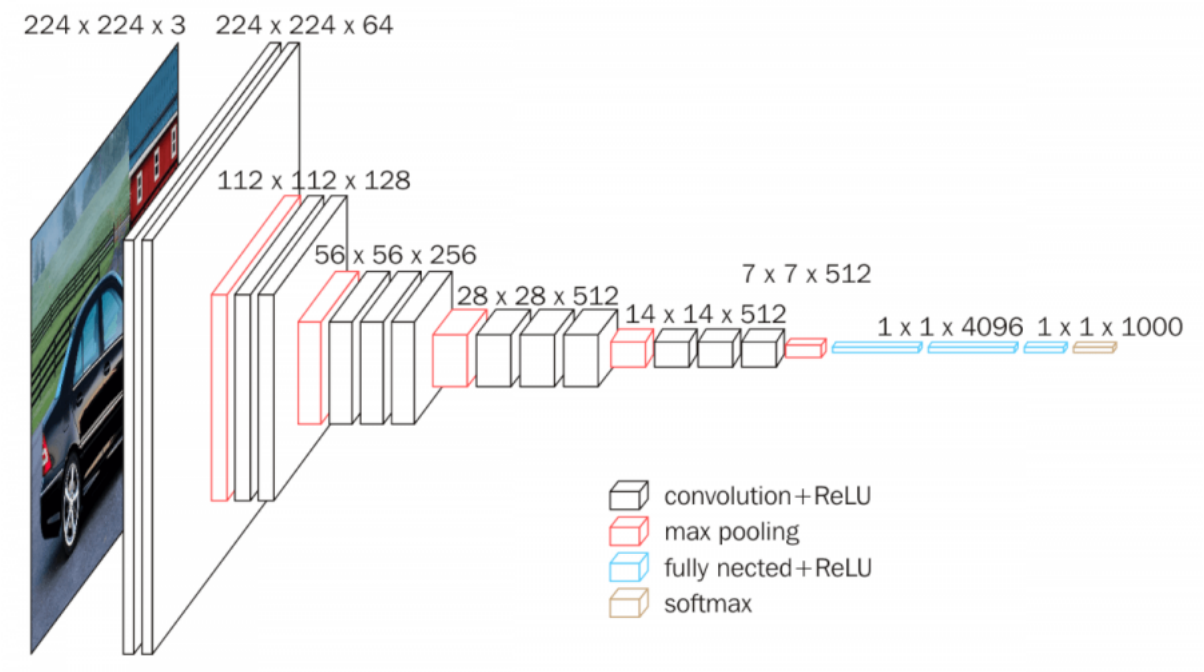

Figura 4.6: Arquitectura red Vgg16 [43]

Utilizamos la parte convolucional de la arquitectura del modelo, eliminando todas las capas completamente conectadas y tras la última capa convolucional agregamos una capa de agrupación promedio global, seguida de una capa de clasificación final que usa la no linealidad softmax.

Una vez se tiene definida la estructura de la CNN, se procede a entrenar la misma. El entrenamiento tiene como objetivo ajustar los pesos de las neuronas y filtros de la red. En cada iteración, se divide un grupo de imágenes de forma aleatoria. Se realizan iteraciones hasta haber entrenado la red con todas las imágenes tantas veces como se indique. Al comenzar el proceso de entrenamiento se tienen las imágenes y sus etiquetas asociadas a unos pesos aleatorios, posteriormente se realiza un ajuste de los pesos mediante una técnica de retropropagación, a partir de una función de coste. Para cada iteración, se pasan las imágenes por la red, y se obtienen sus etiquetas predichas. La arquitectura final se muestra en la Tabla 4.2

Layer	Type	Output shape	Parameters
Block1-conv1	Conv2D	(224, 224, 64)	1792
Block1-conv2	Conv2D	(224, 224, 64)	36928
Block1-pool	MaxPooling2D	(112, 112, 64)	0
Block2-conv1	Conv2D	(112, 112, 128)	73856
Block2-conv2	Conv2D	(112, 112, 128)	147584
Block2-pool	MaxPooling2D	(56, 56, 128)	0
Block3-conv1	Conv2D	(56, 56, 256)	295168
Block3-conv2	Conv2D	(56, 56, 256)	590080
Block3-conv3	Conv2D	(56, 56, 256)	590080
Block3-pool	MaxPooling2D	(28, 28, 256)	0
Block4-conv1	Conv2D	(28, 28, 512)	1180160
Block4-conv2	Conv2D	(28, 28, 512)	2359808
Block4-conv3	Conv2D	(28, 28, 512)	2359808
Block4-pool	MaxPooling2D	(14, 14, 512)	0
Block5-conv1	Conv2D	(14, 14, 512)	2359808
Block5-conv2	Conv2D	(14, 14, 512)	2359808
Block5-conv3	Conv2D	(14, 14, 512)	2359808
Block5-pool	MaxPooling2D	(7, 7, 512)	0

Tabla 4.2: Arquitectura del modelo

A continuación, con la función de coste se compara el resultado obtenido con las etiquetas reales. En este trabajo se utiliza como función de coste la entropía cruzada categórica:

$$\pounds(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{i_j}\log(p_{i_j}) \tag{4.5}$$

Con el resultado de la función de coste, se analiza por retropropagación la dependencia del resultado con cada uno de los pesos. Al tener funciones diferenciables, se pueden determinar las variaciones en la función de coste en función de las variaciones de los pesos, con esta información se van actualizando los pesos en dirección contraria al gradiente de la función de coste. El ajuste de pesos también puede calibrarse a partir del ratio de aprendizaje [44].

