



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



ESCUELA TÉCNICA
SUPERIOR INGENIEROS
INDUSTRIALES VALENCIA

TRABAJO FIN DE MÁSTER EN INGENIERÍA BIOMÉDICA

**DISEÑO Y DESARROLLO DE SISTEMAS
AUTOMÁTICOS DE CLASIFICACIÓN DE
TUMORES MELANOCÍTICOS
SPITZOIDES**

AUTORA: M^a Rocío del Amor del Amor

TUTOR: José Manuel Mossi García

COTUTOR: Adrián Colomer Granero

Curso Académico: 2019-2020

Agradecimientos

“ A Valery, por darme la oportunidad de adentrarme en este maravilloso mundo,
por todo lo que he aprendido, tanto en lo personal como en lo profesional,
y por toda su confianza depositada en mí.

A José, por su cariño
y por sus ganas de emprender este proyecto juntos

A Adri, por su apoyo incondicional
y por hacer que este trabajo fuera posible.

A mi familia, por ser mi mayor pilar,
por quererme tal y como soy,
y por todo el apoyo que he recibido.

A mis amigas, en especial a mis Fasiles Chasiles
porque sin ellas todo esto no sería posible.

A todo el grupo de CVBlab, por ser tan increíbles,
y por haber encontrado el verdadero compañerismo.

A mi Ricis, por no creer en las casualidades, por ser tan coherente,
por el cariño y por hacer que siempre sea la penúltima vez.

A todos, muchas gracias.”

Resumen

El cáncer de piel representa el grupo más común de neoplasias malignas en la población blanca. De hecho, en la actualidad, uno de cada tres cánceres diagnosticados es cáncer de piel. Aunque dos de los cánceres cutáneos más comúnmente diagnosticados son el carcinoma de células basales y el de células escamosas, sin duda, el cáncer de piel más agresivo y peligroso es el melanoma. El melanoma cutáneo es una neoplasia agresiva, con un elevado riesgo metastásico y una extrema complejidad genética y biológica. En la mayoría de tumores melanocíticos, es posible una distinción patológica precisa entre benigno (nevus melanocítico) y maligno (melanoma). Sin embargo, existen neoplasias melanocíticas de difícil diagnóstico clínico debido a que poseen características morfológicas ambiguas. Un ejemplo de ello son los llamados “tumores spitzoides”. Las neoplasias melanocíticas spitzoides son lesiones cutáneas poco frecuentes que constituyen problemas de diagnóstico para los dermatopatólogos. Estos tumores destacan por la proliferación de grandes melanocitos epitelioides o fusocelulares con grandes núcleos de cromatina vesicular y nucléolo prominente. La discrepancia entre su aspecto morfológico, histopatológico y su evolución supone un desafío en el diagnóstico. Esta discrepancia ha servido de estímulo para la realización de estudios moleculares que permitan comprender la biología y el comportamiento de las lesiones spitzoides. Entre estos estudios destaca la metilación del DNA, ya que es una marca distintiva en las lesiones melanocíticas y puede ser clave para su desarrollo y progresión metastásica, convirtiéndose en un marcador diagnóstico, pronóstico y terapéutico.

En el presente Trabajo Fin de Master (TFM) se pretende abordar el desarrollo de algoritmos automáticos de clasificación basados en *deep learning* con los que poder identificar las neoplasias melanocíticas spitzoides benignas y malignas y realizar el pronóstico de aquellas con potencial maligno incierto, utilizando para ello los datos de metilación del ADN y sus imágenes histopatológicas correspondientes. Para llevar a cabo la clasificación en base a los datos de metilación del DNA, se emplearán algoritmos de reducción de la dimensionalidad, así como técnicas de clasificación supervisadas y no supervisadas. Por otro lado, las imágenes histopatológicas serán analizadas con redes neuronales convolucionales y se combinarán con los datos de metilación para el desarrollo de algoritmos de clasificación de lesiones melanocíticas.

Palabras Clave: neoplasias melanocíticas spitzoides, metilación del ADN, imágenes histopatológicas, reducción de la dimensionalidad, *deep clustering*, redes neuronales convolucionales, *autoencoder*, *prototypical learning*.

Resum

El càncer de pell representa el grup més comú de neoplàsies malignes en la població blanca. De fet, en l'actualitat, un de cada tres càncers diagnosticats és càncer de pell. Encara que dos dels càncers cutanis més comunament diagnosticats són el carcinoma de cèl·lules basals i el de cèl·lules escatoses, sens dubte, el càncer de pell més agressiu i perillós és el melanoma. El melanoma cutani és una neoplàsia agressiva, amb un elevat risc metastàsico i una extrema complexitat genètica i biològica. En la majoria de tumors melanocítics, és possible una distinció patològica precisa entre benigne (nevus melanocític) i maligne (melanoma). No obstant això, hi ha neoplàsies melanocítiques de difícil diagnòstic clínic pel fet que posseïxen característiques morfològiques ambigües. Un exemple d'això són els cridats tumors spitzoides. Les neoplàsies melanocítiques spitzoides són lesions cutànies poc freqüents que constitueixen problemes de diagnòstic per als dermatopatòlegs. Estos tumors destaquen per la proliferació de grans melanòcits epitelioides o fusocelulars amb grans nuclis de cromatina vesicular i nuclèol prominent. La discrepància entre el seu aspecte morfològic histopatològic i la seua evolució suposa un desafiament en el diagnòstic. Esta discrepància ha servit d'estímul per a la realització d'estudis moleculars que permeten comprendre la biologia i el comportament de les lesions spitzoides. Entre estos estudis destaca la metilación del ADN, ja que és una marca distintiva en les lesions melanocítiques i pot ser clau per al seu desenrotllament i progressió metastàsica, convertint-se en un marcador diagnòstic, pronòstic i terapèutic.

En el present Treball Fi de Màster (TFM) es pretén abordar el desenrotllament d'algoritmes automàtics de classificació basats en *deep learning* amb els que poder identificar les neoplasias melanocítiques spitzoides benignes i malignes i realitzar el pronòstic d'aquelles amb potencial maligne incert utilitzant les dades de metilación de l'ADN i les seues imatges histopatològiques corresponents. Per a dur a terme la classificació basant-se en les dades de metilación del ADN s'empraran algoritmes de reducció de la dimensionalitat, així com tècniques de classificació supervisades i no supervisades. D'altra banda, les imatges histopatològiques seran analitzades amb xarxes neuronals convolucionals i es combinaran amb les dades de metilació per al desenrotllament d'algoritmes de classificació de lesions melanocítiques.

Paraules clau: neoplàsies melanocítiques spitzoides, metilación de l'ADN, imatges histopatològiques, reducció de la dimensionalidad, *deep clustering*, xarxes neuronals convolucionales, *autoencoder*, *prototypical learning*.

Abstract

Skin cancer represents the most common group of malignant neoplasms in the white population. In fact, one in three cancers diagnosed today is skin cancer. Although two of the most commonly diagnosed skin cancers are basal cell carcinoma and squamous cell carcinoma, without a doubt, the most aggressive and dangerous skin cancer is melanoma. Skin melanoma is an aggressive neoplasm with a high metastatic risk and extreme genetic and biological complexity. In most melanocytic tumors, a precise pathological distinction between benign (melanocytic nevus) and malignant (melanoma) is possible. However, there are melanocytic neoplasms that are difficult to clinically diagnose because of their ambiguous morphological characteristics. An example of this is the so-called "spitzoid tumours". Spitzoid melanocytic neoplasms are rare skin lesions that are diagnostic problems for dermatologists. These tumors are characterized by the proliferation of large epithelioid or fusocellular melanocytes with large nuclei of vesicular chromatin and prominent nucleolus. The discrepancy between their histopathological morphology and their evolution poses a diagnostic challenge. This discrepancy has stimulated molecular studies to understand the biology and behaviour of spitzoid lesions. Among these studies, DNA methylation stands out, since it is a distinctive mark in melanocytic lesions and it can be key to their development and metastatic progression, becoming a diagnostic, prognostic and therapeutic marker.

In the present Master's Degree thesis (TFM), I intend to address the development of automatic classification algorithms based on deep learning to identify benign and malignant spitzoid melanocytic neoplasms and to make the prognosis of those with uncertain malignant potential using DNA methylation data and its corresponding histopathological images. To carry out the classification based on the DNA methylation data, dimensionality reduction algorithms as well as supervised and unsupervised classification techniques will be used. Additionally, the histopathological images will be analyzed with convolutional neural networks and combined with the methylation data to develop classification algorithms for melanocytic lesions.

Keywords: melanocytic spitzoid neoplasms, DNA methylation, histopathological imaging, dimensionality reduction, deep clustering, convolutional neural networks, autoencoder, prototypical learning.

Listado de abreviaturas principales

| | |
|------|---|
| DNA | Deoxyribonucleic acid (Ácido desoxirribonucleico (ADN)) |
| CGI | CpG island (Islas CpG) |
| NGS | Next-Generation Sequencing (Secuenciación de próxima generación) |
| PCR | Polymerase chain reaction (Reacción en Cadena de la Polimerasa) |
| WSI | Whole-Slide-Image (Imágenes completas) |
| PMI | Potencial maligno incierto |
| RRBS | Reduced representation bisulfite sequencing (Secuencia de bisulfito de representación reducida) |
| NAS | Network-attached storage (Almacenamiento conectado a la red) |
| VAE | Variational autoencoder (Autoencoder variacionales) |
| MLP | Multilayer perceptron (Perceptrón multicapa) |
| AM | Attention module (Módulo de atención) |
| PN | Prototypical network (Redes prototípicas) |

Índice general

| | | |
|----------|--|-----------|
| I | Memoria | 1 |
| 1 | Introducción | 3 |
| 1.1 | Motivación y descripción del problema | 4 |
| 1.2 | Marco del proyecto | 5 |
| 1.3 | Objetivos del proyecto | 5 |
| 1.4 | Estructura del documento | 6 |
| 2 | Marco teórico | 7 |
| 2.1 | Neoplasias melanocíticas. Tumores spitzoides | 8 |
| 2.2 | Análisis de imágenes histopatológicas | 10 |
| 2.3 | Análisis de metilación del ADN. | 12 |
| 3 | Materiales | 19 |
| 3.1 | Base de datos de tumores melanocíticos spitzoides | 20 |
| 3.2 | Software utilizado | 22 |
| 3.3 | Hardware utilizado | 22 |
| 4 | Metodología | 25 |
| 4.1 | Preprocesado de los datos | 28 |
| 4.2 | Acondicionamiento de los datos | 32 |
| 4.3 | Partición de los datos | 38 |
| 4.4 | Algoritmos de clasificación para neoplasias melanocíticas spitzoides. | 40 |
| 4.5 | Algoritmos para el pronóstico de neoplasias melanocíticas spitzoides | 50 |
| 5 | Experimentación y Resultados | 55 |
| 5.1 | Experimentos realizados | 56 |
| 5.2 | Métricas utilizadas para la validación de los modelos | 61 |

| | |
|--|-----------|
| 5.3 Resultados y discusión | 63 |
| 6 Conclusiones y líneas futuras | 75 |
| 6.1 Conclusiones | 76 |
| 6.2 Líneas futuras | 78 |
| Bibliografía | 79 |
| | |
| II Presupuesto. | 87 |
| | |
| 1 Presupuesto | 89 |
| 1.1 Objetivo | 90 |
| 1.2 Presupuestos parciales | 90 |
| 1.3 Presupuesto Total | 92 |

Índice de figuras

| | |
|--|----|
| 2.1. Visualización de las células melanocíticas presentes en la capa basal de la epidermis de la piel y en los folículos pilosos. Figura extraída del Instituto Nacional del Cáncer [10]. | 8 |
| 2.2. Taxonomía de las neoplasias melanocíticas. (a) Neoplasias melanocíticas derivadas de melanocitos asociados al epitelio. Las diferentes clases tienen diferentes relaciones con la radiación UV y diferentes distribuciones asociadas a la edad, que se muestran en la parte inferior. (b) Neoplasias melanocíticas derivadas de melanocitos no asociados al epitelio. Este subgrupo no tiene relación con la radiación UV y, con la excepción de los melanomas congénitos asociados con nevus, que ocurren principalmente en niños pre-púberes, tienen una amplia distribución por edad. Abreviaturas: CSD: daño acumulado inducido por el sol. Figura extraída de [9]. | 9 |
| 2.3. Procesos llevados a cabo en la técnica histológica. Figura modificada de [14]. | 11 |
| 2.4. Imagen procedente de un tejido con una lesión melanocítica spitzoide obtenida tras la técnica histológica. Elaboración propia. | 11 |
| 2.5. Progresión de las características histopatológicas en el desarrollo del melanoma spitzoide. Figura extraída de [12]. | 12 |
| 2.6. Proceso de metilación del ADN. Metilación en la posición 5' de la citosina catalizada por DNMT (metiltransferasas), en presencia de S-adenosil-metionina (SAM). Elaboración propia. | 13 |
| 2.7. a) Regiones CpGs desmetiladas dentro de la región promotora. b) Regiones CpGs metiladas dentro de las zonas promotoras. Elaboración propia. | 14 |
| 2.8. Protocolo estándar para la adquisición de datos de metilación mediante la técnica RRBS. Figura modificada de [25]. | 17 |

| | |
|---|----|
| 2.9. Alteración del patrón de metilación del ADN en el cáncer. En las células sanas, las islas CpG de las zonas promotoras de los genes supresores de tumores (1) se encuentran no metiladas, de forma que el gen se expresa normalmente. Por otro lado, las secuencias repetitivas del ADN (2, 3) se encuentran metiladas y por tanto, silenciadas. En el cáncer, el patrón de metilación se invierte, promoviendo la silenciaci3n de genes supresores de tumores, y la reexpresi3n de secuencias repetitivas y protooncogenes. Figura modificada de [27]. | 18 |
| 3.1. Proceso llevado a cabo para obtener el valor de metilaci3n del tejido bajo estudio. Elaboraci3n propia. | 21 |
| 3.2. Anotaci3n realizadas por los pat3logos en muestras con neoplasias benignas, malignas y de potencial maligno incierto. Elaboraci3n propia. | 21 |
| 4.1. Diagrama de flujo de los principales procesos que componen la metodologí a aplicada en este proyecto. | 26 |
| 4.2. Mapa de correlaci3n relativo a los 491 sitios CpG de los que se ha adquirido el valor de metilaci3n. | 28 |
| 4.3. Histograma con ajuste de distribuci3n y grá fico cuantil-cuartil para tres variables representativas de la base de datos. | 30 |
| 4.4. Análisis discriminatorio de tres variables representativas de la base de datos con respecto a la clase. | 30 |
| 4.5. Patches generados de las imagenes WSI. (a) Patch de una zona no anotada por el pat3logo. (b). Patch de una regi3n representativa de las lesiones melanocít icas spitzoides. (c) Patch que no contiene suficiente regi3n de interés. | 31 |
| 4.6. Arquitectura algoritmo <i>autoencoder</i> utilizado para la reducci3n de la dimensionalidad de forma no supervisada. Elaboraci3n propia. | 34 |
| 4.7. Principales diferencias entre un algoritmo basado en <i>autoencoder</i> y aquel basado en un <i>autoencoder</i> variacional. | 35 |
| 4.8. Arquitectura comú n del autoencoder variacional. En este caso, este algoritmo est á optimizado por la minimizaci3n de dos funciones de p3rdidas, una que corresponde a la regularizaci3n del espacio latente y otra correspondiente a la reconstrucci3n de los datos de entrada. Elaboraci3n propia. | 36 |
| 4.9. Utilizaci3n de la metodologí a SMOTE para la generaci3n de muestras sintéticas. a) Se eligen K vecinos en torno al ejemplo dado. b) Se calcula la diferencia entre el vector de característ icas del ejemplo bajo estudio y sus vecinos. c) Se generan nuevos ejemplos en el espacio de característ icas delimitados por los segmentos que unen el ejemplo de estudio y cada uno de sus vecinos. | 38 |
| 4.10. Metodologí a <i>K-fold cross validation</i> implementada en este estudio para llevar a cabo una correcta partici3n de los datos. | 39 |
| 4.11. Arquitectura tí pica de las neuronas que componen un perceptr3n multicapa. | 40 |

| | |
|---|----|
| 4.12. Arquitectura típica de una red neuronal compuesta por capas de entrada, ocultas y de salida para el objetivo de clasificación. Elaboración propia. | 41 |
| 4.13. Funciones de activación más conocidas utilizadas en los clasificadores compuestos por redes neuronales. Elaboración propia. | 41 |
| 4.14. Flujo de trabajo para la clasificación de datos de metilación, donde K es el número de <i>folds</i> , en este caso 5. Elaboración propia. | 42 |
| 4.15. Arquitectura de deep learning propuesta para el problema de clasificación de lesiones melanocíticas. Elaboración propia. | 44 |
| 4.16. Arquitectura básica de una red neuronal convolucional. | 46 |
| 4.17. Arquitectura VGG diseñada para llevar a cabo tareas de clasificación. Figura extraída de [67]. | 47 |
| 4.18. Módulo de atención (<i>attention module</i>) unido a la salida del <i>base model</i> | 48 |
| 4.19. Arquitectura de los bloques Squeeze-and-Excitation. | 49 |
| 4.20. Técnica de <i>majority voting</i> para obtener la predicción a nivel de WSI. | 50 |
| 4.21. Arquitectura de una red híbrida para la fusión de datos. | 51 |
| 4.22. Prototipos c_k computados como la media de los ejemplos embebidos en el espacio latente para cada clase. Los puntos de consulta embebidos se clasifican a través de una <i>softmax</i> de distancias a los prototipos de clase. Elaboración propia. | 53 |
| | |
| 5.1. Arquitectura convolucional propuesta para la clasificación de lesiones melanocíticas spitzoides. Cabe destacar que en algunos casos el tamaño de los bloques no se corresponde con su dimensión real. | 59 |
| 5.2. Arquitectura propuesta para la clasificación de las lesiones melanocíticas a nivel de paciente mediante la fusión de las características de las imágenes histopatológicas y los datos de metilación. | 60 |
| 5.3. Visualización de los cuatro tipos de soluciones en función de la estimación del modelo y la etiqueta de cada muestra. | 61 |
| 5.4. Pérdidas de los modelos de reducción de la dimensionalidad propuestos. (a) Pérdidas de reconstrucción obtenidas con el algoritmo de autoencoder. (b) Pérdidas de reconstrucción y de regularización (KL) obtenidas por el autoencoder variacional. | 63 |
| 5.5. Comparación de las pérdidas obtenidas por el modelo basado en <i>autoencoder</i> (AE) y aquel basado en <i>variational autoencoder</i> (VAE). | 64 |
| 5.6. Representación de las dos primeras variables en distintos espacios de datos. (a) Visualización de las dos primeras variables de los datos de entrada iniciales. (b) Visualización de las dos primeras variables en el espacio latente del algoritmo basado en autoencoder. (c) Visualización de las dos primeras variables en el espacio latente del autoencoder variacional. (d) Visualización de las dos primeras características del espacio latente tras el entrenamiento del algoritmo basado en <i>deep clustering</i> | 65 |

| | |
|--|----|
| 5.7. Curvas ROC correspondientes a los resultados alcanzados por los diferentes métodos propuestos. | 67 |
| 5.8. Representación de las características del espacio latente de los pacientes con potencial maligno incierto. Leyenda: el color azul representa aquellos pacientes que han sido predichos como benignos mientras que el color naranja representa aquellos pacientes que han sido predichos como malignos. Mediante los dos círculos de color cyan y rojo, se representan los centroides de la clase benigna y maligna respectivamente. El acrónimo C hace referencia al número de componente del centroide y de las características del espacio latente que se está representando. | 68 |
| 5.9. Curva ROC de los distintos métodos empleados para la clasificación de lesiones melanocíticas. | 70 |
| 5.10. Mapas de activación de clase para muestras benignas y malignas aleatorias obtenidas por los modelos BM y BM+AM. Cada columna representa los <i>patches</i> de un paciente elegido al azar mientras que cada fila representa los resultados obtenidos por cada uno de los modelos implementados. | 70 |
| 5.11. Visualización del pronóstico basado en modelos de clasificación de imágenes utilizando las primeras componentes del método <i>t-SNE</i> . Leyenda: el color azul representa aquellos pacientes que han sido predichos como benignos mientras que el color naranja representa aquellos pacientes que han sido predichos como malignos. Mediante los dos círculos de color cyan y rojo, se representan los centroides de la clase benigna y maligna respectivamente. | 71 |
| 5.12. Curva ROC de los distintos métodos empleados para la clasificación de lesiones melanocíticas. | 73 |

Índice de tablas

| | |
|---|----|
| 2.1. Esquema de las diversas técnicas que se pueden utilizar para cuantificar la metilación de acuerdo con el pretratamiento que recibe el ADN y el número de nucleótidos modificados y la resolución de la detección (metilación global, de región específica o genómica). Tabla extraída de [22]. | 15 |
| 4.1. Número de <i>patches</i> obtenidos para cada uno de los pacientes con lesiones melanocíticas malignas y benignas. | 32 |
| 4.2. Partición de datos generada mediante la técnica <i>K-fold</i> ajustada. | 39 |
| 5.1. Modelo del <i>autoencoder</i> propuesto para la reducción de la dimensionalidad de los datos de metilación. En esta tabla se expone el número de capas que componen la parte del <i>encoder</i> y el <i>decoder</i> así como el número de neuronas de cada capa. | 56 |
| 5.2. Modelo del <i>variational autoencoder</i> propuesto para para la reducción de los datos de metilación. En esta tabla se expone el número de capas que componen la parte del <i>encoder</i> y <i>decoder</i> así como el número de neuronas de cada capa. | 57 |
| 5.3. Arquitectura del modelo de <i>deep clustering</i> propuesto. Como se puede observar, este algoritmo está compuesto por un <i>autoencoder</i> (parte de <i>encoder</i> y <i>decoder</i>) al que se le ha unido en la salida del <i>encoder</i> una capa de <i>clustering</i> para llevar a cabo el proceso de clasificación no supervisado. | 58 |
| 5.4. Resultados obtenidos para los algoritmos de clasificación propuestos. Cabe destacar que para el caso de los algoritmos supervisados esta tabla muestran los resultados obtenidos en validación. Sin embargo, en el caso de los algoritmos no supervisados, esta tabla muestra los resultados obtenidos tras el entrenamiento de los modelos. | 66 |
| 5.5. Ratio de error y número de ejemplos en los que los diferentes clasificadores dan una respuesta errónea. | 66 |
| 5.6. Pronóstico obtenido para los pacientes con potencial maligno incierto (PMI). | 67 |
| 5.7. Resultados de clasificación sobre las imágenes histopatológicas obtenidos con los algoritmos propuestos. BM:base model; AM:módulo de atención; TM: top model. | 69 |

| | |
|--|----|
| 5.8. Ratio de error y número de ejemplos en los que los diferentes clasificadores dan una respuesta errónea. | 69 |
| 5.9. Pronóstico obtenido para los pacientes con potencial maligno incierto (PMI). | 71 |
| 5.10. Comparación de resultados obtenidos mediante los datos de metilación (DM) y los datos procedentes de imágenes histopatológicas (IH). | 72 |
| 5.11. Resultados obtenidos para los algoritmos de clasificación fusionando los datos (FD). . . | 72 |
| 5.12. Ratio de error y número de ejemplos en los que el clasificador basado en la fusión de datos (FD) da una respuesta errónea. | 73 |
| | |
| 1.1. Descomposición de los costes de personal. | 90 |
| 1.2. Coste del uso del equipo personal. | 91 |
| 1.3. Coste del uso de los servidor del grupo CVBLAB. | 91 |
| 1.4. Descomposición de los costes de hardware. | 91 |
| 1.5. Descomposición de los costes de software. Todos los precios corresponden a las licencias de uso Académico. | 92 |
| 1.6. Presupuesto de Ejecución de Material del proyecto. | 92 |
| 1.7. Presupuesto total del presente proyecto. | 93 |

Parte I

Memoria

Capítulo 1

Introducción

Índice de contenidos

| | |
|---|---|
| 1.1 Motivación y descripción del problema | 4 |
| 1.2 Marco del proyecto | 5 |
| 1.3 Objetivos del proyecto | 5 |
| 1.4 Estructura del documento | 6 |

1.1 Motivación y descripción del problema

El cáncer de piel representa el grupo más común de neoplasias malignas entre la población blanca [1]. La tasa de incidencia del cáncer de piel, melanoma y no melanoma, está aumentando en todo el mundo. De hecho, en la actualidad, uno de cada tres cánceres diagnosticados es cáncer de piel [2].