El ratio de aprendizaje es un hiperparámetro, que junto a otras variables debe ser especificado por el programador para ajustar los algoritmos de entrenamiento, variaciones de estos hiperparámetros puede tener un gran impacto en el rendimiento de la red. Los hiperparámetros modificados en este estudio son [45]:

- Número de epochs o iteraciones. Este hiperparámetro indica el número de veces que el conjunto de datos de entrenamiento ha pasado por la red neuronal durante el proceso de aprendizaje. Una época ha concluído cuando un subconjunto de datos finaliza la etapa *forward-backward* y, en la siguiente época, se realizará el mismo procedimiento pero intentando disminuir la función de error con los pesos actualizados de la época anterior.

- *Batch size*. El *batch size* es el argumento que indica el número de muestras que se agrupan dentro del mismo lote (partición del conjunto de datos que se pasa por la red). Cuando todas las muestras de un *batch size* han pasado por la etapa *forward* y *backward propagation* se dice que han cumplido una época. El tamaño óptimo depende de muchos factores cómo la capacidad de memoria disponible.
- *Learning rate* o ratio de aprendizaje. Se trata de un hiperparámetro que controla cuánto se están ajustando los pesos de la red en cada iteración con respecto al gradiente de pérdida. De forma general cuanto más bajo sea el valor, se desciende más lento y se puede caer en un mínimo local, mientras que cuanto más alto sea, se avanza más rápido (se realizan cambios más grandes en los pesos), pero puede que se salte el mínimo objetivo y dificultar el proceso de aprendizaje.
- Optimizador. Se trata de un parámetro necesario a la hora de compilar y entrenar el modelo. De forma general, el proceso de entrenamiento es un un problema de optimización y, por ello, necesita de un optimizador para actualizar los pesos de la red.

4.2.4 Implementación de una red neuronal híbrida con anotaciones del experto y no experto

Con el fin de obtener una clasificación más precisa, debido a la subjetividad que supone la tarea de anotar y etiquetar las muestras, partimos de la red ya diseñada y modificamos la entrada que recibe el modelo, de forma que obtenemos un modelo híbrido.

Este modelo híbrido permite que el modelo reciba como *dataframe* tanto las imágenes anotadas por el experto como por el no experto, estas imágenes tienen las mismas dimensiones y aportan información complementaria para el entrenamiento, validación y test del modelo. El modelo recibe como entradas tanto los parches como las clases en las que se clasifica en *ground truth*. Para unir las dos entradas utilizamos la capa Concatenate, que nos permite concatenar los datos, obteniendo un modelo con la arquitectura que podemos ver en la tabla 4.3.

Los *dataframes* tanto del anotador experto como del no experto han sido codificados mediante una técnica de *one-hot-encoding*, en forma de vector. Este vector contiene 1 en la clase anotada, y 0 en el resto para cada parche.

Layer	**Type**	**Output shape**	**Parameters**	**Connected to**
imagen de entrada	InputLayer	(224, 224, 3)	0	
vgg16	Model	(7, 7, 512)	14714688	imagen de entrada
global-average pooling	GlobalAverage Pooling2D	(, 512)	0	vgg16
reshape1	Reshape	(, 1, 512)	0	global-average pooling2D
anotaciones no experto	InputLayer	(, 1, 4)	0	
concatenate	Concatenate	(, 1, 516)	0	reshape1 anotaciones no experto
flatten	Flatten	(, 516)	0	concatenate
predictions	Dense	(, 4)	2868	flatten

Tabla 4.3: Arquitectura del modelo híbrido

Partiendo de la concatenación de ambas entradas, antes de aplicar la discriminación, se unen las respuestas por una capa clasificadora común, obteniendo una única salida que es la clasificación de las imágenes. Esta clasificación es evaluada a partir del *dataframe* de ambos anotadores.

Capítulo 5

Experimentos y Resultados

5.1 Estrategia experimental y métricas utilizadas

Como se ha expuesto anteriormente, la finalidad de este proyecto es conseguir un sistema de clasificación semi-automático de tejido canceroso a partir de imágenes histopatológicas de la próstata. A diferencia de los estudios anteriores, entrenamos y evaluamos nuestro modelo sobre la base de anotaciones manuales detalladas de un experto y un no experto conjuntamente.

Para llevar a cabo la estrategia experimental, se han anotado las imágenes histopatológicas generando un *ground truth* no experto. Partiendo de este y del *ground truth* correspondiente a las anotaciones del experto histopatológico, que podemos ver en la tabla 5.1, se ha realizado el preprocesado y la partición de ambas bases de datos mediante el método *hold out* como se ha explicado en el apartado 4.2.2.