Aunque dos de los cánceres de piel más comúnmente diagnosticados son el carcinoma de células basales y el de células escamosas, lesiones no melanocíticas desarrolladas a partir de las células no pigmentadas de la piel, sin duda, el cáncer de piel más agresivo y peligroso es el melanoma [3]. El melanoma cutáneo es una neoplasia agresiva, con un elevado riesgo metastásico y una extrema complejidad genética y biológica, que posee numerosos mecanismos de resistencia frente a agentes terapéuticos [4]. Se trata del cáncer con el mayor aumento de incidencia en el mundo durante los últimos años y, representa el cáncer de piel con el mayor índice de mortalidad, sustentado por su alto potencial metastásico. Por tanto, la comprensión de los mecanismos que permiten su diseminación es un paso esencial en la predicción y prevención terapéutica del proceso metastásico del melanoma [5]. Debido a la alta incidencia y a las tasas de mortalidad de esta patología, el diagnóstico temprano del melanoma se ha convertido en un aspecto crucial, pues la identificación temprana de este puede reducir en gran medida su mortalidad [6].

Las neoplasias melanocíticas incluyen varios tipos de tumores caracterizados por distintas características clínicas, diversos aspectos histopatológicos y diferente comportamiento clínico. En la mayoría de tumores melanocíticos, es posible una distinción patológica precisa entre benigno (nevus melanocítico) y maligno (melanoma). Sin embargo, existen neoplasias melanocíticas de difícil diagnóstico clínico debido a que poseen características morfológicas ambiguas. Esta dificultad conduce a un diagnóstico insuficiente como nevus, a un diagnóstico excesivo como melanoma o a un diagnóstico de “tumor melanocítico de potencial maligno incierto” o “tumor melanocítico límite”. Uno de los principales integrantes de este grupo heterogéneo de tumores melanocíticos son los llamados “tumores spitzoides”.

Las neoplasias melanocíticas spitzoides son lesiones cutáneas poco frecuentes que constituyen problemas de diagnóstico para los dermatopatólogos de forma regular. Estos tumores destacan por la proliferación de grandes melanocitos epitelioides o fusocelulares con grandes núcleos de cromatina vesicular y nucléolo prominente [7]. La discrepancia entre el aspecto morfológico histopatológico y la evolución clínica en este tipo de melanoma supone un desafío en el diagnóstico. Esta discrepancia ha servido de estímulo para la realización de estudios moleculares que permitan comprender la biología y el comportamiento de las lesiones spitzoides.

Por todo lo descrito anteriormente, este proyecto pretende desarrollar algoritmos que combinen la información proporcionada por las imágenes histopatológicas y aquella obtenida mediante estudios moleculares para la clasificación automática de las lesiones melanocíticas spitzoides en nevus y melanoma y, la predicción de los tumores melanocíticos de potencial maligno incierto. Por consiguiente, con la creación de estos algoritmos de clasificación se pretende proporcionar un sistema de ayuda al diagnóstico que facilite la tarea de los patólogos a la hora de diagnosticar patologías melanocíticas.

1.2 Marco del proyecto

Este trabajo se enmarca dentro del proyecto de investigación “*Cloud Artificial Intelligence For Pathology*” (CLARIFY). CLARIFY es un proyecto europeo coordinado por el grupo de investigación CV-BLab (Computer Vision and Behaviour Analysis Lab), perteneciente a la Universidad Politécnica de Valencia. Este proyecto se encuentra subvencionado por la Comisión Europea a través de Horizonte 2020, programa de Financiación de la Investigación y la Innovación en la Unión Europea, para el periodo 2019-2023.

El propósito general de CLARIFY se centra en el desarrollo de un entorno robusto de diagnóstico digital automatizado basado en inteligencia artificial. Para ello, se propone la implementación de algoritmos de datos en la nube que faciliten la interpretación de las imágenes whole-slide (WSI) y que permitan integrar la información aportada por los estudios moleculares. De este modo, se maximizarán los beneficios que aporta la patología digital y los estudios moleculares y, se ayudará a los patólogos en su trabajo diario.

Aunque en la mayoría de casos los patólogos pueden juzgar observando una muestra de tejido bajo un microscopio o una biopsia digitalizada en un ordenador, el análisis manual de las imágenes histopatológicas es típicamente subjetivo, sufre de variabilidad intra e interobservador y consume mucho tiempo. Además, la integración de estudios moleculares en el diagnóstico del cáncer ha demostrado ser un factor determinante en muchos de los casos. Para abordar estos problemas, se desea desarrollar sistemas computarizados que puedan ayudar a los patólogos y proporcionar evaluaciones cuantitativas. Con el objetivo de validar las herramientas y métodos desarrollados, en el proyecto CLARIFY se han seleccionado tres tipos de cáncer específicos y desafiantes que reflejan la variabilidad existente en el diagnóstico del cáncer: cáncer de mama triple negativo, cáncer de vejiga invasivo no muscular de alto riesgo y lesiones melanocíticas Spitzoides.

Dado que dentro del marco de este proyecto, la labor de la Universidad Politécnica de Valencia (UPV), más concretamente del grupo CVBLab, es el diseño y desarrollo de algoritmos automáticos que sean aplicados sobre las imágenes WSI obtenidas y sobre los datos procedentes de estudios moleculares, este trabajo está centrado en la confección de métodos destinados a la clasificación de uno de los tumores estudiados bajo el proyecto CLARIFY, los tumores melanocíticos spitzoides.

1.3 Objetivos del proyecto

El objetivo principal de este trabajo fin de máster (TFM) es el desarrollo de algoritmos automáticos basados en inteligencia artificial para la distinción entre tumores melanocíticos spitzoides benignos (nevus) y malignos. Para lograr la consecución del objetivo principal de este TFM se debe abordar una serie de objetivos específicos:

1. Generar la bases de datos a utilizar en este proyecto con su correspondiente *ground truth*¹. Para ello, por un lado, se han de obtener las imágenes histopatológicas que son objeto de estudio y adecuarlas utilizando la tecnología WSI, que permite analizar imágenes digitales de alta resolución de muestras completas de biopsias. Por otro lado, es necesario adquirir los datos correspondientes al estudio molecular realizado, en este caso los datos obtenidos mediante el estudio de metilación del ADN.

¹*Ground truth: Modelos construidos manualmente utilizados como referencia con los que se comparan los resultados obtenidos por un método automático.*

2. Realizar un estado del arte de los métodos que anterior y actualmente se están utilizando para suplir los objetivos de clasificación planteados en este proyecto.
3. Realizar un preprocesado sobre los datos de metilación de ADN basado en un análisis estadístico, que permita la selección de las mejores características en términos de independencia entre clases y entre variables y, en algoritmos de reducción de la dimensionalidad.
4. Desarrollar modelos de clasificación (supervisados y no supervisados) de tumores melanocíticos spitzoides que utilicen los datos de metilación del ADN basados en aprendizaje profundo.
5. Implementar algoritmos de clasificación basados en redes neuronales convolucionales para el análisis de imágenes histopatológicas.
6. Desarrollo de modelos híbridos que utilicen tanto la información proporcionada por los datos de metilación del ADN como por las imágenes histopatológicas.
7. Predicción de la evolución de los tumores melanocíticos de potencial maligno incierto con los algoritmos de clasificación desarrollados.
8. Efectuar una comparación entre los diferentes métodos de clasificación propuestos en este trabajo y los propuestos por otros autores del estado del arte actual en la medida que sea posible.
9. Identificar y describir los problemas y limitaciones encontrados durante el estudio y, proponer posibles mejoras para investigaciones futuras.

1.4 Estructura del documento

En el capítulo 2 se realiza una extensa descripción de los tumores melanocíticos y en concreto, de los tumores melanocíticos spitzoides puesto que son el objeto de estudio del presente trabajo. Además, se exponen las características fundamentales de las dos técnicas utilizadas en este proyecto, imágenes histopatológicas y estudio de metilación del ADN, así como su relación con el diagnóstico de los tumores melanocíticos spitzoides.

En el capítulo 3 se detalla tanto el material como la metodología empleada para llevar a cabo el desarrollo de los algoritmos de clasificación de tumores melanocíticos spitzoides.

En el capítulo 4 se muestran, se discuten y se comparan respecto al arte los resultados obtenidos con las técnicas propuestas para la clasificación de tumores melanocíticos spitzoides. Además, se aportan los resultados obtenidos tras el análisis de nuevos pacientes totalmente desconocidos por el sistema, tal y como se llevaría a cabo para nuevas muestras en la práctica clínica.

En el capítulo 5 se resumen las conclusiones más importantes extraídas de los estudios realizados y se detallan algunas propuestas, así como posibles líneas de investigación dentro del marco del proyecto.

Capítulo 2

Marco teórico

Índice de contenidos

| | |
|---|----|
| 2.1 Neoplasias melanocíticas. Tumores spitzoides | 8 |
| 2.1.1 Características de las neoplasias melanocíticas. | 8 |
| 2.1.2 Tumores melanocíticos spitzoides | 9 |
| 2.2 Análisis de imágenes histopatológicas | 10 |
| 2.2.1 Técnica histológica | 10 |
| 2.2.2 Rasgos histológicos relacionados con tumores melanocíticos spitzoides | 11 |
| 2.3 Análisis de metilación del ADN. | 12 |
| 2.3.1 Bases moleculares de la metilación del ADN | 13 |
| 2.3.2 Técnicas de adquisición de la metilación del ADN. | 14 |
| 2.3.3 Metilación del ADN y melanoma. | 17 |

2.1 Neoplasias melanocíticas. Tumores spitzoides

2.1.1 Características de las neoplasias melanocíticas

Los melanocitos, células dendríticas encargadas de la producción de melanina (Figura 2.1), pueden dar lugar a un conjunto diverso de neoplasias con distribución anatómica, características clínicas, apariencia histopatológica y comportamiento biológico variable. Aunque los melanocitos son más abundantes en la piel, donde juegan un papel crítico en la pigmentación y en la protección solar, también están presentes en otros lugares del cuerpo como en el epitelio de las mucosas, en la conjuntiva, en la retina y en las leptomeninges, donde tienen funciones adicionales menos conocidas [8]. Las neoplasias melanocíticas surgen con mayor frecuencia a partir de los melanocitos de la piel, sin embargo, también pueden proceder de melanocitos autóctonos de numerosos órganos internos, incluido el sistema nervioso central (SNC) [9].

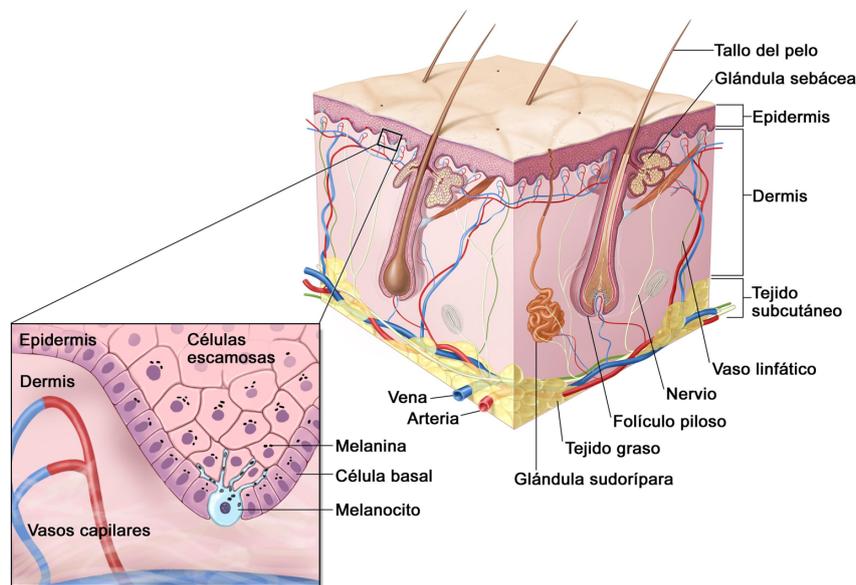


Figura 2.1: Visualización de las células melanocíticas presentes en la capa basal de la epidermis de la piel y en los folículos pilosos. Figura extraída del Instituto Nacional del Cáncer [10].

Como regla general, las neoplasias benignas de linaje melanocítico se denominan nevos melanocíticos (el plural de nevo) y los malignos se denominan melanomas. Aunque los melanomas pueden surgir de los nevos, como se puede inferir de la presencia de un remanente de nevo adyacente a un melanoma, la mayoría de los melanomas primarios no muestran un nevo precursor asociado. En algunos casos, esto se debe a que el nevo precursor fue cubierto por el melanoma durante su progresión, pero en muchos casos probablemente no hubo un estado precursor benigno detectable.

Un sistema de clasificación deseable de neoplasias melanocíticas separa los distintos grupos al considerar las diferencias en las células de origen, la patogénesis, los aspectos clínicos e histológicos, las alteraciones genéticas, etc. La clasificación en base a estos aspectos puede servir como marco para desarrollar enfoques de prevención primaria refinados, algoritmos de diagnóstico y estadificación, y tratamientos terapéuticos personalizados. En la figura 2.2 se expone la clasificación llevada a cabo por [9], en la que se puede observar los distintos grupos de tumores melanocíticos existentes.

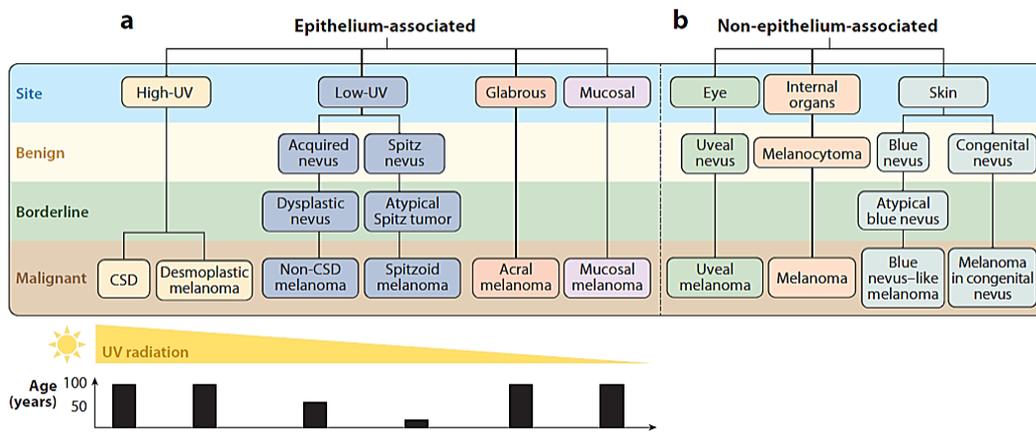


Figura 2.2: Taxonomía de las neoplasias melanocíticas. (a) Neoplasias melanocíticas derivadas de melanocitos asociados al epitelio. Las diferentes clases tienen diferentes relaciones con la radiación UV y diferentes distribuciones asociadas a la edad, que se muestran en la parte inferior. (b) Neoplasias melanocíticas derivadas de melanocitos no asociados al epitelio. Este subgrupo no tiene relación con la radiación UV y, con la excepción de los melanomas congénitos asociados con nevos, que ocurren principalmente en niños prepúberes, tienen una amplia distribución por edad. Abreviaturas: CSD: daño acumulado inducido por el sol. Figura extraída de [9].

La mayoría de tumores melanocíticos tienen unas características microscópicas y genéticas que permiten, generalmente, su clasificación en benignas y malignas. Sin embargo, existen neoplasias melanocíticas con difícil diagnóstico ya que poseen características morfológicas conflictivas. Esta dificultad lleva a un diagnóstico insuficiente como nevus, a un sobrediagnóstico como melanoma o a un diagnóstico de “tumor de potencial maligno incierto” o “límite tumor melanocítico” [11]. Uno de los principales componentes de este heterogéneo grupo son los tumores melanocíticos conocidos como “spitzoides”. Dado que el presente TFM está centrado en el estudio de estas neoplasias melanocíticas, a continuación, se ofrece una descripción sobre este tipo de neoplasias.

2.1.2 Tumores melanocíticos spitzoides

Las neoplasias melanocíticas spitzoides (conocidas como tumores de spitz) son lesiones cutáneas poco frecuentes que constituyen un problema de diagnóstico para los dermatólogos de forma regular [12]. Este tipo de neoplasias se describieron inicialmente en 1948 por Sophie Spitz como “melanomas de la infancia” ya que ocurrían predominantemente en niños y adolescentes. Los melanomas derivados de estas neoplasias se caracterizaban por la proliferación de grandes melanocitos epitelioides o fusocelulares con núcleos grandes de cromatina vesicular y nucléolo prominente [11]. Cuando quedó claro que estos tumores también podían surgir en etapas posteriores de la vida y comportarse de manera benigna, se les cambió el nombre a “Spitz naevi” para indicar su naturaleza benigna. Los tumores melanocíticos con características de spitz e histología atípica que muestran a menudo un curso clínico agresivo similar a los melanomas convencionales se denominaron “melanomas spitzoides”.

El término “tumores de Spitz atípicos”(AST) se refiere a tumores melanocíticos que exhiben características morfológicas de Spitz naevi, así como algunas características asociadas con malignidad pero en un grado insuficiente para clasificarlos como melanomas spitzoides [11]. Para este grupo, se ha aplicado el término de “tumor spitzoide de potencial maligno incierto” (STUMP) que actualmente incluye: 1) nevos melanocíticos que tienen alguna característica común con melanoma, pero que biológicamente son

benignos; 2) tumores atípicos con riesgo desconocido de comportamiento agresivo. Un tercio de ellos se diseminan a los ganglios linfáticos, pero la mayoría de los pacientes no desarrollan metástasis a distancia ni mueren a causa de la enfermedad [13]; 3) un subgrupo de melanoma spitzoide (maligno), con un comportamiento agresivo similar al melanoma convencional. Sin embargo, dado que faltan criterios diagnósticos y pronósticos específicos y útiles, existe una tendencia a diagnosticarlos en exceso como malignos como medida de precaución.

La clasificación patológica de los tumores de Spitz como "Spitz naevus", "AST" o "melanoma spitzoide" y la predicción de su comportamiento clínico puede ser un desafío, incluso para los expertos.

2.2 Análisis de imágenes histopatológicas

El examen histopatológico es el estándar de oro para el diagnóstico de neoplasias melanocíticas spitzoides [11]. A continuación, se expone el proceso de preparación de las muestras histológicas para su posterior análisis, así como la relación existente entre las características morfológicas observadas en las muestras histopatológicas y la progresión de los tumores melanocíticos spitzoides.

2.2.1 Técnica histológica

Para poder observar, examinar y analizar los componentes morfológicos de las diferentes muestras de tumores biopsiados a través de microscopios fotónicos, electrónicos o mediante sistemas digitales, es necesario someter a las muestras a un proceso conocido como técnica histológica. La técnica histológica está compuesta por una serie de procedimientos, mostrados en la Figura 2.3, que tienen como finalidad preparar la muestra y conferirle las condiciones óptimas para su análisis.

- **Fijación de la muestra.** Una vez que se ha adquirido la muestra biopsiada a analizar, se procede a su fijación. La finalidad de este proceso es detener la vida de las células del tejido extraído e impedir las modificaciones autolíticas que pueda sufrir la muestra para mantener la estructura morfológica de las células.
- **Inclusión.** Los tejidos fijados adquieren cierta dureza y consistencia, sin embargo, esta no es suficiente para obtener secciones delgadas de tejido que puedan ser visualizadas. Para ello, se procede a infiltrar los tejidos en diferentes sustancias "de inclusión". De esta forma, estas sustancias se infiltran en el interior de las muestras con la finalidad de darles soporte y conformar un bloque homogéneo en dureza y consistencia. Entre las sustancias más comúnmente utilizadas para llevar a cabo la inclusión se encuentra la parafina.
- **Microtoma.** Tras proporcionarle a los tejidos la dureza necesaria, se lleva a cabo la obtención de los diferentes cortes. Estos cortes o secciones delgadas se obtienen utilizando instrumentos mecánicos, denominados microtomos, diseñados para que, de forma parcialmente automática, seccionen el bloque de parafina en cortes delgados y de grosor uniforme.
- **Tinción.** Una vez que se extraen los cortes de las muestras y son adheridos a los portaobjetos¹ se procede a la coloración o tinción de la muestra. El pigmento más comúnmente utilizado se denomina hematoxilina y eosina (HE). La hematoxilina se encarga de teñir los núcleos de las células proporcionándoles un color violeta característico, mientras que la eosina proporciona al citoplasma y al material extracelular diferentes grados de color rosado.

¹Lámina que sirve de soporte para las preparaciones o los cuerpos que se observan en un microscopio.