	Nº pacientes	Parches tejido benigno	Parches tejido patológico grado 3	Parches tejido patológico grado 4	Parches tejido patológico grado 5
Anotación experto	17	3546	293	255	308
Anotación no experto	17	3219	293	703	187

Tabla 5.1: Número de pacientes e imágenes de cada tejido empleadas en este proyecto

En este trabajo se va a realizar una partición de la base de datos por pacientes, de modo que no se utilicen las imágenes de un mismo paciente para el entrenamiento y la evaluación del modelo, ya que esto hará que el modelo se sobreajuste. Además, buscamos que la información obtenida de ambas bases de datos (anotador experto y no experto) pueda utilizarse conjuntamente para obtener un modelo mejor, por ello en ambas bases de datos un mismo paciente se encontrará en el mismo grupo de entrenamiento, validación o test.

De esta manera, tras aplicar la partición las bases de datos se dividen en tres grupos diferentes:

- Entrenamiento: Contiene el 60 % de las muestras. Estas imágenes son el conjunto de datos real que utilizamos para entrenar el modelo.
  - En la base de datos anotada por el experto, se tienen 3751 muestras.
  - En la base de datos anotada por el no experto, se tienen 3751 muestras.
- Validación. Contiene el 20 % de las muestras. Estas imágenes proporcionan una evaluación imparcial del modelo de entrenamiento ajustando los hiperparámetros del modelo.
  - En la base de datos anotada por el experto, se reservan 325 imágenes.
  - En la base de datos anotada por el no experto, se tienen 325 muestras.
- Test. Contiene el 20 % de las muestras. Estas imágenes se van a emplear para evaluar imparcialmente el modelo final, midiendo la calidad del algoritmo.
  - En la base de datos anotada por el experto, se reservan 391 imágenes.
  - En la base de datos anotada por el no experto, se reservan 391 muestras.

Una vez realizada la partición, se obtiene un *ground truth* para cada base de datos, que será utilizado como entrada para entrenar al modelo y obtener una clasificación de las imágenes. Tras entrenar el modelo, para poder evaluar su precisión se han escogido la métrica de estadística de Kappa de Cohen y la matriz de confusión.

El coeficiente Kappa de cohen se usa ampliamente para medir el acuerdo entre evaluadores. Esta es una medida estadística que mide la concordancia entre dos examinadores en sus correspondientes clasificaciones, tiene en cuenta la posibilidad de que se produzca un acuerdo por casualidad, lo que da como resultado un valor de 0 para el acuerdo que se produce por casualidad y un valor de 1 para un acuerdo perfecto. Para las clases ordenadas, el coeficiente ponderado de kappa cohen es más apropiado porque penaliza más fuertemente el desacuerdo entre anotadores que ocurre entre clases más distantes. Aquí, utilizamos la estadística kappa ponderada cuadrática definida de la siguiente manera:

$$kappa = 1 - \frac{\sum_{i_j} w_{i_j}, O_{i_j}}{\sum_{i_j} w_{i_j}, E_{i_j}}, w_{i_j} = \frac{(i - j^2)}{(N-1)^2} \tag{5.1}$$

donde N es el número total de clases consideradas o puntuaciones de calificación y los índices i , j se refieren a las puntuaciones de calificación ordenados $1 \leq i, j \leq N$, O_{i_j} denota el número de imágenes que recibieron la calificación i por el primer experto y la calificación j por el segundo y E_{i_j} denota el número esperado de imágenes que reciben la calificación i por el primer experto y la calificación j por el segundo, suponiendo que no hay correlación entre las puntuaciones de calificación.

La matriz de confusión es una herramienta que permite la visualización del éxito de la clasificación de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la

clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo las diferentes clases o resultados de la clasificación. Donde la diagonal corresponde a los valores bien detectados y los demás corresponden a los falsos positivos o negativos. Por lo que se busca que la mayor parte de las muestras se encuentren en la diagonal [46].

En el experimento de modelo con entrada de un solo anotador utilizaremos únicamente la base de datos del experto anotador para obtener una clasificación automática de las imágenes. Mientras que en el experimento de modelo con entrada multianotador, se utiliza un modelo híbrido en el cual se pasan como entrada al modelo la base de datos del experto juntamente con la del no experto, de forma que el modelo se entrena con las mismas imágenes que el experimento anterior, pero con el doble de información, con lo cual se presume que obtendremos un modelo que realice una clasificación más precisa.

El resultado obtenido de las particiones de la base de datos se muestran en la tabla 5.2 para el anotador experto y en la tabla 5.1 para el anotador no experto.

Grupo	**NC**	**G3**	**G4**	**G5**
Train	3149	203	149	185
Validation	177	52	39	57
Test	220	38	67	66

Tabla 5.2: Groundtruth experto

Grupo	**NC**	**G3**	**G4**	**G5**
Train	2904	207	448	127
Validation	109	66	108	42
Test	206	20	147	18

Tabla 5.3: Groundtruth no experto

5.2 Comparativa anotador experto contra no experto

En primer lugar, se realiza la comparación entre la anotación del experto y del no experto en la cohorte de test, para comprobar si existe una gran diferencia en las clases anotadas y si esta puede afectar al comportamiento del modelo híbrido. Con este experimento, se quiere comprobar la capacidad de un ingeniero biomédico con conocimientos básicos, al cual se le da una formación de histología de próstata, de detectar patrones cancerosos de forma correcta. La matriz de confusión obtenida para la comparación de ambos anotadores se presenta en la figura 5.1.