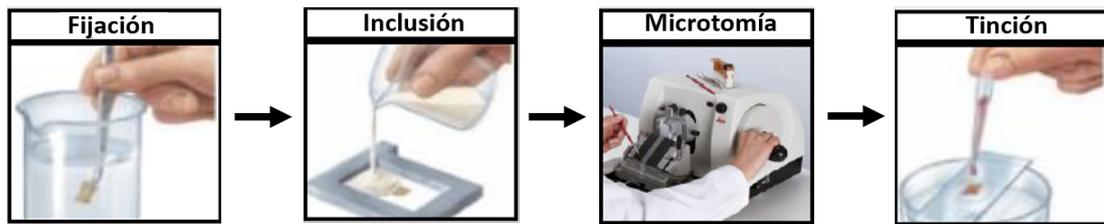


Figura 2.3: Procesos llevados a cabo en la técnica histológica. Figura modificada de [14].

Tras la técnica histológica, se obtienen imágenes como la visualizada en la Figura 2.4 procedente de un tejido que presenta una lesión melanocítica spitzoide.

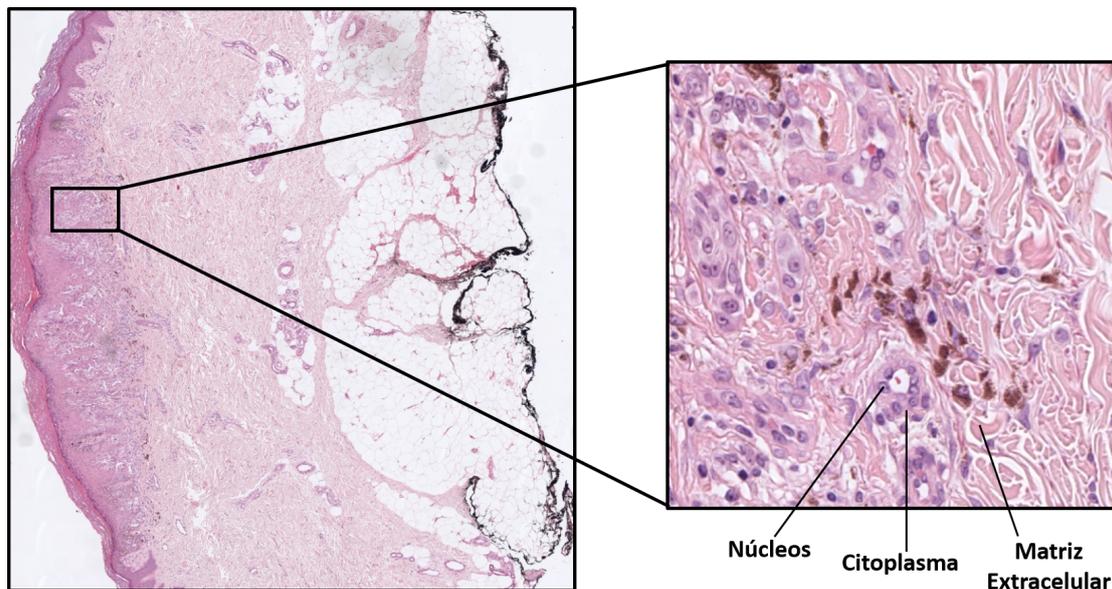


Figura 2.4: Imagen procedente de un tejido con una lesión melanocítica spitzoide obtenida tras la técnica histológica. Elaboración propia.

2.2.2 Rasgos histológicos relacionados con tumores melanocíticos spitzoides

Para identificar la benignidad o malignidad de las muestras de tumores melanocíticos spitzoides a partir del análisis de las imágenes histopatológicas, existen diversas características morfológicas a tener en cuenta.

Un rasgo importante en este tipo de neoplasias es el diámetro de la lesión, el cual se registra en milímetros. La mayoría de tumores spitzoides benignos miden menos de 10 mm y con frecuencia menos de 5–6 mm. Un tamaño más allá de 10 mm generalmente es característico de melanomas spitzoides [12]. Como en todas las lesiones melanocíticas, el aumento de la asimetría en el tejido es anormal y está relacionado con malignidad. Por otra parte, los tumores spitzoides benignos tienden a estar fuertemente circunscritos en su periferia mientras que las lesiones atípicas están a menudo menos circunscritos. La diseminación pagetoide² se observa con poca frecuencia en los tumores spitzoides nevus. Sin embargo, para que dicha

²Pagetoid es un término utilizado en dermatología para referirse a la "propagación ascendente" de células anormales en la epidermis (es decir, de abajo hacia arriba).

diseminación sea benigna debe limitarse a la mitad inferior de la epidermis, debe ser focal y con escaso número de células. La extensión pagetoide más extensa que involucra a la mitad superior de la epidermis es claramente signo de malignidad.

La prominente confluencia de melanocitos así como la alta densidad celular son parámetros estrechamente correlados y se encuentran entre los criterios más importantes para evaluar las lesiones melanocíticas [12]. Desafortunadamente, estas características son subjetivas y por lo tanto, difíciles de reconocer de manera confiable y reproducible. Estos dos parámetros también suelen estar estrechamente vinculados a una maduración disminuida o ausente. Por tanto, agregados o nódulos celulares confluentes de tamaño considerable y con apariencia abarrotada en el componente dérmico, particularmente reemplazando la dermis, y extendiéndose profundamente sin maduración son también signos de malignidad [15]. Por último, la no uniformidad de una lesión así como la presencia de una tasa de proliferación creciente se correlacionan con un comportamiento agresivo o de malignidad. Altas tasas mitóticas, particularmente de más de seis mitosis por mm^2 aumentan considerablemente la malignidad.

Aunque la evaluación de estas características histológicas permite distinguir el nevus de Spitz del melanoma en la mayoría de los casos, hay otros que muestran una superposición morfológica considerable. En estos casos, la asignación definitiva de un diagnóstico benigno o maligno no es posible, lo que genera incertidumbre sobre el tratamiento. La concordancia interobservador en estos casos es deficiente, tanto con respecto al diagnóstico como a la predicción del pronóstico. Como se ha mencionado anteriormente, estos casos de incertidumbre diagnóstica se denominan tumores Spitz atípicos o tumores spitzoides de potencial maligno incierto [15].

En la Figura 2.5 se observa el progreso histológico de los diferentes subtipos de tumores melanocíticos spitzoides, desde el grado de no malignidad hasta el desarrollo de melanomas spitzoides, en el que se pueden identificar las características anteriormente expuestas.

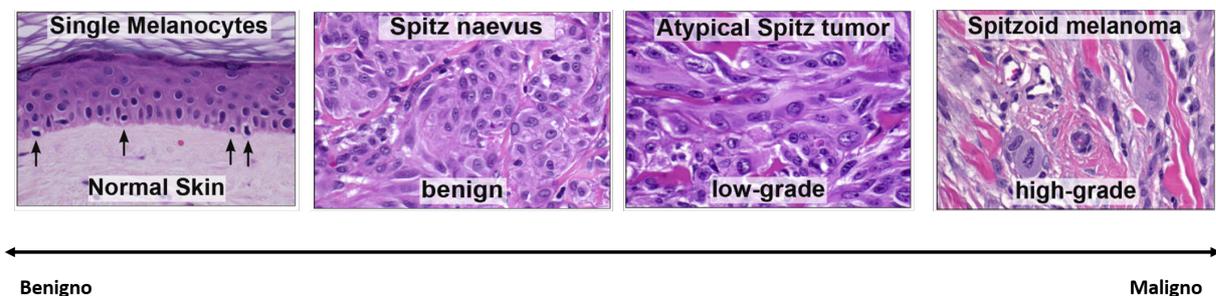


Figura 2.5: Progresión de las características histopatológicas en el desarrollo del melanoma spitzoide. Figura extraída de [12].

2.3 Análisis de metilación del ADN

En la actualidad, el diagnóstico del melanoma spitzoide requiere nuevos métodos que complementen al estudio clínico-histopatológico convencional, con el fin de entender su complejidad biológica, mejorar la precisión del diagnóstico y poder abordar un tratamiento adecuado. Entre estas técnicas se incluyen: hibridación in situ (ISH) [16], arrays de hibridación genómica comparativa (CHG) [9], next-generation sequencing [7], metilación del DNA [17] y expresión de miRNAs [18]. En concreto, la metilación del DNA es una marca distintiva en las lesiones melanocíticas y puede ser clave para el desarrollo de dichas lesiones y de su progresión metastásica [19]. Es por ello que, en el presente proyecto, se procede a

la incorporación de los datos procedentes de estudios de la metilación del ADN en el diagnóstico del melanoma spitzoide.

2.3.1 Bases moleculares de la metilación del ADN

La metilación del ADN es un cambio epigenético³ útil para la distinción de varios tipos de tumores. A nivel bioquímico, la metilación del ADN consiste en la adición covalente de un grupo metilo ($-CH_3$) exclusivamente en la posición 5 del nucleótido de citosina (C), Figura 2.6. La reacción de metilación de las citosinas está mediada por una clase de enzimas llamadas metiltransferasas de ADN (DNMT) que catalizan la transferencia del grupo metilo de S-adenosil-metionina a la citosina. Aunque se han identificado cinco miembros de la familia DNMT en mamíferos: DNMT1, DNMT2, DNMT3a, DNMT3b y DNMT3L, solo DNMT1, DNMT3a y DNMT3b interactúan para producir el patrón global de metilación de citosina [20].

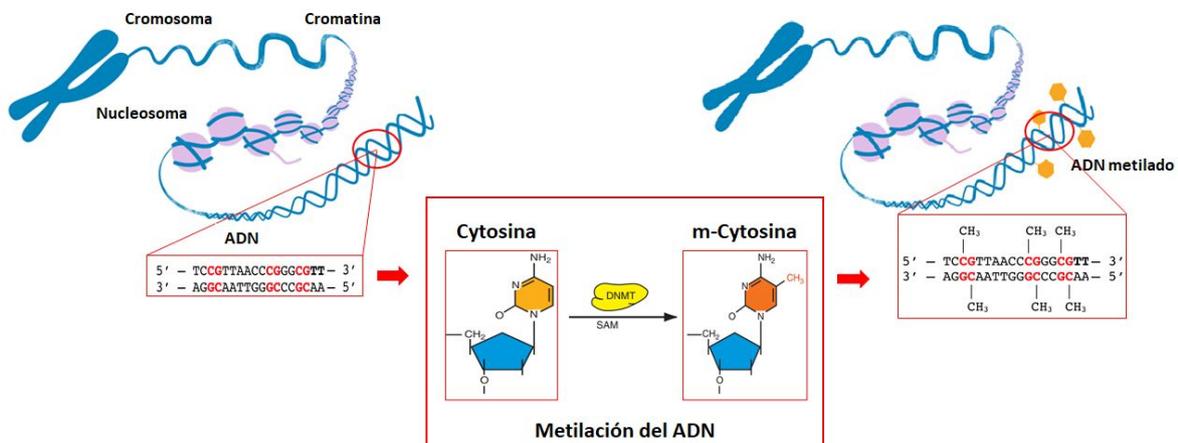


Figura 2.6: Proceso de metilación del ADN. Metilación en la posición 5' de la citosina catalizada por DNMT (metiltransferasas), en presencia de S-adenosil-metionina (SAM). Elaboración propia.

Tal y como se puede observar en la Figura 2.6, esta modificación genética ocurre en el contexto de un dinucleótido (CpG), consecución de los nucleótidos guanina (G) y citosina (C) en la secuencia que compone el ADN. Típicamente, el genoma se encuentra caracterizado por un contenido bastante bajo de CpG. Estos dinucleótidos se concentran en grandes grupos, llamados islas CpG (CGI) que se encuentran principalmente en las regiones promotoras de los genes⁴, Figura 2.7.

Casi el 60% de los promotores humanos se caracterizan por un alto contenido de CpG. Sin embargo, la densidad de CpG en sí misma no influye en la expresión génica. Esta regulación depende únicamente del ADN. Normalmente, las CGI no están metiladas en genes transcripcionalmente activos. Sin embargo, la hipermetilación está asociada con el silenciamiento génico y, juega un papel importante en la inactivación de los genes supresores de tumores [21].

³La epigenética es un campo emergente de la ciencia que estudia los cambios hereditarios causados por la activación y desactivación de los genes sin ningún cambio en la secuencia de ADN subyacente del organismo.

⁴Los genes son segmentos de ADN que contienen información para elaborar una proteína específica. Los genes están compuestos fundamentalmente por una región promotora que marca el inicio de la transcripción, exones que contienen la información genética e intrones que tienen un papel regulatorio.

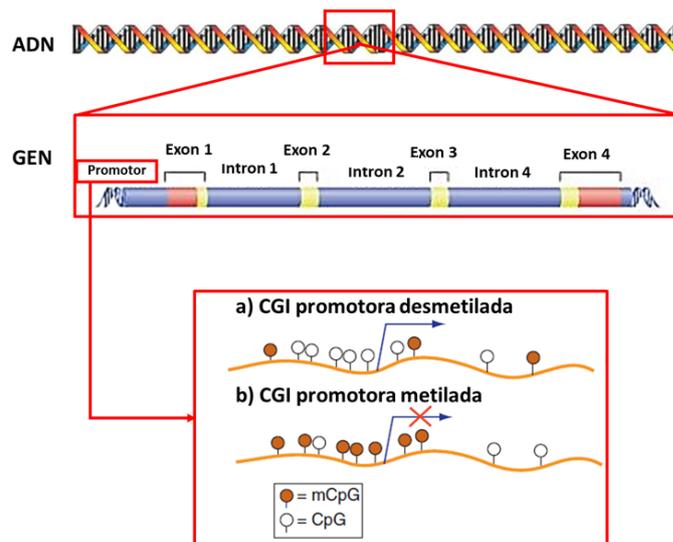


Figura 2.7: a) Regiones CpGs desmetiladas dentro de la región promotora. b) Regiones CpGs metiladas dentro de las zonas promotoras. Elaboración propia.

2.3.2 Técnicas de adquisición de la metilación del ADN

El análisis de la metilación del ADN ha experimentado una revolución durante la última década, debido especialmente a la adaptación de la tecnología de microarrays al estudio de metilación y a la aparición de secuenciación de nueva generación, *Next-Generation Sequencing* (NGS). Existen diferentes técnicas destinadas al estudio del nivel de metilación del ADN y, el método a utilizar depende de los objetivos del experimento. Dado que la información de la metilación del ADN se “borra” tras la reacción en cadena de la polimerasa (PCR), debido a la ausencia de metiltransferasas que mantengan el patrón de metilación, la gran mayoría de técnicas para cuantificar la metilación se basan en un tratamiento metil-dependiente previo a la amplificación o hibridación.

En la tabla 2.1 aparecen de forma esquematizada las principales técnicas utilizadas para obtener los datos de metilación a analizar. Como se puede observar en esta tabla los principios pretratamientos a realizar para obtener los datos de metilación son:

- **Conversión con bisulfito sódico.** Durante el tratamiento con bisulfito, las citosinas no metiladas en el ADN monocatenario se desaminan para dar uracilo, dejando intacta la citosina metilada. Por tanto, este tratamiento convierte un fenómeno epigenético en una diferencia genética y, en consecuencia, analizable mediante diferentes técnicas. La conversión de bisulfito es el enfoque más convencional para el pretratamiento y se considera el estándar de oro para determinar el estado de metilación del ADN porque ofrece una resolución de CpG simple [22].
- **Digestión con enzimas de restricción.** El tratamiento por enzimas de restricción se basa en la capacidad de algunas endonucleasas de restricción⁵ sensibles a la metilación (HpaII, MspI y HhaI) y sus isoenzimas correspondientes (no sensibles a la metilación) para distinguir las citosinas metiladas de las no metiladas. La presencia de metilación en el sitio de restricción puede inhibir la

⁵Una enzima de restricción (endonucleasa) es aquella que puede reconocer una secuencia característica de nucleótidos dentro de una molécula de ADN y cortar el ADN en ese punto en concreto, llamado sitio o diana de restricción.

acción de estas enzimas por lo que su patrón de acción puede ser usado para determinar el estatus de metilación de un determinado locus. Las endonucleasas tienen algunos inconvenientes como la presencia de una alta tasa de falsos positivos debido a una digestión incompleta [22].

- Enriquecimiento por afinidad.** Los métodos basados en el enriquecimiento se basan en la inmunoprecipitación del ADN genómico con anticuerpos específicos para la citosina metilada. Es decir, se captura la fracción metilada del ADN mediante el uso de anticuerpos dirigidos contra las citosinas metiladas, o bien proteínas que se unen al ADN metilado. Un punto importante con respecto a las técnicas basadas en afinidad es que miden la densidad de la metilación en una región específica. Por lo tanto, un tramo metilado de ADN donde los sitios objetivo son escasos podría ser difícil de diferenciar de una región no metilada [22].

La combinación de los diferentes métodos de pretratamiento y de las técnicas de biología molecular posteriores, como los microarrays de ADN y la secuenciación de alto rendimiento, genera una gran cantidad de técnicas para mapear la metilación del ADN de manera factible a escala de todo el genoma.

| Pretratamiento | Cuantitativo global | Locus basado en PCR | Todo el genoma (Array) | Todo el genoma (NGS) |
|---|---|---|------------------------|---------------------------|
| <i>Conversión con bisulfito sódico</i> | Alu/LINE-1-PCR-pyrosequencing ALU/LINE-1-HRM | MSP-PCR | Infinum Golden Gate | RRBS |
| | | MethylLight SMART-MSP MS-HRM Pyrosequencing MassARRAY | | WGSBS oxBS-Seq |
| <i>Digestión con enzimas de restricción</i> | HPLC | HpaII-PCR | DMH | Methyl-Seq |
| | LC-MS | MS-MLPA | MCAM | MCA-Seq |
| | LUMA | MS-FLAG | HELP CHARM MMASS | MSCC HELP-Seq |
| <i>Enriquecimiento por afinidad</i> | 5-mC ELISA | MeDIP-PCR | MeDIP-chip | MeDIP-Seq |
| | 5-mHC ELISA | MIRA | MIRA-chip | MethylCap-Seq MIRA-Seq |

HPLC = High-performance liquid chromatography; LC-MS = liquid chromatography-mass spectroscopy; LUMA = luminometric methylation assay; SMART-MSP = sensitive melting analysis after real-time methylationspecific PCR; MS-HRM = methylation-specific high-resolution melting analysis; CHARM = comprehensive highthroughput arrays for relative methylation; MMASS = microarray-based methylation assessment of single samples; -seq = followed by sequencing; WGSBS = whole-genome shotgun bisulfite sequencing; MSCC = methylationsensitive cut counting; MethylCap = methyl-DNA binding domain capture technique

Tabla 2.1: Esquema de las diversas técnicas que se pueden utilizar para cuantificar la metilación de acuerdo con el pretratamiento que recibe el ADN y el número de nucleótidos modificados y la resolución de la detección (metilación global, de región específica o genómica). Tabla extraída de [22].

Como se puede observar en la Tabla 2.1, la técnica de representación reducida de secuenciación de bisulfito (RRSB) permite obtener el análisis de metilación de todo el genoma. Sin embargo, esta técnica combina la secuenciación de bisulfito con la utilización de enzimas de restricción, lo que permite que esta se dirija a las regiones promotoras de los genes que, como se ha comentado anteriormente, albergan las islas CpG donde principalmente se produce la metilación. Por tanto, se puede considerar una técnica eficiente y de alto rendimiento para el análisis de la metilación. Es por ello que en este proyecto se ha hecho uso de esta técnica para el estudio del nivel de metilación.

2.3.2.1 *Secuenciación de bisulfito de representación reducida*

La secuenciación de bisulfito de representación reducida, *Reduced representation bisulfite sequencing* (RRBS) es un método desatinoado al estudio de metilación del ADN a escala de todo el genoma con resolución de un solo nucleótido. RRBS es una variante de la secuencia de conversión de bisulfito del genoma completo en la que se utiliza la digestión con enzimas de restricción y la selección del tamaño del ADN para enfocar el análisis en un subconjunto del genoma donde se produce la mayor parte de la metilación del ADN. Al centrarse en esta porción del genoma, se genera un conjunto de datos de metilación del ADN en todo el genoma a un costo de secuenciación de ADN más bajo que el WGBS.

El principio básico de RRBS es enfocar la cobertura en las regiones del genoma que contienen sitios CpG, que como se ha comentado anteriormente, son los principales lugares donde se produce la metilación del ADN. De esta forma, se reduce el coste de la secuenciación al no tener que analizar el genoma completo para obtener de forma cuantitativa el valor de la metilación. Con la utilización de enzimas de restricción específicas, es posible obtener datos de metilación de ADN con resolución de nucleótidos para aproximadamente 80 % de las islas de CpGs y más del 50 % de los promotores humanos, secuenciando únicamente el 3 % del genoma. A continuación, se explican los pasos del protocolo estándar para RRBS.

1. **Digestión enzimática:** después de la extracción de ADN, el primer paso en el protocolo RRBS es la digestión con enzimas de restricción insensibles a la metilación. Es fundamental que las enzimas no se vean influenciadas por el estado de metilación de los sitios CpG, ya que esto permite la digestión de las áreas metiladas y no metiladas. La enzima de restricción más común utilizada en RRBS es MspI. Esta enzima se dirige a las secuencias 5'CCGG 3' y escinde los enlaces fosfodiéster del dinucleótido CpG. Cuando se usa esta enzima particular, cada fragmento tiene un CpG en cada extremo. Esta digestión da como resultado fragmentos de ADN de varios tamaños.
2. **Reparación final y ligadura del adaptador:** después de digerir el ADN genómico, se reparan los voladizos de 5'CG y se agregan colas A. La cola A es obligatoria antes de la ligadura del adaptador Illumina porque los adaptadores contienen un saliente 3'-T. Los oligos adaptadores deben tener citosinas metiladas para que no se conviertan en uracilo en la siguiente etapa de conversión de bisulfito.
3. **Selección de fragmentos:** una vez que los adaptadores están ligados, se seleccionan los fragmentos de ADN que tienen el tamaño deseado. La selección de fragmentos puede llevarse a cabo mediante la carga de ADN en un gel de agarosa o con la utilización de bolas magnéticas (*bead-based size selection*) como se ha llevado a cabo en este proyecto, lo que minimiza posibles pérdidas de material. Según Gu et al., los fragmentos de ADN de 40-220 pares de bases son representativos de la mayoría de las secuencias promotoras y de las islas CpG [23].
4. **Conversión de bisulfito:** los fragmentos de ADN se tratan con bisulfito de sodio, que es un proceso que desamina la citosina no metilada en un uracilo. Las citosinas metiladas permanecen sin cambios, debido al grupo metilo que las protege de la reacción.
5. **Amplificación por PCR:** el ADN convertido con bisulfito se amplifica luego usando PCR con cebadores que son complementarios a los adaptadores de secuencia.
6. **Purificación por PCR:** antes de la secuenciación, el producto de PCR debe estar libre de reactivos no utilizados. Por lo tanto, se requiere un paso para la purificación por PCR. Esto se puede hacer con otro gel de electroforesis o con kits diseñados específicamente para la purificación por PCR.

7. **Secuenciación:** se utilizan enfoques de secuenciación de próxima generación. Los flujos de trabajo RRBS más comunes utilizan la plataforma Illumina NGS.
8. **Alineación y análisis de secuencias:** debido a las propiedades únicas de RRBS, se necesita un software especial para la alineación y el análisis [24]. El uso de MspI para digerir el ADN genómico da como resultado fragmentos que siempre comienzan con una C (si la citosina está metilada) o una T (si una citosina no estaba metilada y se convirtió en un uracilo en la reacción de conversión de bisulfito). Esto da como resultado una composición de pares de bases no aleatorias. Además, la composición base está sesgada debido a las frecuencias sesgadas de C y T dentro de las muestras. El alineamiento de secuencias se llevará a cabo empleando alineadores que integran el tratamiento con bisulfito tales como Bismark y BSMAP/RRBSMAP. La alineación con un genoma de referencia permite a los programas identificar pares de bases dentro del genoma que están metilados. Los datos resultantes de metilación del DNA serán visualizados con UCSC Genome Browser y con el Integrative Genomics Viewer.

En la Figura 2.8 aparecen de forma esquematizada los pasos que componen el protocolo estándar para la adquisición de datos de metilación mediante la técnica RRBS.

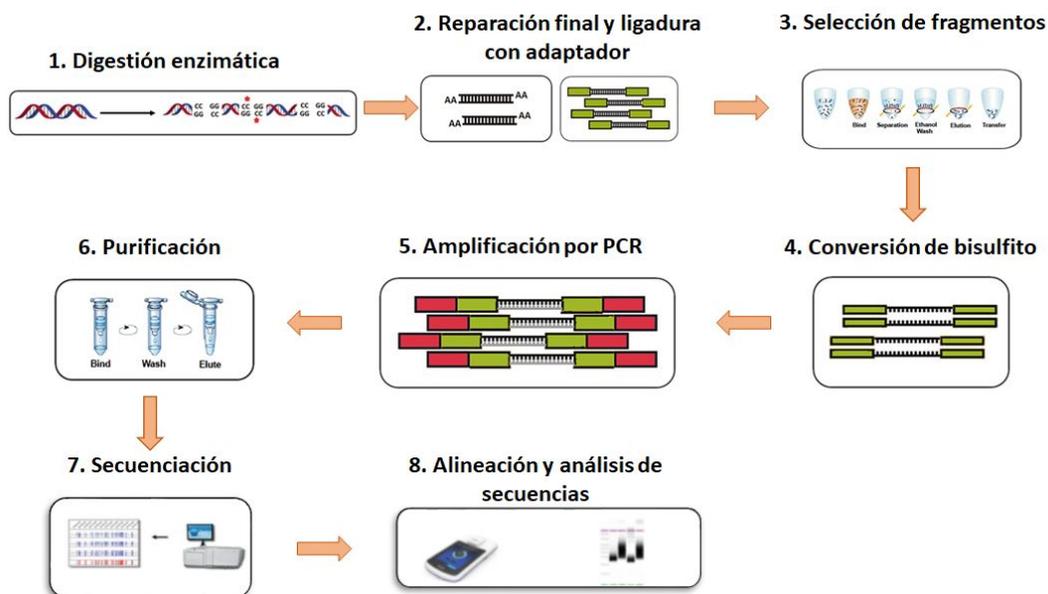


Figura 2.8: Protocolo estándar para la adquisición de datos de metilación mediante la técnica RRBS. Figura modificada de [25].

2.3.3 Metilación del ADN y melanoma

Como se ha comentado a lo largo de esta sección, la metilación del ADN ocurre típicamente en el contexto de los llamados sitios CpG. Los dinucleótidos CpG se encuentran infrarrepresentados en el genoma, concentrándose en la región promotora de hasta un 70% de los genes y formando agrupaciones denominadas islas CpG [26]. Estas islas CpG se encuentran habitualmente no metiladas, favoreciendo la expresión génica en presencia de factores de transcripción. Sin embargo, en el contexto de diversas enfermedades, fundamentalmente en el cáncer, se produce una hipermetilación de estas islas, lo que conlleva la silenciación génica [27]. Si este efecto tiene lugar en los genes supresores de tumores o en aquellos que controlan el ciclo celular y la apoptosis, la metilación del ADN desempeña un papel clave

en la carcinogénesis. Por otro lado, las secuencias repetitivas del ADN, que se encuentran en las regiones codificantes de los genes, se encuentran habitualmente metiladas y silenciadas. El mantenimiento de esta metilación es fundamental para preservar la integridad cromosómica. La hipometilación global del ADN, y por tanto la expresión de estas secuencias repetitivas, es un fenómeno frecuente en el cáncer, y contribuye a las alteraciones cromosómicas características en la carcinogénesis, Figura 2.9.

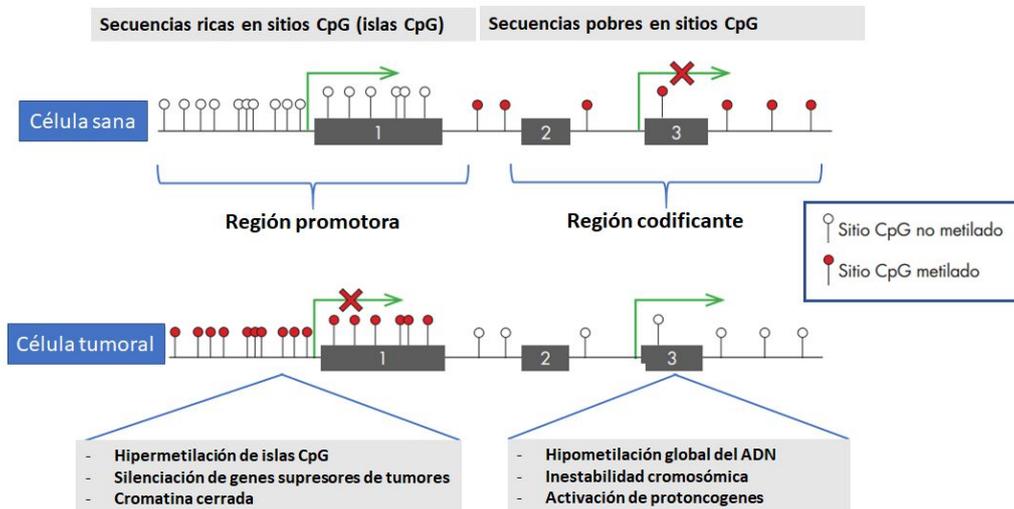


Figura 2.9: Alteración del patrón de metilación del ADN en el cáncer. En las células sanas, las islas CpG de las zonas promotoras de los genes supresores de tumores (1) se encuentran no metiladas, de forma que el gen se expresa normalmente. Por otro lado, las secuencias repetitivas del ADN (2, 3) se encuentran metiladas y por tanto, silenciadas. En el cáncer, el patrón de metilación se invierte, promoviendo la silenciación de genes supresores de tumores, y la reexpresión de secuencias repetitivas y protooncogenes. Figura modificada de [27].

Se han reconocido diversos patrones anormales de metilación del ADN en el melanoma en los últimos años y, hasta ahora se han descubierto más de 70 genes hipermetilados de forma aberrante en el melanoma cutáneo [26]. Se ha encontrado que estos genes están hipermetilados, ya sea por examen directo de los CpG metilados o indirectamente por su activación tras el tratamiento con agentes desmetilantes. La fuente de metilación aberrante del ADN en el melanoma, es decir, la hipometilación global y la hipermetilación focal, sigue siendo difícil de comprender [28]. En general, estos cambios podrían estar mediados por (1) un proceso activo, ya sea a través de una mayor actividad o sobreexpresión de enzimas DNMT; (2) una focalización aberrante de las enzimas DNMT, debido a un *splicing* alterado, la unión alterada o la expresión alterada de proteínas de andamiaje; o (3) un proceso pasivo debido a cambios en otras modificaciones epigenéticas que regulan la focalización de la metilación del ADN.

Por tanto, diversos estudios sugieren que existe una fuerte evidencia de que la metilación del ADN de genes seleccionados está relacionada con el melanoma y aumenta cuantitativa y cualitativamente con las etapas del tumor [26], [28], [29]. Por ello, una parte del presente TFM está dedicada al estudio de la relación existente entre los niveles de metilación de las islas CpG con un tipo concreto de tumores melanocíticos, los tumores melanocíticos spitzoides.

Capítulo 3

Materiales

Índice de contenidos

| | |
|---|----|
| 3.1 Base de datos de tumores melanocíticos spitzoides | 20 |
| 3.2 Software utilizado | 22 |
| 3.3 Hardware utilizado | 22 |

3.1 Base de datos de tumores melanocíticos spitzoides

Para el desarrollo de este proyecto, se ha contado con dos bases de datos obtenidas a partir de muestras de pacientes recopiladas en el Servicio de Anatomía Patológica del Hospital Clínico Universitario de Valencia desde 1990 hasta 2017. El primer conjunto de datos contiene información relevante al estudio molecular de metilación del ADN, mientras que el otro conjunto, está compuesto por imágenes histopatológicas. La información contenida en ambas bases de datos pertenece a los mismos pacientes, en concreto, 39 pacientes con lesiones melanocíticas spitzoides. De los 39 pacientes bajo estudio, 8 están diagnosticados como lesión melanocítica maligna (melanoma), 12 como lesión melanocítica benigna (nevus) y 19 como potencial maligno incierto (PMI). A continuación, se describen las principales características de cada base de datos.

■ *Base de datos de metilación de ADN*

El primer paso para obtener los datos de metilación de los 39 pacientes bajo estudio consiste en la extracción de los ácidos nucleicos. Para ello, se marca la zona tumoral óptima sobre la laminilla teñida de hematoxilina-eosina de cada paciente, se traslada dicha zona al bloque de parafina correspondiente y se realiza un cilindro (*punch*) empleando un *tissue arrayer* manual. Una vez se tiene la selección de la región de interés en el bloque de parafina, se procede a la extracción del ADN utilizando el kit *DNA-RNA Allprep FFPE extraction* de Qiagen. Tras la extracción el ADN, este se hibrida con el microarray (OncoScan) cuyas sondas se utilizan para capturar los alelos de más de 220.000 SNPs de regiones genómicas seleccionadas a través de todo el genoma en aproximadamente 900 genes relacionados con cáncer. Seguidamente, se realiza una digestión enzimática de las regiones genómicas seleccionadas y se amplifican las muestras obtenidas mediante PCR. Posteriormente, se lleva a cabo el estudio de metilación del DNA a nivel genómico mediante la metodología *Reduced Representation Bisulfite Sequencing* (RRBS), que ha sido explicada en la sección 2.3.2. En la Figura 3.1, aparece de forma esquematizada los pasos a seguir para obtener el porcentaje de metilación de las zonas CpG del tejido tumoral bajo estudio, valor entre 0 y 100, que es la información con la que posteriormente se trabajará. Tras la realización de estos procesos, se obtiene, para cada paciente, el porcentaje de metilación de 491 sitios CpG de los genes seleccionados mediante el microarray OncoScan. Para llevar a cabo un procedimiento de preselección y obtener los sitios CpG con la expresión de metilación más diferencial, eliminando aquellos que no se relacionan con la enfermedad bajo estudio, es necesario la realización de un análisis estadístico previo al desarrollo de cualquier método de clasificación.

■ *Base de datos de imágenes histopatológicas*

Tras digitalizar las muestras de hematoxilina-eosina de las biopsias que han sido utilizadas para obtener los datos de metilación, los patólogos utilizan una aplicación llamada '*MicroDraw*' para anotar las regiones de interés de cada imagen histológica [30]. *MicroDraw* es una aplicación web desarrollada sobre la librería *OpenSeadragon* que permite una visualización multi-resolución de las imágenes histológicas sin pérdida de calidad. La plataforma *MicroDraw* está escrita en JavaScript, empleando HTML5, CSS3 y jQuery, y ha sido modificada por los ingenieros de CVBLab para realizar anotaciones en las imágenes de tumores benignos, malignos o de potencial maligno incierto (véase Figura 3.2). La anotación de estas imágenes es necesaria pues toda la muestra extraída no tiene rasgos histopatológicos que definan la existencia de tumores melanocíticos. Además, para el aprendizaje y validación de los métodos de clasificación propuestos, es necesario contar con una etiqueta que defina a qué clase pertenece cada muestra.

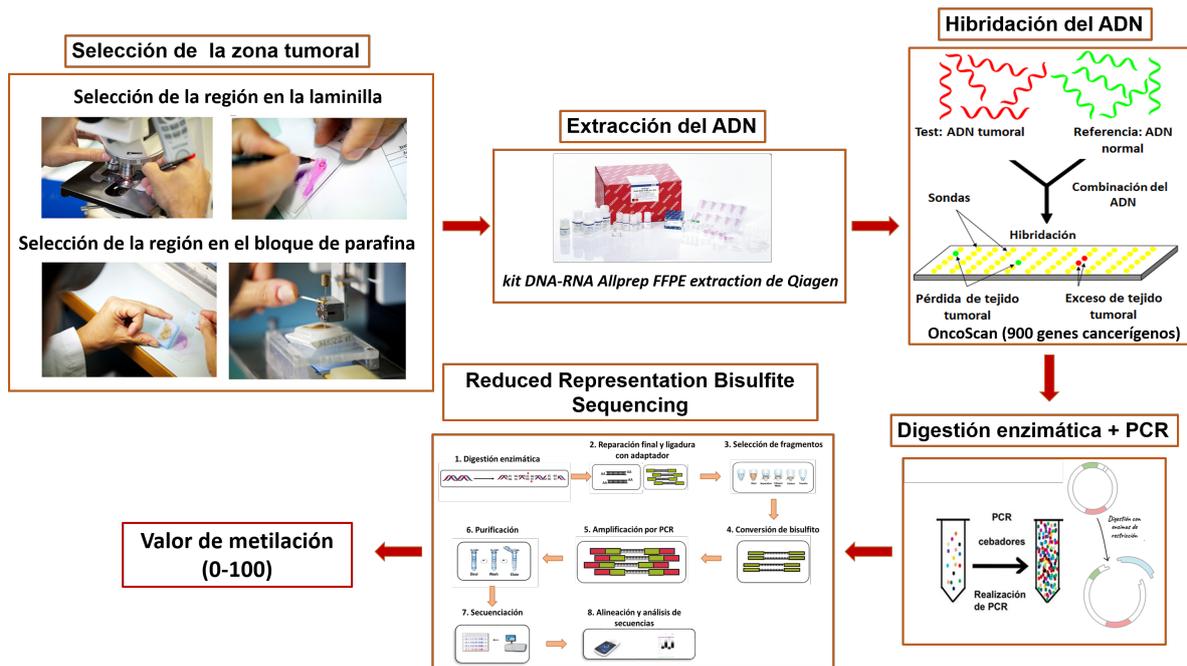


Figura 3.1: Proceso llevado a cabo para obtener el valor de metilación del tejido bajo estudio. Elaboración propia.

Para poder utilizar la aplicación de MicroDraw, en primer lugar, 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 *.tif, en otras con diferentes niveles de resolución en formato *.dzi. Esto permite que MicroDraw pueda ir cargando la imagen *.dzi y el patólogo pueda cambiar el nivel de aumento en función de sus necesidades, sin modificar en ningún momento la calidad original de la imagen. De esta forma, es mucho más sencillo identificar la región de interés.

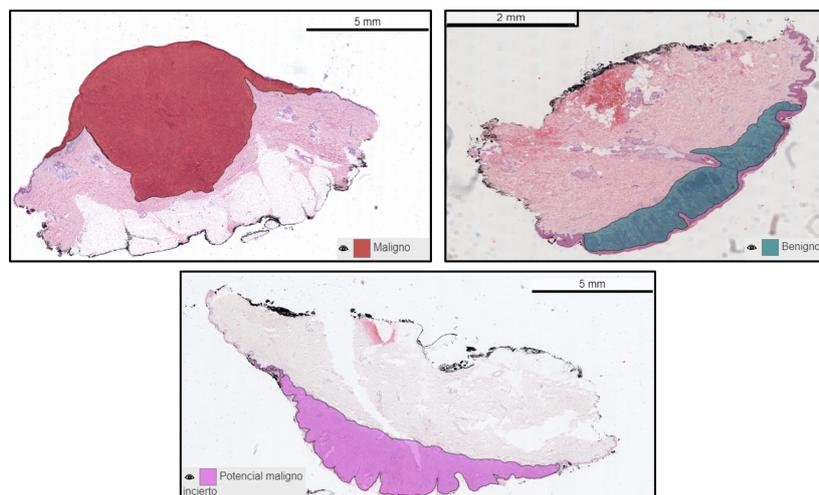


Figura 3.2: Anotación realizadas por los patólogos en muestras con neoplasias benignas, malignas y de potencial maligno incierto. Elaboración propia.

Las imágenes histopatológicas con las que se trabaja en este proyecto se encuentran digitalizadas a una magnificación de 40x con el objetivo de disponer de una resolución espacial lo más alta posible y captar así los detalles más finos de las imágenes. Debido a este nivel de detalle, el tamaño de una imagen

histopatológica completa es del orden de los *gigabyte*, lo que hace computacionalmente imposible procesarla y analizarla. Debido a la carga computacional que suponen estos datos, es necesario llevar a cabo un pre-procesado de las imágenes histopatológicas con el objetivo de extraer *patches*¹ característicos procedentes del tejido tumoral bajo estudio.

3.2 Software utilizado

Para la implementación de los algoritmos propuestos y el análisis de los resultados obtenidos en este proyecto, se han empleado los entornos de programación Python3.5 y MATLAB® v.R2019a. Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel, que utiliza una semántica dinámica [31]. El entorno de programación de Python admite la integración de módulos y paquetes, lo que fomenta la modularidad del programa y la reutilización de código. Uno de los módulos ampliamente conocido, que ha sido utilizado en este proyecto, es Keras. Este módulo permite el desarrollo de algoritmos basados en aprendizaje profundo a alto nivel, creando un entorno de programación mucho más amigable para el usuario. Para la etapa de entrenamiento de los diferentes algoritmos de clasificación basados en aprendizaje profundo, se ha hecho uso de la librería NVIDIA Cuda® Deep Neural Network (cuDNN) [32]. Como entorno de desarrollo integrado compatible con el lenguaje de programación Python se ha utilizado PyCharm y, dado que el entrenamiento de los algoritmos se ha ejecutado en un servidor remoto por las altas prestaciones computacionales que se requieren, se ha hecho uso del programa MobaXterm [33]. Esta herramienta permite generar una conexión SSH (Secure SHell) facilitando la comunicación segura entre el cliente (el ordenador de trabajo) y el servidor (donde se ejecutan los distintos algoritmos).

Con el objetivo de preparar los datos bajo estudio y validar los resultados obtenidos, se ha hecho uso de software MATLAB. MATLAB®, cuyo nombre proviene de *MATrix LABORatory*, es una plataforma desarrollada para resolver problemas relacionados con la rama de la ingeniería y las ciencias que incorpora un lenguaje de alto nivel basado en operaciones con matrices. Se utiliza para el desarrollo de aprendizaje automático y visión artificial, procesamiento de señales e imágenes, finanzas computacionales y otros muchos campos. Además, se encuentra disponible para las plataformas Unix, Windows, Mac OS X y GNU/Linux [34].

3.3 Hardware utilizado

La mayor parte del proyecto ha sido desarrollado en un equipo compuesto por un procesador intel® Core™ i7-4770K @3.4 GHz con un sistema operativo Windows 10 de 64 bit. En cuanto a la memoria RAM, la capacidad es de 16 GB y, además, el equipo cuenta con una tarjeta gráfica NVIDIA GeForce GTX650.

Como se ha expuesto anteriormente, para el entrenamiento de los algoritmos desarrollados, se requieren servidores que cuenten con unidades de procesamiento gráfico más potentes. Es por ello que, para llevar a cabo este proceso, se ha hecho uso de un servidor de computación de altas prestaciones perteneciente al grupo de investigación CVBLab. Este está compuesto por un procesador Intel i7 @4.20GHz, 32GB de RAM y una tarjeta gráfica NVIDIA Titan V que cuenta con 12 GB de memoria HBM2, 5120 núcleos CUDA y 640 núcleos tensor que precisamente están destinados a multiplicar por 9 su rendimiento en el área del *deep learning* y de la inteligencia artificial.

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

Por otro lado, dado que en este proyecto se ha trabajado con gran cantidad de imágenes que ocupan mucho espacio en disco, se ha hecho uso del servidor NAS Synology DS918 de 32 TB de capacidad para almacenarlas. Así se ha establecido una diferencia entre el servidor de computación, donde se han almacenado y ejecutado los *scripts* correspondientes al código implementado y, el servidor NAS de datos, en el que se han almacenado todos los datos necesarios para llevar a cabo el proceso de entrenamiento, validación y testeado de los algoritmos.

Capítulo 4

Metodología

Índice de contenidos

| | | |
|-------|--|----|
| 4.1 | Preprocesado de los datos | 28 |
| 4.1.1 | Selección de características relevantes para la base de datos de metilación | 28 |
| 4.1.2 | Generación de patches para la base de datos de imágenes histopatológicas | 31 |
| 4.2 | Acondicionamiento de los datos | 32 |
| 4.2.1 | Reducción de la dimensionalidad de los datos de metilación | 32 |
| 4.2.2 | Balanceo y creación de muestras sintéticas para la base de datos de imágenes histológicas. | 37 |
| 4.3 | Partición de los datos | 38 |
| 4.4 | Algoritmos de clasificación para neoplasias melanocíticas spitzoides | 40 |
| 4.4.1 | Algoritmos de clasificación supervisada para los datos de metilación | 40 |
| 4.4.2 | Algoritmos de clasificación no supervisada para los datos de metilación | 43 |
| 4.4.3 | Algoritmos de clasificación supervisada para imágenes histopatológicas | 45 |
| 4.4.4 | Algoritmos de clasificación combinando modalidades | 50 |
| 4.5 | Algoritmos para el pronóstico de neoplasias melanocíticas spitzoides | 50 |
| 4.5.1 | Algoritmos para el pronóstico basados en datos de metilación. | 51 |
| 4.5.2 | Algoritmos para el pronóstico basados en imágenes histopatológicas | 53 |

Con el objetivo de presentar de forma clara y ordenada las diferentes fases que componen esta sección, en la Figura 4.1 se expone un diagrama de flujo donde se pueden observar las diferentes etapas correspondientes al proceso de clasificación y pronóstico de las lesiones melanocíticas spitzoides.