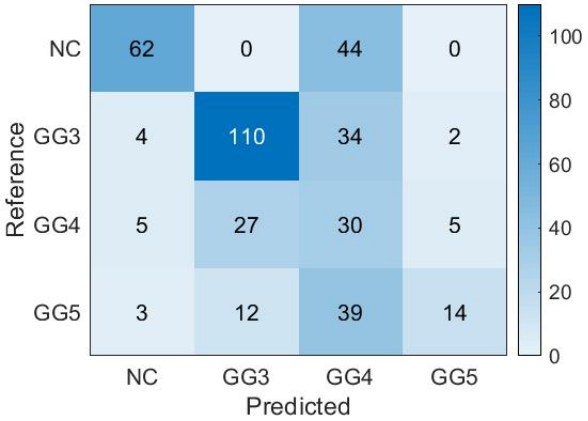

Figura 5.1: Comparación de los datos anotados por el experto y el no experto

Donde observamos que existe un alto grado de acuerdo entre los anotadores, aunque se pueden ver diferencias en la anotación principalmente en la predicción del no experto a la hora de anotar el grado 4 de Gleason. Obteniendo un coeficiente kappa de cohen de 0,440.

5.3 Modelo basado en CNNs con entrada de un solo anotador

En este experimento utilizamos únicamente las anotaciones del experto patólogo para entrenar al modelo.

En cuanto al diseño del modelo, que hemos visto en 4.2.3, en este estudio se ha utilizado una arquitectura de red basada en VGG16, para la cual se han ajustado los hiperparámetros ya explicados, en función de las necesidades del modelo utilizando los conjuntos de entrenamiento y validación.

Se ha escogido el optimizador Adam, este es un método de *learning rate* adaptativo que adapta el ratio de aprendizaje en función de cómo estén distribuidos los parámetros. Si los parámetros están muy dispersos (*sparse*), el ratio de aprendizaje aumentará. Este método busca solucionar el problema de otros optimizadores en los que no se indica la distancia entre los pesos. De esta manera, en el inicio se actualizarán los pesos de forma más abrupta e irán más rápido en la disminución de la función de coste. Para implementarlo se ha aplicado un *learning rate* de 1e-4, utilizando un *batch size* de 32 puesto que es el valor óptimo a nivel computacional y el número de épocas que se han utilizado para entrenar el modelo es de 100.

La matriz de confusión obtenida al aplicar el modelo con los datos del conjunto de test del experto, utilizando el mejor modelo obtenido en el conjunto de validación se presenta en la figura 5.2.

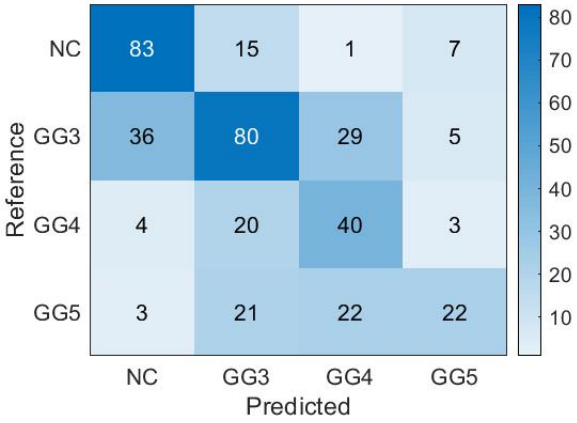

Figura 5.2: Predicción del modelo con datos del experto

Aquí podemos ver que existe un gran porcentaje de acuerdo entre el experto y el modelo, con una exactitud del 57 %, donde la mayoría de las clasificaciones erróneas se encuentran dentro de los patrones vecinos de Gleason. Un 28 % de los parches anotados como no canceroso habían sido predichos como Gleason 3, pero solo el 3 % se pronostican como patrones de Gleason 4 o 5. Esto nos indica que existe una alta probabilidad de que el modelo acierte en las predicciones sobre la anotación. Para esta comparación, el kappa cuadrático de cohen obtenido es de 0,571.

5.4 Modelo con entrada multianotador

Finalmente, para llevar a cabo este segundo estudio se utilizan como entrada las anotaciones del experto y no experto de forma conjunta, modificando el modelo como se ha explicado en 4.2.4 para convertirlo en un modelo híbrido, utilizando los mismos hiperparámetros explicados en el apartado anterior. De manera que este modelo tiene más información de la que aprender sin necesidad de introducir un mayor número de imágenes de entrada.

La matriz de confusión obtenida tras la implementación del modelo híbrido con los datos de entrada del conjunto de test de ambos anotadores se presenta en la figura 5.3.

Estos resultados reflejan que existe mejor predicción por parte del modelo que en el estudio anterior, mejorando la exactitud hasta un 64 %, ya que como podemos ver los errores del modelo se encuentran principalmente en los vecinos cercanos y son cuantitativamente menores que los del experimento anterior, excepto en el caso del grado 4 donde un 65 % de las muestras fueron predichas como grado 3, esto podría deberse a las diferencias de anotación entre el experto y no experto en este grupo, que hemos visto anteriormente. El valor kappa obtenido es de 0,685. Este valor junto con los buenos resultados ofrecidos por la matriz de confusión nos llevan a afirmar que este modelo es considerablemente más preciso que el anterior, gracias a la introducción de un segundo anotador.