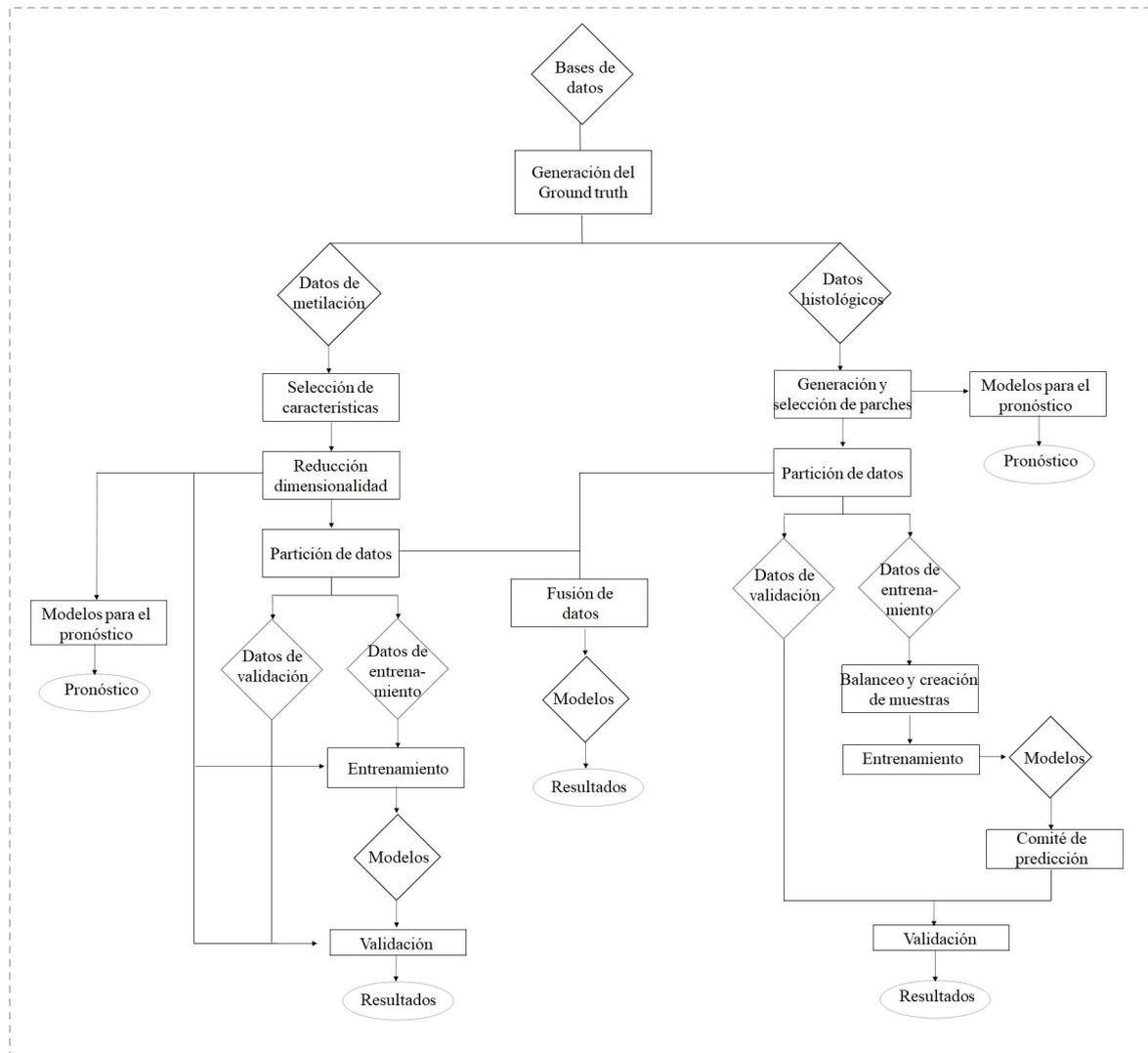


Figura 4.1: Diagrama de flujo de los principales procesos que componen la metodología aplicada en este proyecto.

Como se ha explicado en la sección 3.1 para el desarrollo de este trabajo se ha contado con dos bases de datos en relación a las lesiones melanocíticas spitzoides. Por un lado, se tiene información procedente de los estudios de metilación del ADN, mientras que por otro lado, se tienen datos correspondientes a las imágenes histopatológicas. El primer proceso a desarrollar consiste en el preprocesado de los datos de lesiones melanocíticas malignas y benignas. Para ello, se desarrollan algoritmos de selección de las características para los datos de metilación y la generación y selección de *patches* para las imágenes histopatológicas. Posteriormente, se desarrollarán algoritmos de acondicionamiento de los datos. Para ello, se implementarán métodos de reducción de la dimensionalidad sobre toda la base de datos de metilación de lesiones melanocíticas benignas y malignas. Tras el desarrollo de estos algoritmos se procede a la partición de estos datos para entrenar los algoritmos supervisados que se desarrollarán con los datos de metilación. Esta partición de datos será trasladará a los *patches* de las imágenes histológicas

y sobre los datos de entrenamiento se llevará a cabo el acondicionamiento de los datos, que en este caso, consiste en el balanceo y en la creación de nuevas muestras. Con respecto a los modelos de clasificación para las lesiones melanocíticas basadas en los datos de metilación, estos estarán basados en algoritmos supervisados y no supervisados. Por tanto, los algoritmos no supervisados no necesitarán la partición de datos mencionada y serán entrenados y validados con toda la base de datos correspondiente a los datos con lesiones melanocíticas benignas y malignas. En contraposición, los algoritmos supervisados se entrenarán con los datos de entrenamiento y se validarán con los datos de validación con el fin de obtener los resultados. En el caso de las imágenes histopatológicas, los algoritmos de clasificación se entrenarán con los diferentes *patches* de los pacientes de los *set* de entrenamiento. Para obtener los resultados por paciente correspondientes a esta etapa, se desarrollarán algoritmos que lleven a cabo un comité de predicción sobre las muestras de validación. Cabe destacar que tras la partición de datos implementada, se desarrollarán algoritmos de clasificación que combinen ambas modalidades de datos.

Tras el desarrollo de los modelos de clasificación de lesiones melanocíticas malignas y benignas, se desarrollarán algoritmos para el pronóstico de las lesiones melanocíticas de potencial maligno incierto. En la implementación de algoritmos para el pronóstico basados en datos de metilación, se utilizarán los algoritmos de selección de características y reducción de la dimensionalidad entrenados con las muestras benignas y malignas para obtener las características discriminatorias de los pacientes con PMI. Posteriormente, se utilizarán estas características para desarrollar algoritmos de pronóstico. Con respecto a las imágenes histopatológicas, se aplicarán los algoritmos de generación de *patches* y los modelos entrenados con muestras benignas y malignas para predecir las características discriminatorias de los pacientes con PMI que serán utilizadas para el desarrollo de algoritmos de pronóstico.

4.1 Preprocesado de los datos

A continuación, se exponen los pasos previos realizados sobre las bases de datos bajo estudio con el objetivo de preparar, para su posterior clasificación, tanto los datos de metilación del ADN como las imágenes histopatológicas.

4.1.1 Selección de características relevantes para la base de datos de metilación

El objetivo de esta etapa es seleccionar de los sitios CpG de los que se ha adquirido los valores de metilación, en total **491**, únicamente aquellos que presenten una información relevante de cara a resolver el problema de clasificación binario propuesto: benigno vs maligno. Cabe destacar que, como se ha comentado en la sección 3.1, se ha obtenido el nivel de metilación de los genes relacionados con varios tipos de cáncer que se encuentran en el microarray OncoScan, por lo que todos los sitios CpG obtenidos no tienen por qué estar relacionados con las lesiones melanocíticas spitzoides. Para seleccionar los sitios CpG de interés, se establece una serie de contrastes de hipótesis cuya finalidad es descartar las variables que aportan información redundante y aquellas que no guardan una relación de dependencia con la clase. Como se ha expuesto en la sección 3.1, la técnica RRBS informa del porcentaje de metilación obtenido, proporcionando un valor entre 0 y 100. Por tanto, el primer paso para llevar a cabo el análisis de características consiste en normalizar los valores de metilación de cada sitio CpG al rango 0-1, dividiéndolos entre 100.

Una vez que los datos se encuentran normalizados, se analiza la independencia entre variables, independencia entre sitios CpG, con el objetivo de descartar aquellas que no aportan información en el desarrollo de los modelos. Para realizar este análisis, se ha utilizado la correlación en valor absoluto, ya que interesa eliminar las variables correladas independientemente del sentido de la tendencia. En la Figura 4.2 se expone la matriz de correlación en la que se representa el coeficiente de correlación obtenido al enfrentar pares de variables de la base de datos de metilación del ADN.

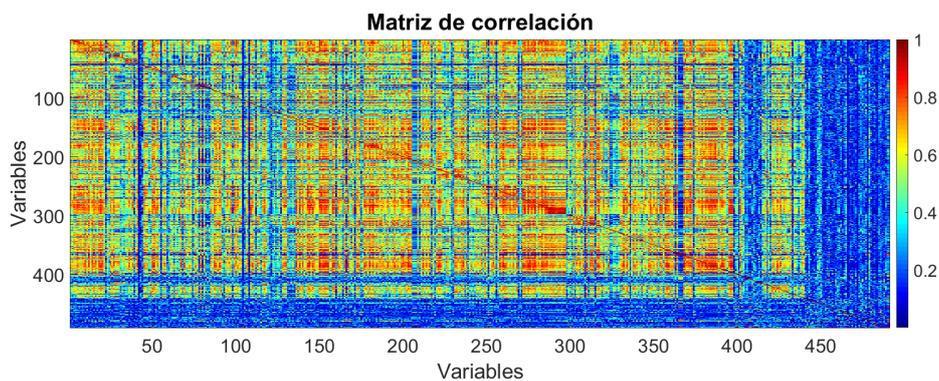


Figura 4.2: Mapa de correlación relativo a los 491 sitios CpG de los que se ha adquirido el valor de metilación.

A nivel cualitativo, los tonos más cálidos de la matriz de correlación representan los coeficientes de correlación más altos, mientras que los tonos más fríos, se corresponden con los coeficientes de correlación más bajos. Como se puede observar, a nivel cualitativo no es posible llevar a cabo con exactitud una selección de las variables independientes. Es por ello que se realiza un análisis cuantitativo consistente en la realización de un contraste de hipótesis. Para ello, se plantean dos hipótesis: la hipótesis nula representada como H_0 , que es la hipótesis que se desea contrastar y, aquella que se mantendrá a no ser

que los datos indiquen su falsedad, y la hipótesis alternativa, H_1 , que niega la hipótesis nula. En este caso, $H_0 = \text{Las variables son independientes}$ y $H_1 = \text{Las variables no son independientes}$. Además, se fija el nivel de significancia $\alpha = 0.05$, el cual determina la región de rechazo y aceptación, indicando para que valores de la muestra se rechaza o no la hipótesis nula. El cálculo de la correlación proporciona dos valores: el $p\text{-value}$ y un porcentaje que determina el grado de correlación. En función del $p\text{-value}$ se acepta o no la hipótesis nula, es decir, se determina si dos variables están correlacionadas o no. No obstante, para este proyecto se define adicionalmente un umbral $t = 0.90$ para descartar únicamente las variables que estén correlacionadas más de un 90%. Por tanto, en este estudio, se descartan aquellas variables que tienen un $p\text{-value}$ menor al nivel de significancia (α) pero que además aportan menos de un 10% de información nueva.

Una vez que se ha llevado a cabo el estudio de independencia entre variables, se realiza el estudio de dependencia entre las variables y la clase para establecer que variables tienen mayor capacidad discriminativa. Para ello, en primer lugar, es necesario conocer la distribución que sigue cada variable. Si la variable sigue una distribución normal, se realizará un estudio $t\text{-student}$ en el que se compararán las medias de los valores de dicha variable asociadas a cada clase. Por el contrario, si la variable no sigue una distribución normal, como sucede en la mayoría de variables que representan información sobre la metilación, la comparación se llevará a cabo en base a la mediana, es decir, se realizará un *Wilcoxon Rank-Sum Test*. Para conocer la distribución de cada variable, se realiza un estudio denominado *test de Kolmogorov-Smirnov* que permite comparar la distribución de la variable en cuestión con otra de referencia de media $\mu = 0$ y desviación típica $\sigma = 1$. Previamente al estudio de la distribución de cada variable es necesario tipificarlas. Para ello, se utiliza la fórmula del $z\text{-score}$:

$$z_i = \frac{x_i - \bar{x}}{s} \quad (4.1)$$

donde x_i hace referencia a cada una de las variables, \bar{x} representa su valor medio, s la desviación típica y z_i es el resultado tras el proceso de tipificación.

Tras la tipificación de todas las variables, se lleva a cabo el contraste de hipótesis para averiguar la distribución de cada variable. Como se ha explicado anteriormente, se define un nivel de significancia, en este caso $\alpha = 0.05$, y se establece $H_0 = \text{La variable sigue una distribución normal}$ y $H_1 = \text{La variable no sigue una distribución normal}$. Por tanto, si tras realizar el test de *Kolmogorov-Smirnov*, se obtiene un $p\text{-value}$ menor al nivel de significancia propuesto, se determina que con el 95% de confianza se rechaza la hipótesis nula y por tanto, la variable dada no sigue una distribución normal. En este caso, se analizará la mediana de la variable para establecer su dependencia con respecto a la clase. En el caso de que el $p\text{-value}$ sea mayor que el nivel de significancia, no se puede rechazar la hipótesis nula y por tanto, se comparará la media. En la Figura 4.3, se exponen ejemplos representativos de tres variables de la base de datos estudiada, en las que visualmente se puede comprobar si estas siguen o no una distribución normal. Respecto al histograma de ajuste de distribución representado en la Figura 4.3, si el histograma que representa cada variable concuerda con una campana de Gauss, se puede afirmar que la variable sigue una distribución normal. En el caso de la gráfica cuantil-cuantil, si la nube de puntos se adapta a la línea recta especificada por los puntos rojos, se puede concluir que la variable dada sigue una distribución normal.

A nivel cualitativo, la Figura 4.4 muestra si una variable presenta mayor o menor poder discriminativo. Para ello, se han seleccionado de forma aleatoria 3 variables de la base de datos. Este diagrama, denominado diagrama de cajas y bigotes, permite enfrentar para una misma características los valores que toma en

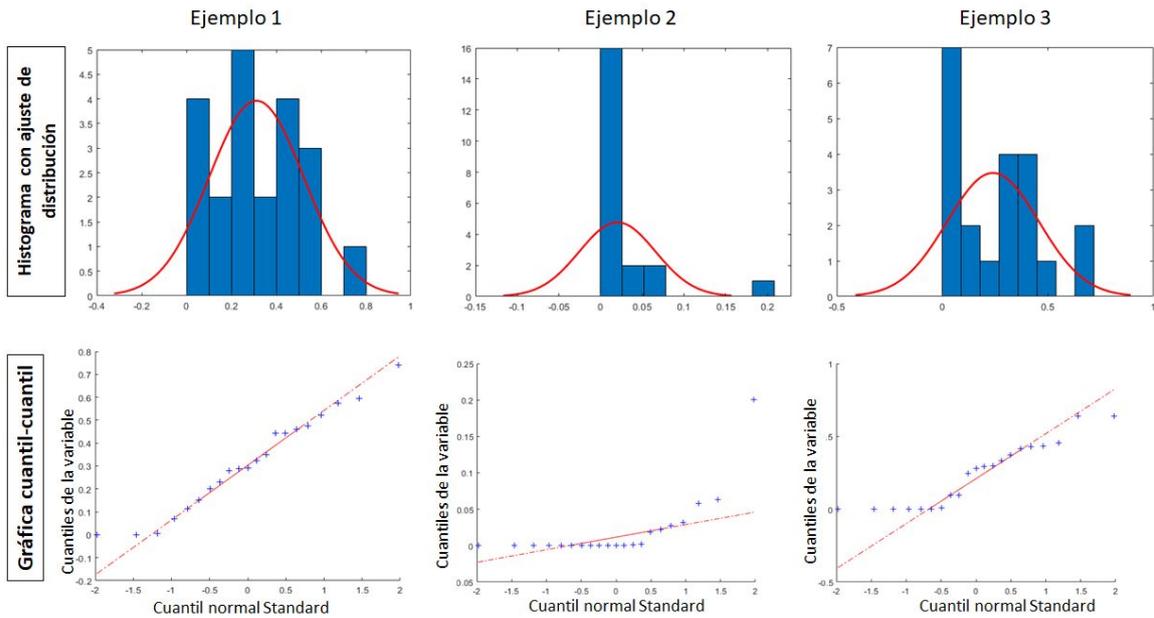


Figura 4.3: Histograma con ajuste de distribución y gráfico cuantil-cuantil para tres variables representativas de la base de datos.

función de cada clase. Si el rango intercuartil de los datos, que representa el 50% de los datos, con respecto a cada clase se encuentra solapado, la variable estudiada no tiene poder discriminatorio para separar las clases estudiadas. Si por el contrario, el rango intercuartil de cada variable aparece separado, esa variable tendrá poder discriminatorio para separar las clases.

Tras el estudio de independencia entre variables y dependencia con respecto a la clase, se obtienen un total de **62 sitios CpG discriminantes**. Estas características serán el input de la siguiente etapa. Cabe destacar que el análisis de independencia de variables y dependencia con respecto a la clase, se ha llevado a cabo sobre los pacientes con lesiones melanocíticas malignas y benignas, que son los datos que se utilizarán para entrenar y validar los modelos. Posteriormente, los sitios CpG discriminativos se transferirán a los pacientes con potencial maligno incierto, con el objetivo de pronósticar su evolución en maligno o benigno, tal y como se realizaría con nuevas muestras a predecir por los algoritmos.

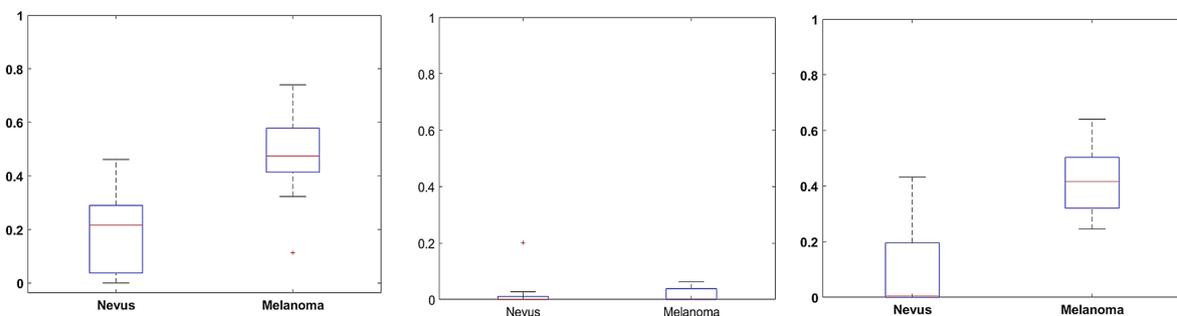


Figura 4.4: Análisis discriminatorio de tres variables representativas de la base de datos con respecto a la clase.

4.1.2 Generación de patches para la base de datos de imágenes histopatológicas

Con el objetivo de poder analizar las imágenes WSI procedentes de las biopsias de tumores melanocíticos spitzoides, se procede a la generación de sub-imágenes o *patches* de las mismas, Figura 4.5. En este caso, se han generado *patches* con una dimensión de 512x512x3 píxeles. Una vez que se han obtenido los *patches* procedentes de las imágenes WSI, se procede a la selección de aquellos que serán utilizados para la distinción de los tumores melanocíticos spitzoides. Para ello, se eliminan todos los *patches* que no se encuentran anotados por el patólogo, Figura 4.5 (a), ya que no pertenecen a la región de interés bajo estudio. Una vez que se obtienen los *patches* anotados por los patólogos, se seleccionan de estos aquellos que contienen más de un 20% de tejido y más de un 25% de anotación (región roja sobre los *patches* de las imágenes). De esta forma se eliminan todos los *patches* como los mostrados en la Figura 4.5 (c). Como se puede observar en la Figura 4.5, estos *patches* se caracterizan por la presencia de una gran cantidad de fondo, por lo que contienen poca información relativa al tejido histológico. Además, la región anotada por el patólogo no es representativa dentro del *patch*, por lo que se puede concluir que estos *patches* no son distintivos para la clasificación de lesiones melanocíticas spitzoides. Por tanto, el input a los algoritmos de clasificación desarrollados serán aquellos parches como los representados en la Figura 4.5 (b), que tienen poca información relativa al fondo y que contienen una región de interés representativa.

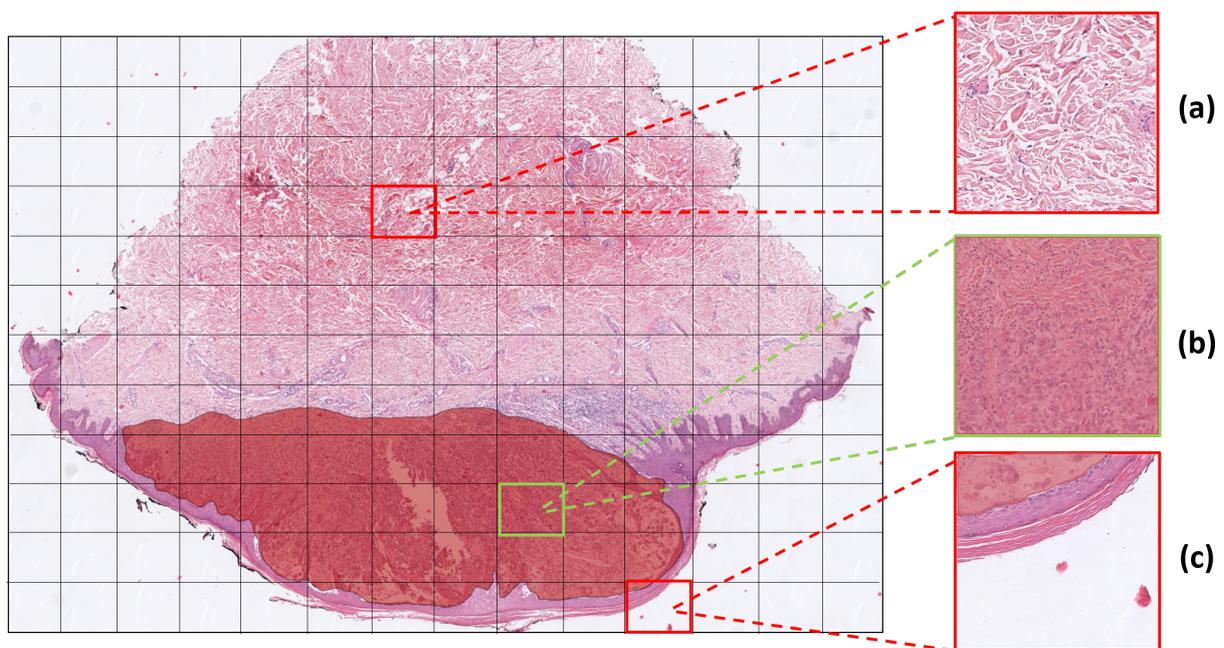


Figura 4.5: Patches generados de las imágenes WSI. (a) Patch de una zona no anotada por el patólogo. (b). Patch de una región representativa de las lesiones melanocíticas spitzoides. (c) Patch que no contiene suficiente región de interés.

En la Tabla 4.1 se muestra el número de *patches* de cada uno de los pacientes con melanoma y tumores benignos que se utilizarán para el entrenamiento de los algoritmos de clasificación desarrollados.

Como se puede observar en esta tabla, el número de *patches* obtenidos varía considerablemente entre los distintos pacientes. Esto se debe a causa de las irregularidades que presentan estas lesiones, el tamaño de los tumores es muy diferente entre pacientes. Este hecho, unido a que el número de pacientes utilizado en este estudio es relativamente bajo, hace imprescindible la implementación de técnicas que permitan

| Melanoma | | Nevus | |
|-------------|------------|------------|------------|
| ID | N_ Patches | ID | N_ Patches |
| 15B0000744 | 248 | 00B009136 | 28 |
| 15B0012535 | 23 | 09B0001931 | 320 |
| 17B0025620 | 43 | 10B008110 | 29 |
| 18B002795A | 41 | 11B0008865 | 70 |
| 18B007048A | 99 | 12B0027444 | 22 |
| 18B0013319A | 152 | 13B0027110 | 15 |
| 18B0013387A | 21 | 13B0029948 | 38 |
| 18B0016343A | 226 | 14B009803 | 16 |
| | | 14B0013206 | 39 |
| | | 17B001080 | 8 |
| | | 17B0003166 | 28 |
| | | B334721 | 11 |

Tabla 4.1: Número de *patches* obtenidos para cada uno de los pacientes con lesiones melanocíticas malignas y benignas.

lidar con el problema de desbalanceo entre los distintos *patches* de los pacientes de cada clase de neoplasia melanocítica spitzoide.

4.2 Acondicionamiento de los datos

Una vez que se llevado a cabo el preprocesado de los datos de metilación así como la selección de los *patches* significativos de las imágenes WSI, es necesario realizar un acondicionamiento de los datos antes de implementar cualquier algoritmo de clasificación. A continuación, se lleva a cabo una descripción del acondicionamiento de los datos llevado a cabo para cada una de las bases de datos bajo estudio.

4.2.1 Reducción de la dimensionalidad de los datos de metilación

El análisis de metilación del ADN ha experimentado una revolución durante la última década debido, especialmente, a la adaptación de la tecnología de microarrays al estudio de la metilación y a la aparición de la secuenciación de próxima generación [35], [36]. Estos avances tecnológicos combinados con el desarrollo de técnicas como la secuenciación de bisulfito de alto rendimiento han permitido la realización de análisis de metilación del ADN a nivel molecular [37]. Este desarrollo hace que los estudios de metilación actuales generen una gran cantidad de datos. Las dimensiones extremadamente altas de los datos de metilación, comparadas con el número generalmente pequeño de muestras disponibles, hace necesario el desarrollo de algoritmos de reducción de la dimensionalidad de los datos, antes de implementar algoritmos de clasificación, que permitan obtener una dimensión de estos al número de muestras disponibles [38]. En este contexto, Yuvaraj et al. presentaron diferentes algoritmos para la reducción de la dimensionalidad de los datos de metilación basados en el análisis de componentes principales (PCA) y en el criterio de Fisher [38]. Sin embargo, los conjuntos de datos de metilación del ADN no gaussianos no pueden describirse eficientemente mediante estos métodos en los que la gaussianidad de los datos es asumida. Jazayer et al. usaron una factorización de matriz no negativa (NMF) para la reducción de la dimensión de los datos de metilación, procedentes de tejidos mamarios, con la posterior utilización de los clasificadores supervisados *Extreme Learning Machine* (ELM) y *Support Vector Machines* (SVM). Sin embargo, con el algoritmo NMF no es posible transferir los datos de entrada a un espacio dimensio-

nal más pequeño que el número de datos totales ya que este método transfiere los datos a una segunda dimensión menor o igual al rango de los datos originales [39].

Recientes avances en el campo de la inteligencia artificial han permitido el desarrollo de algoritmos de aprendizaje profundo que llevan a cabo la incrustación de estados de metilación de los sitios CpG para extraer características biológicamente significativas de menor dimensión. Zhongwei et al. presentaron un conjunto de capas *Random Boltzmann Machine* (RBM) con el objetivo de reducir la dimensionalidad de datos de metilación de tejido mamario [40]. Khwaja et al. propusieron un algoritmo no supervisado denominado *autoencoder* para la reducción de la dimensionalidad de datos de metilación de varios tipos de cáncer: cáncer de mama, carcinoma de pulmón, leucemia linfoblástica y tumores urológicos [41]. Titus et al. propusieron un marco de aprendizaje profundo no supervisado con *variational autoencoder* (VAE) para aprender representaciones latentes de la metilación del ADN a partir de tres conjuntos de datos de tumores de mama [42].

Como se puede observar, los métodos más novedosos presentes en el estado del arte que se han aplicados para la reducción de la dimensionalidad de los datos de metilación del ADN en distintos tipos de cáncer son algoritmos basados en *autoencoders*. Es por ello que, en el presente TFM, se proponen por primera vez algoritmos basados en *autoencoder y variational autoencoder* para la reducción de la dimensionalidad de los datos de metilación obtenidos a partir de neoplasias melanocíticas spitzoides.

■ *Autoencoder*

El método de *autoencoder* (AE) es uno de los algoritmos más significativos en la representación de datos no supervisada. El objetivo de este método consiste en entrenar una función de mapeo que asegure el mínimo error de reconstrucción entre la entrada y la salida del algoritmo [43]. Como se puede observar en la Figura 4.6, la arquitectura de un *autoencoder* se compone principalmente de dos etapas: etapa de codificador (*encoder*) y etapa de decodificador (*decoder*). El codificador se encarga de transformar los datos de entrada \mathbf{X} , en este caso el valor de metilación de los sitios CpG de cada paciente, en una representación latente \mathbf{Z} mediante un mapeo no lineal, $\mathbf{Z} = f_{\phi}(\mathbf{X})$, donde ϕ son parámetros a optimizar en la etapa del codificador. Con el objetivo de evitar la llamada "maldición de la dimensionalidad", las capas que conforman la parte del codificador van reduciendo su dimensión hasta llegar a la dimensión del espacio latente \mathbf{Z} que es mucho menor que la de los datos de entrada [44]. Dado que el espacio latente es una combinación no lineal de los datos de entrada y tiene una menor dimensionalidad, este espacio representa las características más destacadas de los datos, dando menos importancia a aquellas que no son tan importantes. Tras la codificación de los datos, se lleva a cabo la decodificación de los mismos. En esta etapa, se produce la reconstrucción de los datos en función de las características incrustadas en el espacio latente, $\mathbf{R} = g_{\theta}(\mathbf{Z})$. Lo que se pretende es que la representación reconstruida \mathbf{R} sea lo más similar posible a \mathbf{X} . Por lo tanto, dado un conjunto de muestras de datos x_i , donde $i = 1 \dots n$ siendo n el número de muestras disponibles, el modelo de *autoencoder* se optimiza mediante el error cuadrático medio o *mean squared error* (MSE):

$$\min_{\theta, \phi} L_{rec} = \min \frac{1}{n} \sum_{i=1}^n \|x_i - g_{\theta}(f_{\phi}(x_i))\|^2 \quad (4.2)$$

donde θ y ϕ denotan los parámetros del codificador y del decodificador respectivamente.

De esta forma, el algoritmo de *autoencoder* habrá encontrado la solución óptima cuando el término L_{rec} sea el menor posible. Así, la entrada y la salida de este algoritmo serán lo más parecidas posible, lo que

indicará que las características incrustadas en el espacio latente representan con fiabilidad los datos de entrada mediante una dimensión menor, lo cual es el objetivo del presente proyecto.

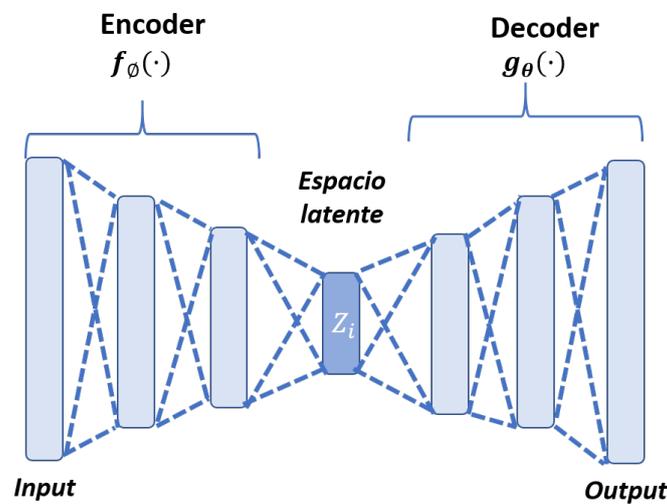


Figura 4.6: Arquitectura algoritmo *autoencoder* utilizado para la reducción de la dimensionalidad de forma no supervisada. Elaboración propia.

La arquitectura del *autoencoder* puede variar entre un perceptrón multicapa, una red de memoria a largo plazo (LSTM) o una red neuronal convolucional (CNN), según el caso de uso. En caso de que los datos de entrada tengan una dimensión 1-D y no estén relacionados en el tiempo, como ocurre en el caso de la metilación del ADN, ya que para cada paciente se tiene un vector 1-D con los valores de metilación de los distintos sitios CpG, tanto el codificador como el decodificador generalmente están contruidos por una red neuronal completamente conectada formada por capas *fully connected layer* o capas densas. De esta forma, cada nodo que compone las capas densas está conectado a cada uno de los nodos de la siguiente capa oculta. Por tanto, en la etapa de codificación cada capa densa tendrá un número de nodos menor hasta llegar a la dimensión que se quiera que tenga el espacio latente. En la etapa de decodificación se producirá un aumento de los nodos que componen cada capa, hasta llegar a la última capa, donde el número de nodos será igual a la dimensión de las características de entrada. Dado que cada nodo de cada capa densa es una combinación de los nodos de la capa anterior, el espacio latente es una representación combinada y reducida de las características de entrada.

■ *Variational autoencoder*

Al igual que el algoritmo basado en *autoencoder* convencional, un autoencoder variacional es una arquitectura compuesta por una parte que codifica los datos (*encoder*) y una parte encargada de su decodificación (*decoder*), con el objetivo de minimizar el error de reconstrucción entre los datos iniciales y los datos decodificados. Sin embargo, la principal diferencia entre un *autoencoder* convencional y uno variacional reside en que en este último se introduce una regulación en el espacio latente con el fin de mejorar sus propiedades.

En un autoencoder variacional en lugar de codificar los datos de entrada como puntos del espacio latente z_i , tal y como se llevaba a cabo en el *autoencoder* convencional, los datos se codifican como una distribución normal multivariante $p(z|x)$ ¹ alrededor de un punto en el espacio latente, de modo que el co-

¹En probabilidad y estadística, una distribución normal multivariante, también llamada distribución gaussiana multivariante $N(\mu, \sigma)$, es una generalización de la distribución normal unidimensional a dimensiones superiores.

difificador es optimizado para obtener la media y la matriz de covarianza que describen estos gaussianos, Figura 4.7 [45]. Los algoritmos basados en *autoencoder* variacionales suponen que no hay correlación entre ninguna de las dimensiones del espacio latente y, por lo tanto, que la matriz de covarianza es diagonal. Esto significa que el codificador solo necesita asignar cada entrada a un vector medio y un vector de varianza y no necesita preocuparse por la covarianza entre las dimensiones. Además, se asigna el logaritmo de la varianza, ya que este puede tomar cualquier número real en el rango $(-\infty, \infty)$, mientras que los valores de la varianza siempre son positivos.

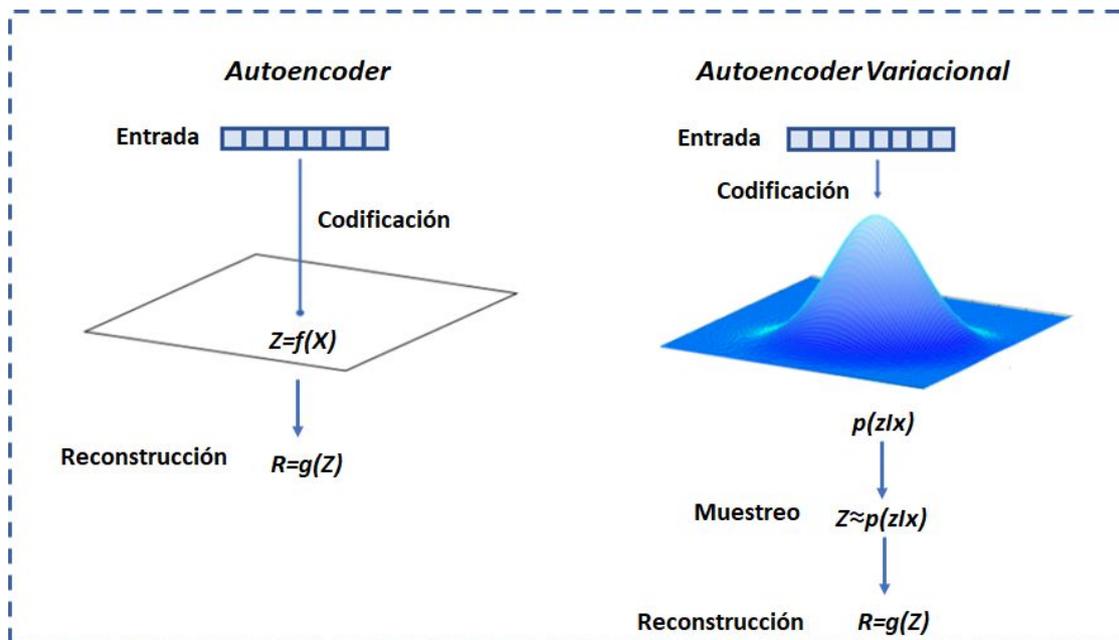


Figura 4.7: Principales diferencias entre un algoritmo basado en *autoencoder* y aquel basado en un *autoencoder* variacional.

Como se ha mencionado, en el caso del algoritmo basado en *autoencoder* variacionales, a primera instancia la salida del codificador es la media y el logaritmo de la varianza. Con el fin de aportar al espacio latente continuidad, los puntos que pertenecen a este espacio deberían concentrarse de tal forma que dos puntos cercanos no proporcionen dos contenidos completamente diferentes una vez decodificados, y completitud, un punto muestreado desde el espacio latente debería dar contenido significativo una vez decodificado, es necesario regularizar tanto el logaritmo de la varianza como la media de las distribuciones devueltas por el codificador. En la práctica, esta regularización se realiza haciendo cumplir que las distribuciones devueltas por el codificador estén cerca de una distribución normal estándar (centrada $\mu = 0$ y reducida $\sigma = 1$).

Tras la obtención y optimización de los parámetros de media y varianza de las distribuciones, es necesario obtener los puntos del espacio latente. Para ello, se toman muestras de la distribución obtenida, mediante la siguiente ecuación:

$$z \approx p(z|x) = \mu + \sigma \cdot \varepsilon \quad (4.3)$$

donde ε se muestrea de forma aleatoria a partir de una distribución normal estándar y $\sigma = \exp(\frac{\log(\sigma^2)}{2})$.

La función de pérdidas que se minimiza al entrenar un algoritmo basado en *autoencoder* variacionales (VAE) se compone de un "término de reconstrucción" (en la capa final), cuyo objetivo es que el esquema

de codificación-decodificación sea lo más eficaz posible, y un "término de regularización"(en el capa latente), que tiende a regularizar la organización del espacio latente al hacer que las distribuciones devueltas por el codificador se aproximen a una distribución normal estándar, Figura 4.8. Ese término de regularización se expresa como la divergencia *Kullback-Leibler* que mide la diferencia entre la distribución de probabilidad predicha y la distribución normal estándar en términos de las medias y varianza de las dos distribuciones [46]:

$$D_{KL}[N(\mu, \sigma) || N(0, 1)] = \frac{1}{2} \sum (1 + \log(\sigma^2) - \mu^2 - \sigma^2) \quad (4.4)$$

Cabe destacar que las operaciones se realizan sobre todas las dimensiones del espacio latente. La función de *Kullback-Leibler* se minimiza a 0 cuando $\mu = 0$ y $\log(\sigma^2) = 0$ para todas las dimensiones. A medida que estos dos términos comienzan a diferir de 0, esta pérdida aumenta. La compensación entre el error de reconstrucción y la divergencia KL es un parámetro a ajustar en este tipo de arquitecturas. Si la ponderación llevada a cabo sobre la pérdida de reconstrucción es demasiado elevada, la pérdida correspondiente al término de KL no tendrá el efecto regulador deseado [46]. De esta forma, las distribuciones no tenderán a "superponerse", y por tanto, no se satisficieran las condiciones de continuidad e integridad esperadas. Si el término de ponderación es demasiado pequeño, la pérdida de divergencia KL dominará y los datos reconstruidos serán pobres, por lo que no representarán la información contenida en el *input* de características [45].

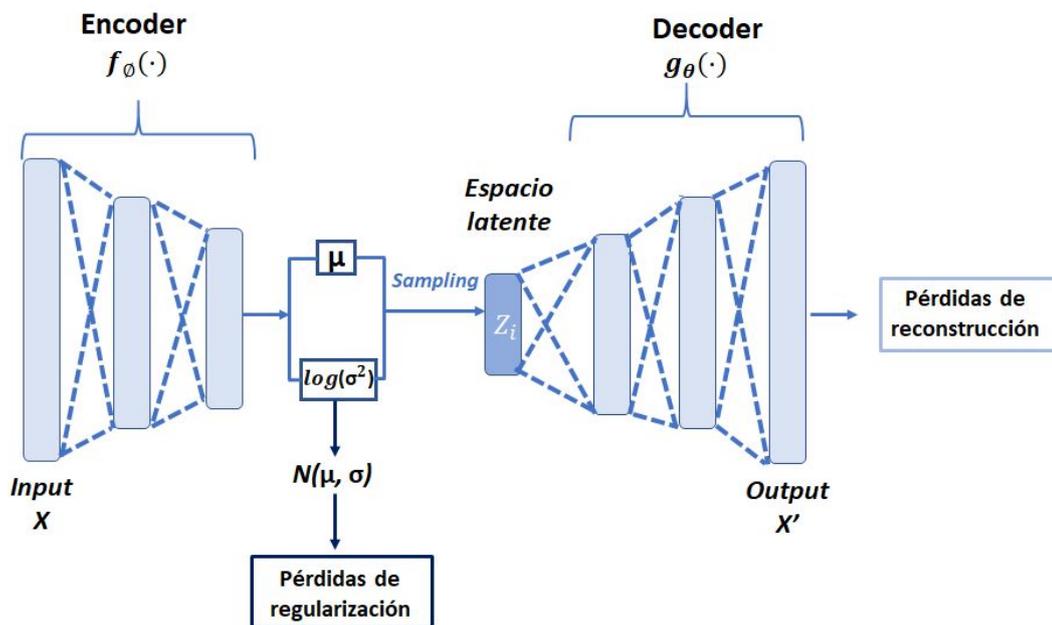


Figura 4.8: Arquitectura común del autoencoder variacional. En este caso, este algoritmo está optimizado por la minimización de dos funciones de pérdidas, una que corresponde a la regularización del espacio latente y otra correspondiente a la reconstrucción de los datos de entrada. Elaboración propia.

4.2.2 *Balanceo y creación de muestras sintéticas para la base de datos de imágenes histológicas*

Como se ha podido observar en la Tabla 4.1, el número de *patches* en los pacientes que presentan neoplasias benignas, en general, es inferior que en aquellos que presentan melanoma. Esto lleva consigo la consideración de la clase benigna como la clase minoritaria y por ello, un desbalanceo entre las diferentes clases con las que se entrenarán los modelos. El desbalanceo entre clases afecta, por lo general, a los algoritmos de clasificación en su proceso de generalización de la información, perjudicando a las clases minoritarias. Si se entrena un algoritmo supervisado, por ejemplo una red neuronal, con el 90% de muestras de la clase A y solo el 10% de la clase B, cuando este sea testeado con nuevas muestras, lo más probable es que se limite a asignar todas las muestras a la clase A, puesto que así tuvo más de un 90% de acierto durante su fase de entrenamiento. Dado que en el presente TFM se pretende desarrollar algoritmos de clasificación de imágenes supervisados que tienden a sobreajustarse a la clase mayoritaria cuando estas están desbalanceadas, se ha llevado a cabo una combinación de las técnicas de *undersampling* y *oversampling* para mitigar este problema. Cabe destacar que estas técnicas han sido empleadas únicamente a los datos con los que se entrenará el modelo.

■ *Undersampling*

El término *undersampling* o submuestreo se puede definir como la eliminación de algunas de las muestras de la clase mayoritaria de forma aleatoria, hasta que la clase minoritaria se convierta en un porcentaje específico de la clase mayoritaria. El submuestreo puede ser una buena opción cuando se tiene gran cantidad de datos. Sin embargo, el principal inconveniente de esta técnica se encuentra en que se está eliminando información que puede ser valiosa, lo que podría conducir a una falta de ajuste y una generalización deficiente del conjunto de prueba. Dado que para este estudio el número de muestras es limitado, se ha decidido combinar esta técnica con una aumento de la clase minoritaria u *oversampling*.

■ *Oversampling*

El término de *oversampling* o sobremuestreo se puede definir como la agregación de copias de la clase minoritaria. El sobremuestreo puede ser una buena opción cuando no se tienen muchos datos para entrenar los modelos. Aunque existen varios métodos para llevar a cabo un aumento en el número de muestras de la clase minoritaria, como la duplicación de muestras en el *set* de entrenamiento o la creación de nuevos ejemplos de entrenamiento mediante la realización de ciertas operaciones como rotaciones y traslaciones, uno de los métodos más utilizados y que se emplea en este proyecto es el método de SMOTE (*Synthetic Minority Over-sampling Technique*).