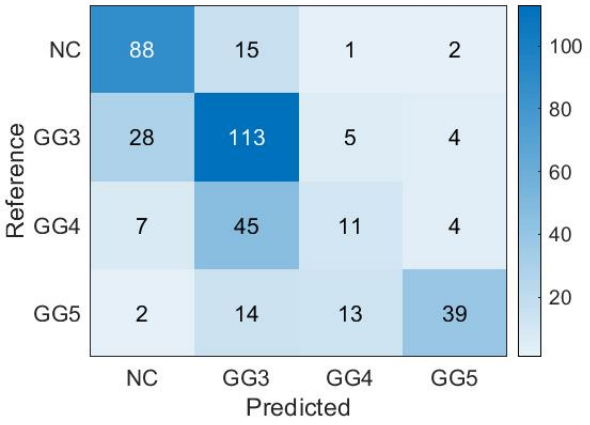

Figura 5.3: Predicción del modelo híbrido con datos del experto y no experto

5.5 Comparación y estado del arte

Los resultados de nuestros experimentos mostraron diferencias en la precisión y exactitud de los modelos al utilizar una o dos anotaciones como entrada. Esto demuestra la importancia de evaluar el rendimiento de la clasificación entre múltiples anotadores en lugar de uno solo.

Basándonos en la literatura existente vemos que los resultados obtenidos apoyan la idea defendida por Nir G, et al. sobre la importancia de evaluar los datos de más de un experto para reducir la variabilidad interobservador y aumentar la precisión del modelo sin añadir nuevas imágenes. Además, estos resultados son muy cercanos a los obtenidos por Arvaniti, et al. en el estudio de clasificación automática de Gleason en microarrays de tejido de cáncer a partir, donde se estudia el comportamiento del modelo tomando en cada estudio las anotaciones de un experto. En la tabla 5.5 podemos ver una comparación entre estos estudios.

Fuente	**Tipo de clasificación**	**Datos**	**Multiples Anotadores**	**Resultados (Kappa Cohen)**
Arvaniti et al. (2018)	Grados de Gleason 3, 4 y 5	886 pacientes	No	0,49-0,55
Nir G et al. (2018)	Grados de Gleason 3, 4 y 5	231 pacientes, 16 000 parches	Sí (entrada única: voto mayoritario entre 6 expertos)	0,30-0,58 para un anotador 0,60 voto mayoritario
Beser et al. (2020)	Benigno y grados de Gleason 3, 4 y 5	17 pacientes, 4 402 parches	Sí (entrada mixta: modelo híbrido	0,57 para un anotador 0,68 multianot.

Tabla 5.4: Comparación estado del arte

Los resultados obtenidos con el modelo utilizando únicamente la anotación del experto como entrada son similares a los de la literatura previa. No obstante, al introducir la segunda anotación del no experto los resultados mejoran notablemente, obteniendo una clasificación mucho más precisa.

Capítulo 6

Conclusiones

En este estudio, nos centramos en un conjunto de datos bien anotado de microarrays de tejido de cáncer de próstata y demostramos que una red neuronal convolucional puede ser entrenada con éxito como un anotador de puntuación de Gleason.

Primero se ha estudiado la capacidad de un anotador no experto de clasificar imágenes de tejido prostático según la literatura aprendida sobre los grados de Gleason. A continuación se ha hecho uso de las redes neuronales convolucionales para realizar un modelo semi automático mediante aprendizaje supervisado que sea capaz de clasificar imágenes que no ha visto nunca a partir del aprendizaje de imágenes anotadas. Finalmente se ha evaluado la precisión del modelo al introducirle el conjunto de anotaciones del no experto.

Los resultados obtenidos son satisfactorios, ya que vemos que existe un valor estadístico de kappa de cohen de 0,571 para el modelo con entradas anotadas únicamente por el experto y al introducir conjuntamente las entradas anotadas por el experto y el no experto este mejora hasta obtener un valor estadístico de 0,685. También se ha comparado el acuerdo interanotador obteniendo un valor estadístico de kappa de cohen de 0,440.

Podemos concluir que la capacidad del no experto para detectar los patrones cancerosos según su grado de Gleason es limitada, sobretodo para el grado 4, pero aún así podemos ver que tras la combinación de ambas anotaciones a la CNN, la precisión del método mejora considerablemente. Esta es una posible solución para reducir la carga de trabajo de patólogos sin que la precisión del sistema se vea tan afectada como en literatura previa.

Futuras investigaciones se centrarán en conseguir más anotaciones por parte de anotadores no expertos, para realizar una comparativa más robusta con más datos. Además, otro enfoque para usar datos de múltiples anotadores podría realizarse asignando pesos a diferentes etiquetas en función de la información previa o aprendida con respecto a la calidad de las anotaciones de cada experto, por ejemplo utilizando la combinación de anotaciones realizadas por distintos no expertos a modo de *crowd-sourcing*.

Parte II

Presupuesto

Capítulo 7

Presupuestos

El objetivo de este capítulo es presentar una valoración económica del proyecto realizado, utilizando la herramienta Arquímedes aprendida durante las prácticas de grado. El presupuesto del TFG basado en el desarrollo de un sistema de clasificación semiautomática de WSI se separa en distintas secciones.