El algoritmo de SMOTE fue desarrollado por Chawla et. al [47] y consiste en la creación de ejemplos sintéticos realizando operaciones en el espacio latente de características en lugar de en el espacio de datos. En este caso, la clase minoritaria se sobremuestra tomando una muestra, considerando sus k vecinos más cercanos (en el espacio de características) e introduciendo ejemplos sintéticos a lo largo de los segmentos que unen la muestra en consideración a cualquiera/todos de los k vecinos, Figura 4.9.

Para crear un punto de datos sintético, se toma la diferencia entre el vector de características de la muestra elegida y entre uno de esos k vecinos, x_{zi} . Posteriormente, se multiplica esa diferencia por un valor β aleatorio entre 0 y 1, y se agrega al vector de características en consideración para generar una nueva muestra x_{new} :

$$x_{new} = x_i + \beta(x_{zi} - x_i) \quad (4.5)$$

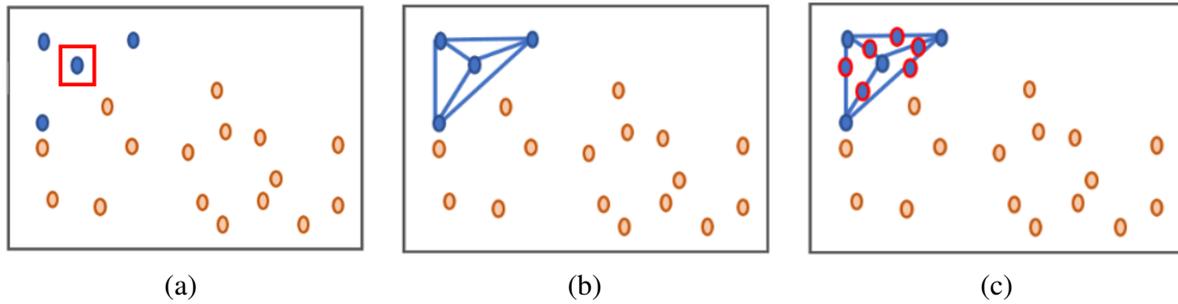


Figura 4.9: Utilización de la metodología SMOTE para la generación de muestras sintéticas. a) Se eligen K vecinos en torno al ejemplo dado. b) Se calcula la diferencia entre el vector de características del ejemplo bajo estudio y sus vecinos. c) Se generan nuevos ejemplos en el espacio de características delimitados por los segmentos que unen el ejemplo de estudio y cada uno de sus vecinos.

Este enfoque obliga a que la región de decisión de la clase minoritaria sea más general evitando así el sobreajuste de los modelos a los datos.

■ *Data Augmentation*

Una vez que se ha llevado a cabo el proceso de balanceo, se realiza un aumento de los datos de ambas clases mediante la técnica denominada *data augmentation*. Esta técnica consiste en la creación de muestras sintéticas mediante diversas transformaciones geométricas tales como rotaciones, traslaciones y volteos de manera que se aumenta significativamente la diversidad de datos disponibles para la capacitación de los modelos sin recopilar nuevos datos.

4.3 Partición de los datos

En los algoritmos de clasificación supervisada para que pueda llevarse a cabo el proceso de entrenamiento de los modelos, es fundamental determinar qué muestras del conjunto de datos se utilizarán para este proceso. A su vez, para garantizar la robustez de los modelos entrenados, es necesario realizar una partición de los datos para entrenar y validar los modelos de clasificación propuestos. Si dichos modelos se entrenan con los mismos datos que posteriormente se utilizan para validarlos, los resultados obtenidos serían inmejorables, pero muy poco fiables, ya que los modelos obtendrían peores resultados al predecir datos que nunca han sido vistos.

Debido al reducido número de pacientes (12 benignos y 8 malignos) con los que se cuenta en este estudio, se ha llevado a cabo la técnica conocida como *k-fold cross validation* para poder obtener la predicción de todos los pacientes bajo estudio. Cabe destacar que, tras el entrenamiento y validación/testeo de los modelos con los pacientes benignos y malignos, se han utilizado los pacientes diagnosticados como potencial maligno incierto (PMI) como pacientes de testeo para predecir sus pronóstico en malignos o benignos.

La metodología *k-fold cross validation* utilizada en este estudio queda ejemplificada en la Figura 4.10. En este caso, los dos conjuntos de datos con los que se cuenta en este estudio, base de datos de metilación y de imágenes histopatológicas, se han dividido en N sub-conjuntos que representan los distintos pacientes malignos y benignos con los que se cuenta, en total $N = 20$. La utilización de estos pacientes como datos de entrenamiento o de validación se ha ido alternado entre los distintos *K-fold*, en este caso $K = 5$ de la metodología implementada. El reparto de los pacientes entre los distintos *folds* ha sido el mismo tanto

para la base de datos de metilación como para la de base de datos de imágenes histopatológicas para la posterior fusión de ambas modalidades.

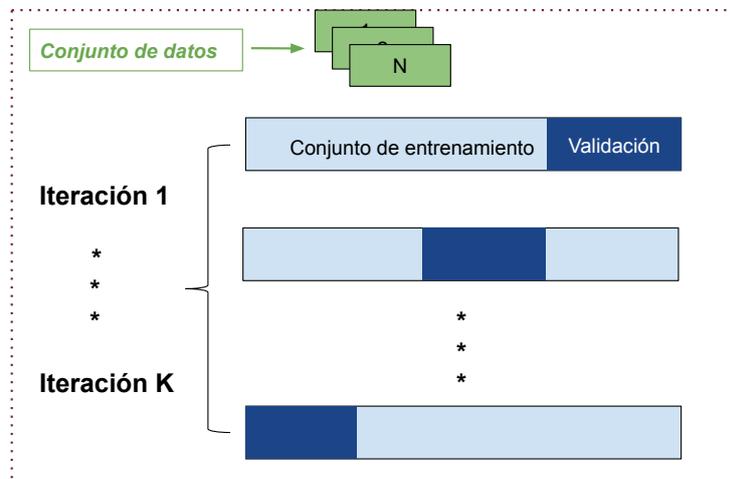


Figura 4.10: Metodología *K-fold cross validation* implementada en este estudio para llevar a cabo una correcta partición de los datos.

Cabe destacar que aunque se ha utilizado la técnica *K-fold* para llevar a cabo la partición de los datos, se han propuesto una serie de modificaciones que serán explicadas a continuación. Como se ha expuesto en la sección 4.1.2, para cada uno de los pacientes se utilizan varios *patches* para el entrenamiento de los modelos. Como se muestra en la Tabla 4.1, el número de *patches* entre los distintos pacientes es dispar puesto que las características y el tamaño de este tipo de tumores es muy irregular. Con el objetivo de intentar balancear lo máximo posible las distintas clases (benignos y malignos) en cada uno de los *K-fold* e intentar igualar el número de imágenes utilizadas dentro de cada partición, se ha llevado a cabo una partición semialeatoria en lugar de la partición totalmente aleatoria realizada por la técnica *K-fold cross validation*. Para ello, como el número de *K-fold* implementados en este modelo es 5, lo que corresponde a un 20% de los datos en cada conjunto de validación, se ha intentado que cada *set* contuviese un 10% de pacientes con lesiones melanocíticas malignas y un 10% con lesiones melanocíticas benignas, obteniendo la partición mostrada en la Tabla 4.2. Como se puede observar en esta tabla, a pesar de la implementación de esta partición modificada, todavía existe un desbalanceo entre las clases de los distintos set, por lo que será necesario el desarrollo de los algoritmos explicados en la sección 4.2.2, que permitan llevar a cabo un balanceo de los distintos *patches* entre las diferentes clases.

| K-fold | Set Entrenamiento | | | Set Validación | | |
|--------|-------------------|---------|---------|------------------|---------|---------|
| | Imágenes Totales | Clase 0 | Clase 1 | Imágenes Totales | Clase 0 | Clase 1 |
| Set 1 | 1214 | 536 | 678 | 262 | 73 | 175 |
| Set 2 | 1093 | 509 | 584 | 383 | 114 | 269 |
| Set 3 | 887 | 303 | 584 | 589 | 320 | 269 |
| Set 4 | 1301 | 547 | 754 | 175 | 77 | 99 |
| Set 5 | 1409 | 597 | 812 | 67 | 26 | 41 |

Tabla 4.2: Partición de datos generada mediante la técnica *K-fold* ajustada.

Como se ha comentado anteriormente, la misma partición por paciente que se ha obtenido para la base de datos de imágenes histopatológicas se ha utilizado en la base de datos de datos de metilación con el objetivo de la posterior fusión de los datos. Cabe destacar, que en lo relativo a los datos de metilación esta partición solo se ha utilizado para el desarrollo de los métodos supervisados.

4.4 Algoritmos de clasificación para neoplasias melanocíticas spitzoides

A continuación, se exponen los algoritmos de clasificación implementados para la distinción de las neoplasias melanocíticas benignas y malignas, tanto para los datos correspondientes a la metilación del ADN como para las imágenes histopatológicas.

4.4.1 Algoritmos de clasificación supervisada para los datos de metilación

Tras llevar a cabo la reducción de la dimensionalidad de la base de datos de metilación correspondiente a los pacientes con neoplasias melanocíticas malignas y benignas, se lleva a cabo la clasificación de estos datos. Para llevar a cabo esta tarea, se ha hecho uso del método comúnmente conocido como perceptrón multicapa, ya que ha demostrado ser un algoritmo que ofrece resultados óptimos.

El perceptrón multicapa (MLP) (del término *multi-layer perceptron* en inglés) es una red neuronal artificial (RNA) formada por múltiples capas que genera un conjunto de salidas a partir de un conjunto de entradas. Este algoritmo es una generalización del perceptrón simple y surgió como consecuencia de las limitaciones de dicha arquitectura en lo que al problema de la separabilidad no lineal se refiere. El elemento básico del perceptrón multicapa es lo que se conoce como “neurona”, que se compone de una función de agregación y de una función de activación no lineal, Figura 4.11.

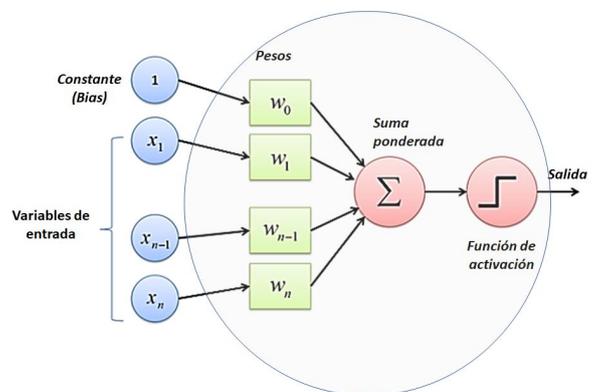


Figura 4.11: Arquitectura típica de las neuronas que componen un perceptrón multicapa.

En un perceptrón multicapa, las neuronas se disponen en una serie de capas, siendo la primera la capa de entrada, la última la capa de salida, y las intermedias, las capas ocultas, Figura 4.12. La capa de entrada está compuesta por tantas neuronas como variables se quieren analizar. La capa de salida, en cambio, está conformada por tantas neuronas como número de categorías en las que se quiera realizar la clasificación, excepto en casos especiales en los que se utilicen neuronas binarias que solo estará compuesta por una. Entre ambas capas, se encuentran las capas ocultas, encargadas de transformar la información a la entrada en niveles superiores de abstracción con el objetivo de poder dividir datos de manera precisa [48].

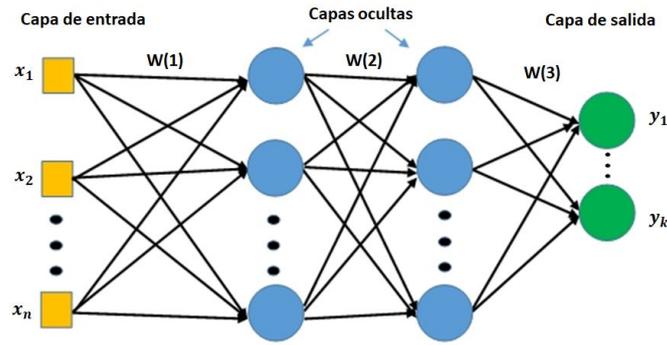


Figura 4.12: Arquitectura típica de una red neuronal compuesta por capas de entrada, ocultas y de salida para el objetivo de clasificación. Elaboración propia.

Para transmitir la información entre las diferentes capas del perceptrón, se aplica una función de activación a la suma de los productos de las activaciones procedentes de las capas anteriores por sus pesos correspondientes. Las funciones de activación más conocidas se muestran en la Figura 4.13.

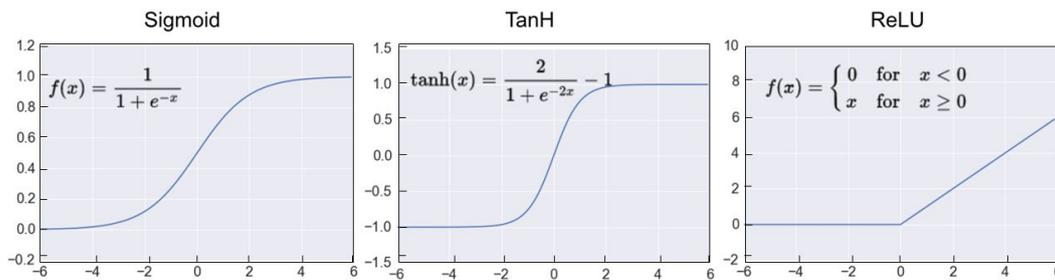


Figura 4.13: Funciones de activación más conocidas utilizadas en los clasificadores compuestos por redes neuronales. Elaboración propia.

A estas funciones de activación se les une la función de activación *softmax* que es comúnmente utilizada para las neuronas que se encuentran en la capa de salida.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{n=1}^{N_c} e^{z_n}} \quad (4.6)$$

Esta función de activación transforma las salidas a una representación en forma de probabilidades, de tal manera que el sumatorio de todas las probabilidades de las salidas es igual a la unidad, $\sum \sigma(z)_j = 1$. Por tanto, por ejemplo, en un problema de clasificación binaria que contenga esta función en la etapa de salida, esta capa estará formada por dos neuronas donde cada una de ellas establecerá la probabilidad de pertenencia a cada clase.

Los parámetros a optimizar para conseguir el objetivo de clasificación propuesto, son los pesos de la red neuronal. El proceso de aprendizaje de la red neuronal o también llamado fase de entrenamiento se compone de cuatro etapas: paso hacia delante o (*Foward pass*), cálculo del error, retropropagación o (*backward pass*) y actualización de los pesos. Durante el *Foward pass*, se transmite la información de las entradas de la primera capa al resto de neuronas del modelo que conforman el resto de capas. Una vez que la información llega a la capa de salida, dado que se trata de aprendizaje supervisado, la salida del

modelo es comparada con la etiqueta real de cada entrada y se calcula el error de predicción obtenido. En este momento da comienzo la fase de propagación hacia atrás o *backward pass*, donde se actualizan los pesos de las diferentes neuronas, en función de la tasa de aprendizaje o *learning rate*, tratando de encontrar el valor óptimo que minimice el error de propagación. En el caso que se lleve a cabo una clasificación binaria la función de pérdidas que se utiliza es la *binary cross-entropy*:

$$L_{BC} = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i) \quad (4.7)$$

donde N es el número de muestras a la salida del modelo, y_i es el valor de *ground truth* de la i -th muestra mientras que \hat{y}_i representa la predicción realizada por el algoritmo.

Cabe destacar que la clasificación se lleva a cabo en el espacio latente de características y no en el espacio de datos inicial. Por tanto, el input del perceptrón implementado es el espacio latente obtenido tras entrenar el modelo de *autoencoder* o *autoencoder* variacional, Figura 4.14. Como se ha comentado anteriormente, el perceptrón multicapa se trata de un algoritmo supervisado por lo que es necesario utilizar una parte de los datos para la fase de entrenamiento en la que los input del algoritmo sean los datos y sus correspondientes etiquetas para llevar a cabo la optimización de los pesos del modelo. Por tanto, aunque los algoritmos no supervisados de reducción de la dimensionalidad propuestos en este estudio transforman en características latentes toda la base de datos correspondientes a los pacientes con neoplasias benignas y malignas, sin necesidad de las etiquetas correspondientes, la partición de datos explicada en la sección 4.3 es utilizada para poder entrenar de manera supervisada los modelos y obtener la clasificación de cada uno de los pacientes, Figura 4.14.

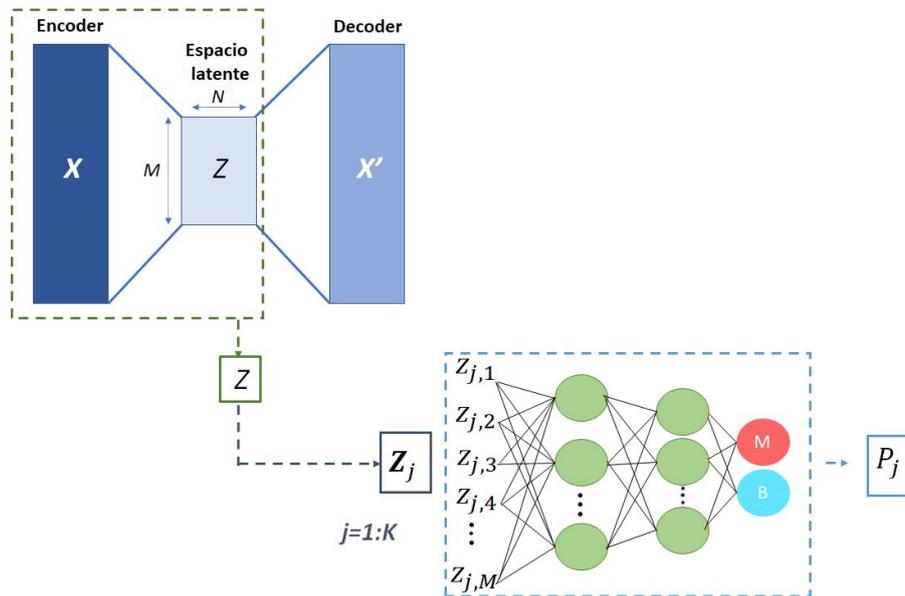


Figura 4.14: Flujo de trabajo para la clasificación de datos de metilación, donde K es el número de *folds*, en este caso 5. Elaboración propia.

4.4.2 Algoritmos de clasificación no supervisada para los datos de metilación

Como se ha explicado anteriormente, el uso de algoritmos supervisados para la clasificación de lesiones melanocíticas spitzoides requiere de una partición de datos posterior a la reducción de la dimensionalidad. De este modo, las tareas de reducción y clasificación de los datos no se pueden aunar para desarrollar algoritmos que optimicen al mismo tiempo la reducción de los datos junto con la tarea de clasificación. Como solución a este problema se presenta un algoritmo no supervisado basado en *deep clustering* que permite mejorar la representación de los datos en el espacio latente mientras que optimiza la clasificación de los mismos.

El concepto de *deep clustering* fue introducido por primera vez por [49] para la separación de fuentes acústicas. Gradualmente, esta familia de algoritmos ha ganado popularidad mejorando los resultados presentes en el estado del arte en varias tareas como clasificación de imágenes [44], [50], [51], desarrollo de algoritmos de segmentación [52], separación de voz [53] o secuenciación del RNA [54]. Los algoritmos basados en *deep clustering* son una familia de métodos de clustering que adoptan redes neuronales profundas para aprender representaciones de los datos que favorecen la agrupación. Dado que integran la reducción de la dimensionalidad con la tarea de clasificación, la función de pérdidas (objetivo de optimización) de estos métodos generalmente se compone de dos partes: pérdida de reconstrucción L_{rec} y pérdida de agrupación $L_{cluster}$. El término de la función de pérdidas correspondiente a la reconstrucción L_{rec} se utiliza para mejorar la representación de características, mientras que el término de pérdidas de *clustering* $L_{cluster}$ fomenta que los puntos de características formen grupos o se vuelvan más discriminatorios [43]. La pérdida de reconstrucción puede ser la pérdida de reconstrucción de un autoencoder, la pérdida de variación de un *autoencoder* variacional (VAE) o la pérdida de una *generative adversarial network* (GAN). La pérdida de agrupamiento puede ser pérdida de k-means, pérdida de agrupamiento aglomerativo, pérdida de preservación de la localidad, etc. En este caso, el método propuesto para la clasificación de tumores melanocíticos basados en los datos de metilación se muestra en la Figura 4.15. Tras una exhaustiva búsqueda en la literatura, se ha podido comprobar que no existen trabajos previos aplicados a la metilación del DNA basados en este tipo de algoritmos.

- *Arquitectura del modelo basado en deep clustering*

Como se puede observar en la Figura 4.15, el método está compuesto por un *autoencoder* para llevar a cabo la reducción de la dimensionalidad unido a una capa de clustering, que se utiliza para la clasificación no supervisada en el espacio latente. El proceso de entrenamiento de este algoritmo se compone de dos sub-etapas. La primera sub-etapa consiste en el pre-entrenamiento del *autoencoder*, sin la capa de *clustering*, con el objetivo de inicializar los centroides mediante las características extraídas del espacio latente. Para ello, se utiliza el algoritmo de *K-means* en el espacio latente obtenido una vez pre-entrenado minimizando pérdidas de reconstrucción. La segunda sub-etapa consiste en la capacitación del modelo propuesto (*autoencoder* + proceso de agrupamiento) minimizando los términos de las pérdidas correspondiente al proceso de agrupamiento y al de reconstrucción de los datos. La pérdida del *autoencoder* se minimiza para obtener la máxima similitud entre la salida del autoencoder y los datos de entrada, mientras que la pérdida de agrupamiento se minimiza para obtener el refinamiento de los centroides.

Dado que en la sección 4.2.1 se ha expuesto el modo de actuación del algoritmo basado en *autoencoder*, a continuación, se explican los principales fundamentos de la capa de *clustering* implementada.