7.1 Presupuestos parciales

El informe de presupuestos parciales consta de tres cuadros de precios: costes de mano de obra, costes de maquinaria y costes de materiales, que han sido utilizados en el TFG.

7.1.1 Mano de obra

A continuación, se describen los recursos humanos necesarios para el desarrollo del proyecto. Se realiza una estimación de los costes en función del tiempo dedicado al trabajo, teniendo en cuenta que este proyecto se ha llevado a cabo a lo largo de 6 meses.

Se considera la contribución de: D^{a}. Valery Naranjo Ornedo (como directora del proyecto y tutora), D. Julio José Silva Rodríguez (como cotutor del proyecto) y D^{a}. María Beser Robles (como estudiante del Grado de Ingeniería Biomédica y autora del proyecto).

Denominación	**Uds**	**Precio Unitario**	**Horas**	**Total**
Tutora (Catedrática)	h	42,000	15,000	630,00
Cotutor (Doctorando)	h	17,200	42,000	722,40
Autora (Estudiante GIB)	h	13,000	400,000	5.200,00
		Total mano de obra:		**6.552,40**

Tabla 7.1: Cuadro de precios mano de obra

7.1.2 Maquinaria

En este apartado se detallan los costes correspondientes a los recursos hardware y software necesarios para llevar a cabo la totalidad del proyecto. Dado que estas herramientas no se han obtenido específicamente para la elaboración del trabajo, es necesario tener en cuenta el periodo de amortización para cada una de ellas. Ese periodo se relaciona con la vida útil del material y el intervalo de tiempo amortizado para cada herramienta.

Entre estas maquinarias se incluyen los recursos gratuitos utilizados, plataformas de código abierto como Python o Overleaf. También se incluyen los servicios proporcionados por el grupo CVBLab que corresponden al servidor.

Denominación	**Uds.**	**Precio unitario**	**Cant.**	**Periodo de amortización (años)**	**Intervalo amortizado (meses)**	**Total**
MicroDraw	h	0,040	30	1	2	1,20
Licencia MATLAB v.R2019	u	800,000	1	1	8	533,33
Python 3.8.1	u	0,000	1	1	2	0,00
Overleaf	u	0,000	1	1	1	0,00
Hp intel Core i5 82500U	u	550,000	1	5	8	73,33
Procesador servidor intel Core i7 @4.20GHz	u	344,000	1	5	2	11,46
Tarjeta gráfica servidor NVIDIA Titan V	u	3.300,000	1	5	2	660,00
Disco servidor SSD 250 GB	u	77,000	1	5	2	2,56
Servidor Synology CVBLab	h	0,040	60	1	8	2,40
					Costes maquinaria:	**1284,28**

Tabla 7.2: Cuadro de precios de maquinaria

7.1.3 Materiales

Para llevar a cabo este presupuesto se ha tenido en cuenta, por una parte, el coste de la realización de una prueba de biopsia, a partir de la cual se obtienen las imágenes con las que se trabaja en este proyecto. El número de biopsias se toma como el número de pacientes que han sido analizados en este estudio.

Por otra parte, se considera el coste que supone la adquisición de cada muestra relativa al corte histológico donde se presentan las imágenes de tejido prostático teñidas con hematoxilina y eosina (H&E). En este caso, las muestras corresponden a cada uno de los *slides* donde se fijan las secciones de tejido extraídas al realizar la biopsia.

Denominación	**Uds**	**Precio unitario**	**Cantidad**	**Total**
Biopsia	u	600,000	17,000	10.200,00
Muestras	u	10,00	17,000	170,00
		Total materiales:		**10.370,00**

Tabla 7.3: Cuadro de precios de materiales

7.2 Presupuesto total

Para el cálculo del presupuesto total del proyecto, es necesario tener en cuenta los capítulos en los cuales se ha dividido este proyecto para obtener el presupuesto de ejecución material.

Además, se deben añadir el porcentaje de gastos generales (13 %) y el asociado al beneficio industrial (6 %). A continuación, se añadirá al precio total bruto el impuesto del IVA (21 %), obteniendo como resultado el presupuesto de ejecución por contrata. Con todo ello tendríamos el presupuesto total que supondría la realización del presente TFG.

Capítulo	**Importe**
1. Definición del proyecto	494,81
2. Investigación del estado del arte	5.865,85
3. Adquisición de muestras	11.227,00
4. Desarrollo del modelo	2.539,98
5. Redacción y defensa del TFG	3.071,46
Presupuesto de ejecución material	**23.199,10**
13 % de gastos generales	3.015,88
6 % de beneficio industrial	1.391,95
Suma	27.606,93
21 %	5.797,46
Presupuesto de ejecución por contrata	**33.404,39**

Tabla 7.4: Cuadro de presupuestos totales

Bibliography

[1] "Datos y cifras de cáncer", *https://www.who.int/es/news-room/fact-sheets/detail/cancer*,

[2] "Las cifras del cáncer en España 2018", *https://seom.org/es/noticias/106525-las-cifras-del-cancer-en-espana-2018*,

[3] "Causas del cáncer", *https://www.aecc.es/es/todo-sobre-cancer/que-es-cancer/factores-riesgo*,

[4] "Global Cancer Observatory", *https://gco.iarc.fr/*,

[5] "Cáncer de próstata", *https://www.cancer.net/es/tipos-de-c %C3 %A1ncer/c %C3 %A1ncer-de-pr %C3 %B3stata/estad %C3 %ADsticas.*,

[6] B. Martínez-Amores Martínez, M. Durán Poveda, M. Sánchez Encinas y R. Molina Villaverde, "Actualización en cáncer de próstata", *Medicine - Programa de Formación Médica Continuada Acreditado*, vol. 11, n.º 26, págs. 1578-1587, feb. de 2013, ISSN: 03045412. DOI: 10.1016/S0304-5412(13)70509-2. dirección: https://linkinghub.elsevier.com/retrieve/pii/S0304541213705092.

[7] "Diagnóstico de cáncer de próstata para hombres, biopsia", *https://www.cancer.net/es/tipos-de-c %C3 %A1ncer/c %C3 %A1ncer-de-pr %C3 %B3stata/diagn %C3 %B3stico*,

[8] F. José Núñez Sánchez-Agustino, S. Kanaan, I. Carles y V. Royo, "Diseño de un sistema de reconocimiento automático de matrículas de vehículos mediante una red neuronal convolucional", inf. téc.

[9] "Diseño y desarrollo de un sistema automático de clasificación de estructuras glandulares en imágenes histológicas de próstata", *http://hdl.handle.net/10251/109246*,

[10] D. Komura y S. Ishikawa, "Machine Learning Methods for Histopathological Image Analysis", *Computational and Structural Biotechnology Journal*, vol. 16, págs. 34-42, 2018, ISSN: 20010370. DOI: 10.1016/j.csbj.2018.01.001. dirección: https://linkinghub.elsevier.com/retrieve/pii/S2001037017300867.

[11] "Sicap CVBLAB", *http://www.cvblab.webs.upv.es/es/project/sicap-2/*,

[12] W. C. Allsbrook, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, M. B. Amin, D. G. Bostwick, P. A. Humphrey, E. C. Jones, V. E. Reuter, W. Sakr, I. A. Sesterhenn, P. Troncoso, T. M. Wheeler y J. I. Epstein, "Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists", *Human Pathology*, vol. 32, n.º 1, págs. 74-80, ene. de 2001, ISSN: 00468177. DOI: 10.1053/hupa.2001.21134. dirección: https://linkinghub.elsevier.com/retrieve/pii/S0046817701914327.

[13] G. Nir, D. Karimi, S. L. Goldenberg, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, D. J. S. Thompson, P. C. Black y S. E. Salcudean, "Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images", *JAMA Network Open*, vol. 2, n.º 3, e190442, mar. de 2019, ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2019.0442. dirección: http://jamanetworkopen.jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2019.0442.

[14] A. Basavanhally, S. Ganesan, S. Agner, J. Monaco, M. Feldman, J. Tomaszewski, G. Bhanot y A. Madabhushi, "Computerized Image-Based Detection and Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology", *IEEE Transactions on Biomedical Engineering*, vol. 57, n.º 3, págs. 642-653, mar. de 2010, ISSN: 0018-9294. DOI: 10.1109/TBME.2009.2035305. dirección: http://ieeexplore.ieee.org/document/5306163/.

[15] A. W. Wetzel, R. Crowley, S. Kim, R. Dawson, L. Zheng, Y. M. Joo, Y. Yagi, J. Gilbertson, C. Gadd, D. W. Deerfield y M. J. Becich, "<title>Evaluation of prostate tumor grades by content-based image retrieval</title>", ene. de 1999, págs. 244-252. DOI: 10.1117/12.339826. dirección: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=975147.

[16] R. Farjam, H. Soltanian-Zadeh, K. Jafari-Khouzani y R. A. Zoroofi, "An image analysis approach for automatic malignancy determination of prostate pathological images", *Cytometry Part B: Clinical Cytometry*, vol. 72B, n.º 4, págs. 227-240, jul. de 2007, ISSN: 15524949. DOI: 10.1002/cyto.b.20162. dirección: http://doi.wiley.com/10.1002/cyto.b.20162.

[17] J. Diamond, N. H. Anderson, P. H. Bartels, R. Montironi y P. W. Hamilton, "The use of morphological characteristics and texture analysis in the identification of tissue composition

in prostatic neoplasia", *Human Pathology*, vol. 35, n.º 9, págs. 1121-1131, sep. de 2004, ISSN: 00468177. DOI: 10.1016/j.humpath.2004.05.010. dirección: https://linkinghub.elsevier.com/retrieve/pii/S0046817704003223.

[18] J. Quinlan, "Decision trees and decision-making", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, n.º 2, págs. 339-346, 1990, ISSN: 00189472. DOI: 10.1109/21.52545. dirección: http://ieeexplore.ieee.org/document/52545/.

[19] R. A. Jacobs, M. I. Jordan, S. J. Nowlan y G. E. Hinton, "Adaptive Mixtures of Local Experts", *Neural Computation*, vol. 3, n.º 1, págs. 79-87, feb. de 1991, ISSN: 0899-7667. DOI: 10.1162/neco.1991.3.1.79. dirección: https://www.mitpressjournals.org/doi/abs/10.1162/neco.1991.3.1.79.

[20] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen van de Kaa, P. Bult, B. van Ginneken y J. van der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis", *Scientific Reports*, vol. 6, n.º 1, pág. 26 286, sep. de 2016, ISSN: 2045-2322. DOI: 10.1038/srep26286. dirección: http://www.nature.com/articles/srep26286.

[21] L. Gorelick, O. Veksler, M. Gaed, J. A. Gomez, M. Moussa, G. Bauman, A. Fenster y A. D. Ward, "Prostate Histopathology: Learning Tissue Component Histograms for Cancer Detection and Classification", *IEEE Transactions on Medical Imaging*, vol. 32, n.º 10, págs. 1804-1818, oct. de 2013, ISSN: 0278-0062. DOI: 10.1109/TMI.2013.2265334. dirección: http://ieeexplore.ieee.org/document/6522505/.

[22] "H&E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer", *https://www.biorxiv.org/content/10.1101/064279v9*,

[23] "Step by step VGG16 implementation in Keras", *https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c*,

[24] "Large scale digital prostate pathology image analysis combining feature extraction and deep neural network", *https://arxiv.org/abs/1705.02678*,

[25] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff y M. Claassen, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning", *Scientific Reports*, vol. 8, n.º 1, pág. 12 054, dic. de 2018, ISSN: 2045-2322. DOI: 10.1038/s41598-018-30535-1. dirección: http://www.nature.com/articles/s41598-018-30535-1.

[26] G. Jhan y P. R. Arteaga, "Aplicación del aprendizaje profundo", inf. téc.

[27] "Mathworks Matlab", *https://es.mathworks.com/products/matlab.html*,

[28] Andrea, "Facultad de medicina de Cantabria", inf. téc.

[29] 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, mayo de 2012, ISSN: 01678655. DOI: 10.1016/j.patrec.2011.10.001. dirección: https://linkinghub.elsevier.com/retrieve/pii/S0167865511003230.

[30] O. Hassan y A. Matoso, "Clinical significance of subtypes of Gleason pattern 4 prostate cancer", *Translational Andrology and Urology*, vol. 7, n.º S4, S477-S483, sep. de 2018, ISSN: 22234683. DOI: 10.21037/tau.2018.02.06. dirección: http://tau.amegroups.com/article/view/18596/20796.

[31] J. I. Epstein, "A new contemporary prostate cancer grading system", *Annales de Pathologie*, vol. 35, n.º 6, págs. 474-476, dic. de 2015, ISSN: 02426498. DOI: 10.1016/j.annpat.2015.09.002. dirección: https://linkinghub.elsevier.com/retrieve/pii/S0242649815002023.

[32] "Nueva clasificación", *http://www.revurologia.sld.cu/index.php/rcu/article/view/270*,

[33] G. García, A. Colomer y V. Naranjo, "First-Stage Prostate Cancer Identification on Histopathological Images: Hand-Driven versus Automatic Learning", *Entropy*, vol. 21, n.º 4, pág. 356, abr. de 2019, ISSN: 1099-4300. DOI: 10.3390/e21040356. dirección: https://www.mdpi.com/1099-4300/21/4/356.

[34] "Machine Learning: Cómo Desarrollar un Modelo desde Cero", *https://medium.com/datos-y-ciencia/machine-learning-c%C3%B3mo-desarrollar-un-modelo-desde-cero-cc17654f0d48*,

[35] "Técnicas de regularización para redes", *https://medium.com/metadatos/t%C3%A9cnicas-de-regularizaci%C3%B3n-b%C3%A1sicas-para-redes-neuronales-b48f396924d4*,

[36] "Deep learning & Convolutional Neuronal Network: qué es", *https://itelligent.es/es/deep-learning-convolutional-neuronal-network-cnn-consiste*,

[37] "Fully Connected in CNN", *https://missinglink.ai/guides/convolutional-neural-networks/fully-connected-layers-convolutional-neural-networks-complete-guide/*,

[38] "Convolutional Neural Networks (CNNs / ConvNets)", *https://cs231n.github.io/convolutional-networks/*,

[39] C. F. Kweldam, G. J. van Leenders y T van der Kwast, "Grading of prostate cancer: a work in progress", *Histopathology*, vol. 74, n.º 1, págs. 146-160, ene. de 2019, ISSN: 03090167. DOI: 10.1111/his.13767. dirección: http://doi.wiley.com/10.1111/his.13767.

[40] "ReLU: Activación", *https://sitiobigdata.com/2019/06/22/relu-funciones-activacion/*,

[41] "Convolutional Neural Network Tutorial (CNN) – Developing An Image Classifier In Python Using TensorFlow", *https://www.edureka.co/blog/convolutional-neural-network/*,

[42] "Regularizando nuestra red: DropOut", *https://mc.ai/regularizando-nuestra-red-dropout/*,

[43] "Deep Learning básico con Keras (Parte 3): VGG", *https://enmilocalfunciona.io/deep-learning-basico-con-keras-parte-3-vgg/*,

[44] "Función de coste – Redes neuronales", *http://www.diegocalvo.es/funcion-de-coste-redes-neuronales/*,

[45] "Hiperparámetros", *https://www.buguroo.com/es/blog/optimizacion-de-hiperparametros*,

[46] "Machine Learning: La matriz de confusión", *https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/*,