La capa de *clustering* propuesta se aplica una vez que los datos \mathbf{X} se han transformado en características del espacio latente \mathbf{Z} . Por tanto, dada una estimación inicial del mapeo no lineal f_ϕ y los centroides

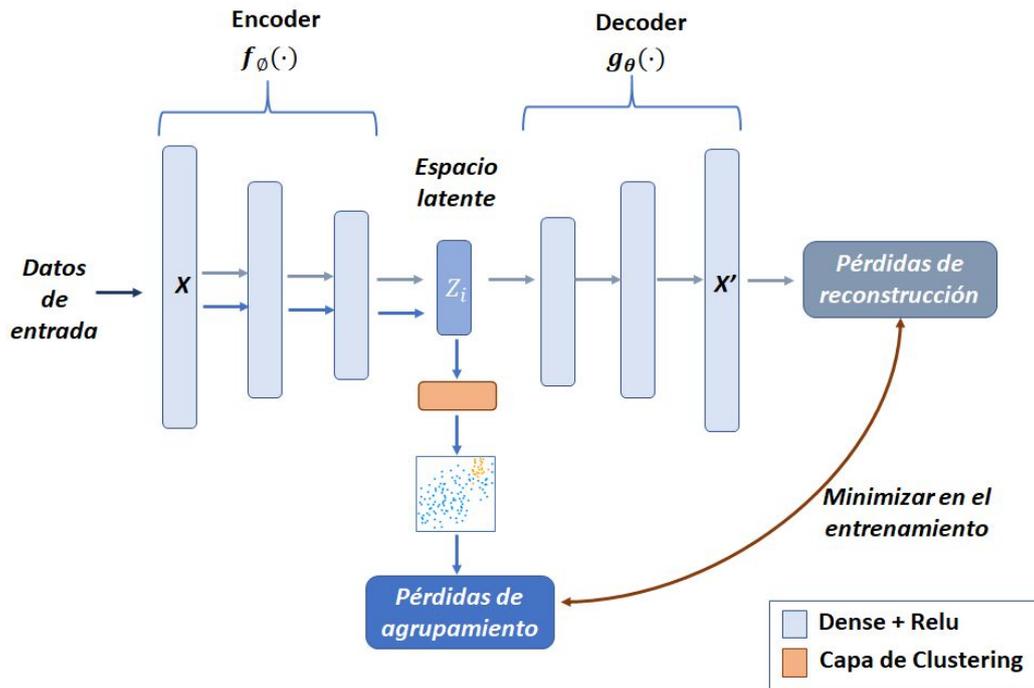


Figura 4.15: Arquitectura de deep learning propuesta para el problema de clasificación de lesiones melanocíticas. Elaboración propia.

iniciales de los *cluster* k $\{\mu_j\}_{j=1}^k$, el proceso de *clustering*, al igual que en [44], consiste en dos etapas principales: (1) asignación probabilística de los puntos refinados del espacio latente a los centroides de los *cluster* $\{\mu_j\}_{j=1}^k$ (2) refinamiento de los centroides de los *cluster* mediante el aprendizaje de las asignaciones de alta confianza utilizando una distribución objetivo auxiliar.

- *Asignación de los puntos refinados del espacio latente a los centroides de los clusters*

Siguiendo el método propuesto en [44], el algoritmo utiliza la distribución *t-Student's* para medir la similitud entre los puntos del espacio latente z_i y los centroides μ_j :

$$q_{i,j} = \frac{(1 + \|z_i + \mu_j\|^2)^{-1}}{\sum_{j'} (1 + \|z_i + \mu_{j'}\|^2)} \quad (4.8)$$

donde $q_{i,j}$ es la probabilidad de asignar la muestra i del espacio latente al *cluster* j .

- *Refinamiento de los centroides de los cluster*

Con el objetivo de refinar iterativamente los centroides de los *cluster*, se lleva a cabo un proceso de aprendizaje de las asignaciones de alta confianza llevadas a cabo por una distribución objetivo auxiliar $p_{i,j}$, que derivan de la asignación probabilística realizada por el algoritmo, Ecuación (4.8). La distribución objetivo se calcula elevando q_i a la segunda potencia y normalizándola por la frecuencia de cada *cluster* [44]. Por lo tanto, la distribución objetivo se puede expresar con la siguiente fórmula:

$$p_{i,j} = \frac{q_{i,j}^2 / f_j}{\sum_{j'} q_{i,j'}^2 / f_{j'}} \quad (4.9)$$

donde $f_j = \sum_i q_{i,j}$ son las frecuencias de cada *cluster*.

Las asignaciones de cada punto del espacio latente a cada *cluster* se optimizan haciendo que la asignación probabilística llevada a cabo por la distribución *t-Student's* coincida con la asignación de la distribución objetivo. Para lograr este propósito, se utiliza la función de pérdidas *Kullback–Leibler (KL divergence)* entre las asignaciones probabilísticas q_i y la distribución auxiliar $p_{i,j}$.

$$L_{cluster} = KL(P||Q) = \sum_i \sum_j p_{i,j} \log \frac{p_{i,j}}{q_{i,j}} \quad (4.10)$$

Durante el proceso de entrenamiento del algoritmo propuesto, a medida que se refina la asignación de los puntos del espacio latente a cada uno de los *cluster*, también se optimizan las representaciones del espacio latente dadas por el *autoencoder* mediante la función de pérdidas *mean square error (MSE)*, L_{rec} . Por tanto, la función de pérdidas de este método esta formada por dos términos, pérdida de *autoencoder* L_{rec} y pérdida de agrupación $L_{cluster}$:

$$L = L_{cluster} + \beta L_{rec} \quad (4.11)$$

donde β pondera la importancia de las pérdidas debido a la reconstrucción de los datos.

Durante la optimización, dado un *batch-size*² m y una tasa de aprendizaje lr , los centroides de los *cluster* son actualizados de acuerdo a la siguiente fórmula:

$$\mu_j = \mu_j - \frac{lr}{m} \sum_{i=1}^m \frac{\partial L_{cluster}}{\partial \mu_j} \quad (4.12)$$

donde $\frac{\partial L_{cluster}}{\partial \mu_j}$ representa el gradiente de la función de pérdidas correspondiente al término de *clustering* frente a los centroides μ_j .

Los pesos del *decoder* W' son actualizados como:

$$W' = W' - \frac{lr}{m} \sum_{i=1}^m \frac{\partial L_{rec}}{\partial W'} \quad (4.13)$$

Mientras que los pesos del *encoder* W son actualizados como:

$$W = W - \frac{lr}{m} \sum_{i=1}^m \left[\frac{\partial L_{rec}}{\partial W} + \beta \frac{\partial L_{cluster}}{\partial W} \right] \quad (4.14)$$

4.4.3 Algoritmos de clasificación supervisada para imágenes histopatológicas

De entre todos los métodos existentes para el análisis de imágenes, los algoritmos basados en redes neuronales convolucionales (CNN) son el método de aprendizaje profundo más utilizado para la clasificación de imágenes histopatológicas [55]. Estos algoritmos se han aplicado con éxito para la detección de varios tipos de cáncer como de cáncer de mama [56]-[58], cáncer de próstata [59], [60], cáncer de pulmón [61]-[63] y melanoma común [55]. Sin embargo, hasta el momento no existe ningún algoritmo que haya sido aplicado a imágenes histológicas para la clasificación del melanoma spitzoide.

²El *batch-size* define el número de muestras que se propagarán a través de la red.

- *Redes neuronales convolucionales y transfer learning*

Una red neuronal convolucional es un tipo de red neuronal artificial (RNA) que emplea capas convolucionales para extraer automáticamente a partir de los datos de entrada, imágenes en este caso, características discriminatorias capaces de resolver ciertas tareas de visión por computador (clasificación, detección, segmentación, etc.). Como se puede observar en la Figura 4.16, un clasificador CNN consta de dos partes fundamentales: (i) *base model* o extractor de características y (ii) *top model* o clasificador [64].

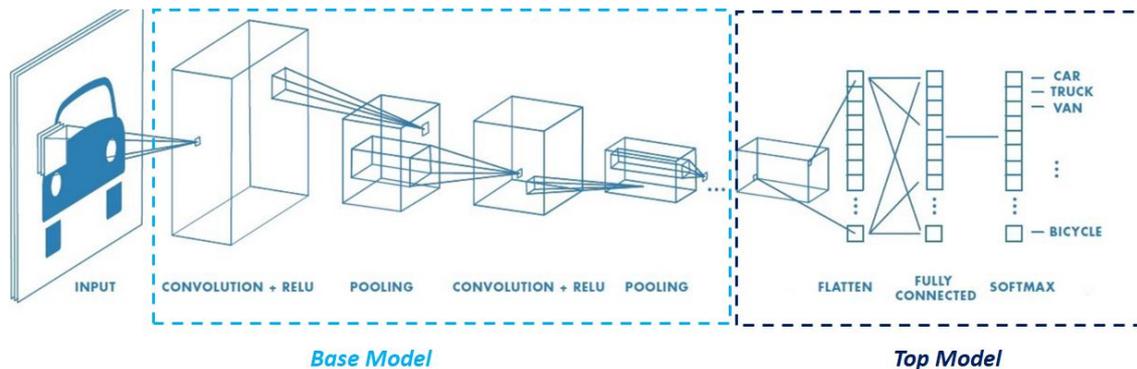


Figura 4.16: Arquitectura básica de una red neuronal convolucional.

El extractor de características o *base model* suele estar compuesto por una serie de capas convolucionales, unidas a funciones de activación, encargadas de extraer características de la imagen a distintos niveles de abstracción, y capas de *pooling* que reducen la dimensión de los mapas de características extraídos por las capas convolucionales. El clasificador o *top model* es el encargado de llevar a cabo la clasificación de las imágenes a partir de las características extraídas por el *base model*. Esta parte de la arquitectura suele estar compuesta por una capa *flatten*, que transforma los mapas de características extraídos por el *base model* en un vector 1-D, y una serie de capas completamente conectadas complementadas con una función de activación, comúnmente *softmax*, para establecer la probabilidad de pertenencia a cada clase. Del mismo modo que una red neuronal artificial, la CNN es un algoritmo supervisado que lleva a cabo el proceso de aprendizaje de los modelos en cuatro etapas: paso hacia delante, cálculo de la función de pérdidas, paso hacia atrás y actualización de los pesos.

El entrenamiento de una red neuronal convolucional profunda desde cero es difícil ya que requiere una gran cantidad de datos de entrenamiento etiquetados, un requisito que puede ser complicado de cumplir en el ámbito médico ya que la anotación por parte de los expertos es costosa y conlleva una gran cantidad de tiempo [65]. Para intentar suplir este problema surge la técnica conocida como *transfer learning*. Esta técnica consiste en la utilización de arquitecturas que previamente han sido entrenadas con grandes bases de datos etiquetados. De esta forma, se puede hacer uso de los valores de pesos de los filtros ya ajustados y optimizados para la base de datos original con la que fue entrenada la arquitectura. Así, se puede utilizar de forma más efectiva redes con un número elevado de capas convolucionales, lo cual es necesario cuando se quiere resolver problemas ambiguos, aunque el número de datos no sea tan grande como el esperado. La idea de las técnicas basadas en *transfer learning* es dejar los pesos del *base model* 'congelados' y añadir un *top model ad-hoc* para la aplicación propuesta, donde la última capa tendrá tantas neuronas como clases en las que se quiere clasificar. Sin embargo, con esta técnica, la precisión que se tiene en problemas complejos es relativamente baja ya que el *base model* no es capaz de adaptarse a nuevos datos de entrada. Como solución a este problema, aparte de añadir un nuevo

top-model se re-entrenan algunos bloques convolucionales del *base-model*. Esta técnica, que se engloba dentro de *transfer learning*, se conoce como *fine-tuning* y permite un ajuste más fino en la clasificación de los modelos sin que se produzca *overfitting*. Comúnmente, para llevar a cabo la técnica de *fine-tuning* se utilizan las arquitecturas convolucionales resultantes del concurso anual ImageNet, públicamente disponibles. Entre ellas las más destacadas son: AlexNet [66], VGG16 y VGG19 [67], Inception [68] y ResNet [69].

Dado que el número de pacientes con los que se cuenta en este estudio es reducido, el método implementado para el análisis de lesiones melanocíticas spitzoides malignas y benignas está basado en técnicas de transferencia del conocimiento (*transfer learning*). En concreto, se ha utilizado la técnica de *fine-tuning* sobre la arquitectura VGG, mostrada en la Figura 4.17, ya que ha que ha mostrado los mejores resultados en otros estudios en los que se ha llevado a cabo tareas de clasificación con imágenes histopatológicas [59]. Esta arquitectura está compuesta por un total de trece capas convolucionales con filtros 3x3 distribuidos en 5 bloques para dotar de profundidad a la red. Estas capas convolucionales se encuentran seguidas de la función ReLU para la activación de los pesos. Además, entre cada bloque convolucional, se utiliza una capa de *maxpooling* en ventanas 2x2 con paso 2.

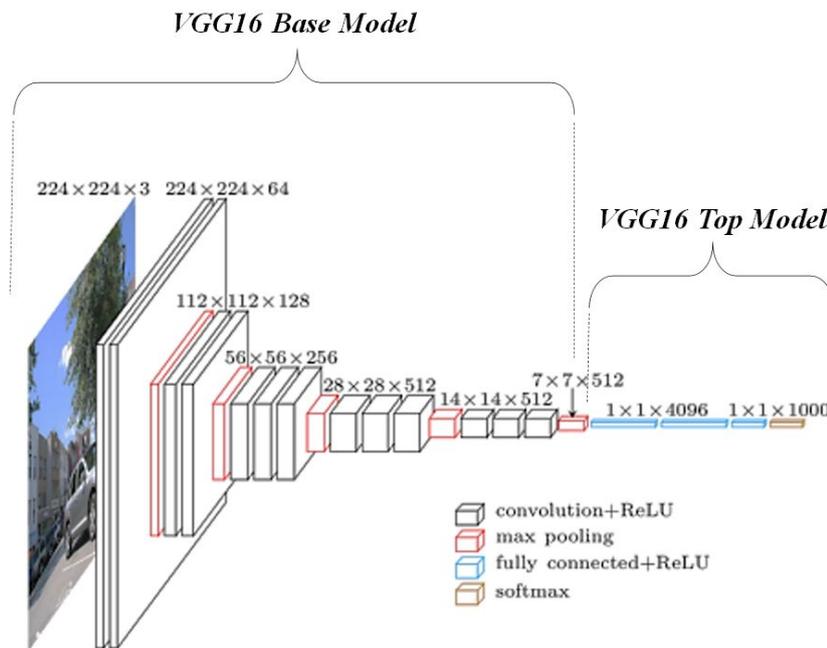


Figura 4.17: Arquitectura VGG diseñada para llevar a cabo tareas de clasificación. Figura extraída de [67].

- *Módulo de atención*

Las imágenes médicas siempre contienen mucha información irrelevante que puede perturbar la toma de decisiones. Es por ello que, cuando se quiere resolver problemas de clasificación ambiguos, es necesario refinar las características extraídas del *base model*. De esta forma, las características finales serán más discriminatorias y se podrá mejorar la precisión en la clasificación. Es por ello que, en este proyecto, se ha propuesto un módulo de atención que imita el comportamiento clínico al centrarse en las características clave para la predicción [70].

En este caso, el *attention module* o módulo de atención tiene como entrada la última capa del *base model* implementado $M \in \mathbb{R}^{H \times W \times C}$ donde H y W son la altura y la anchura espacial del mapa de característi-

cas y C el número de canales en la dimensión Z . Los principales componentes de este módulo son un conjunto de capas convolucionales con un tamaño de *kernel* de (1×1) y con un número de filtros decrecientes hasta llegar a 1. De esta forma, se intenta realizar un reducción de la dimensionalidad de las características, de la misma forma que se hace con los algoritmos basados en *autoencoder* y *variational autoencoder*, combinando las características en la dirección espacial Z , Figura 4.18. Por tanto, los mapas de características obtenidos a la salida de cada una de estas capas convolucionales tendrán la misma altura, H , y anchura, W , que el mapa de características anterior con la diferencia de que el número de canales se habrá reducido. La última capa de esta reducción, que como se ha comentado presenta un único filtro, está activada mediante la función sigmoide. De esta forma, esta función le aporta un peso a cada uno de los píxeles del mapa de características. La salida de esta capa se encuentra unida a una capa convolucional que presenta el mismo número de filtros que la capa de entrada a este módulo, siendo la salida de este módulo el mapa de características $A \in H \times W \times C$. La salida del módulo de atención se pondera con la salida del *base model*. Por tanto, el objetivo de este conjunto de capas convolucionales consiste en un refinamiento de las características en la dimensión espacial [70].

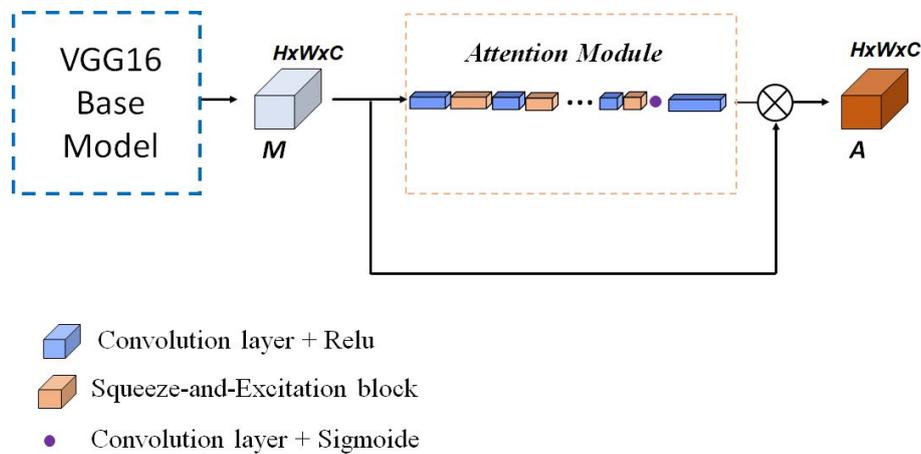


Figura 4.18: Módulo de atención (*attention module*) unido a la salida del *base model*.

Con el fin de explorar las interdependencias que existen entre los diferentes canales de características así como la información contextual se implementan los bloques denominados *Squeeze-and-Excitation* entre las diferentes capas convolucionales del módulo de atención (véase Figura 4.19) [71]. La entrada a este módulo $G \in \mathbb{R}^{H \times W \times R}$ es embebida en un vector $z \in \mathbb{R}^{1 \times 1 \times R}$ mediante una capa de *global Max pooling*, que ofrece una distribución global de respuestas por canales. Cabe destacar que el número de filtros R , corresponde al número de filtros que se tiene tras aplicar las diferentes capas convolucionales del *attention module*. En el paso siguiente, z es transformado a $\hat{z} = W_1(\partial(W_2 z))$ donde $W_1 \in \mathbb{R}^{R \times \frac{R}{r}}$ y $W_2 \in \mathbb{R}^{\frac{R}{r} \times R}$ son los pesos de dos capas completamente conectadas y ∂ la función Relu. El parámetro r es el ratio de reducción para la reducción de la dimensionalidad, indicando el cuello de botella. Después de pasar a través de una capa con función de activación sigmoide $\sigma(\hat{z})$, las activaciones de \hat{z} están limitadas entre $[0,1]$ y son utilizadas para recalibrar el input $\mathbf{G} = [g_1, g_2, \dots, g_c]$ donde $g_i \in \mathbb{R}^{H \times W}$. El mapa de características de salida de este bloque es $\mathbf{G}_{se} = [\sigma(\hat{z}_1)g_1, \sigma(\hat{z}_2)g_2, \dots, \sigma(\hat{z}_c)g_c]$.

Para crear los mapas de características refinados que serán la entrada del *top model*, se lleva a cabo la multiplicación de los mapas de características extraídos del *base model* M con aquellos que se obtienen tras el módulo de atención para dar lugar al mapa de características refinado A . Recordemos que esta

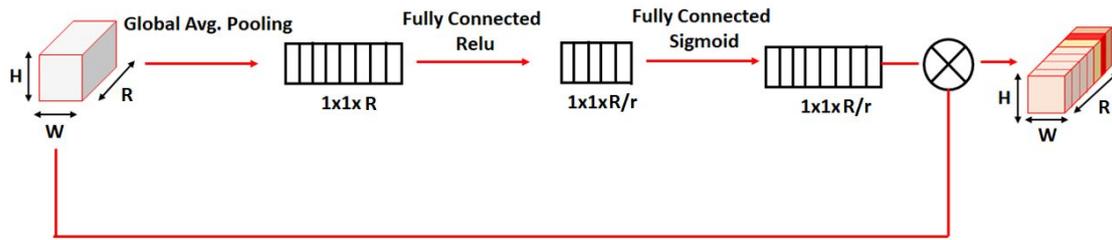


Figura 4.19: Arquitectura de los bloques Squeeze-and-Excitation.

multiplicación es posible ya que tanto la salida del *base model* como la salida del *attention net* o módulo de atención tienen unas dimensiones de $H \times W \times C$.

- *Predicción de las WSIs: majority voting*

Como se ha comentado en la sección 4.1.2, se han generado varios *patches* para cada imagen WSI de los diferentes pacientes, con el objetivo de entrenar los algoritmos implementados. Por tanto, el entrenamiento y la predicción de los modelos basados en *deep learning* aplicados a imágenes histopatológicas se ha llevado a cabo a nivel de *patch*. Para poder obtener la predicción de estos algoritmos a nivel de paciente, es decir de toda la imagen WSI, se ha utilizado la técnica conocida como *majority voting*. Con esta técnica, se realiza una clasificación de la imagen completa I en benigna o maligna en función del número de *patches* de cada clase que hayan sido predichos por el algoritmo.

Sea Pa_i , $i = \{1, 2, \dots, n\}$ un *patch* extraído de la imagen I de un paciente, donde n es el número total de *patches*, y λ la clase del *patch* Pa_i predicha por el algoritmo tal que $\lambda \in [0, 1]$, la función de conteo $F_i^{(k)}$ para cada una de las clases puede ser definida como:

$$F_i^{(k)} = \begin{cases} 1 & \text{si } \lambda = k \\ 0 & \text{si } \lambda \neq k \end{cases} \quad (4.15)$$

donde $k \in [0, 1]$ corresponde a las clases evaluadas, *ground truth*. El número de votos para cada clase C^k puede ser expresado como:

$$C^k = \sum_{i=1}^n F_i^{(k)} \quad (4.16)$$

Por tanto, la clasificación para cada imagen completa I y por tanto, para cada paciente, se puede expresar mediante la siguiente fórmula:

$$T = \operatorname{argmax}(C^k) \quad (4.17)$$

En la Figura 4.20 se muestra una imagen representativa del modo de actuación de este algoritmo. En esta imagen se muestran las diferentes etiquetas de *patch* para una imagen de entrada predicha por el algoritmo. Como se puede observar, el número de *patches* con la etiqueta de la clase 0 (benigna) en predicción es superior al número de parches estimados como clase 1 (maligno), por lo que mediante este algoritmo se predice la etiqueta de la imagen completa como 0.

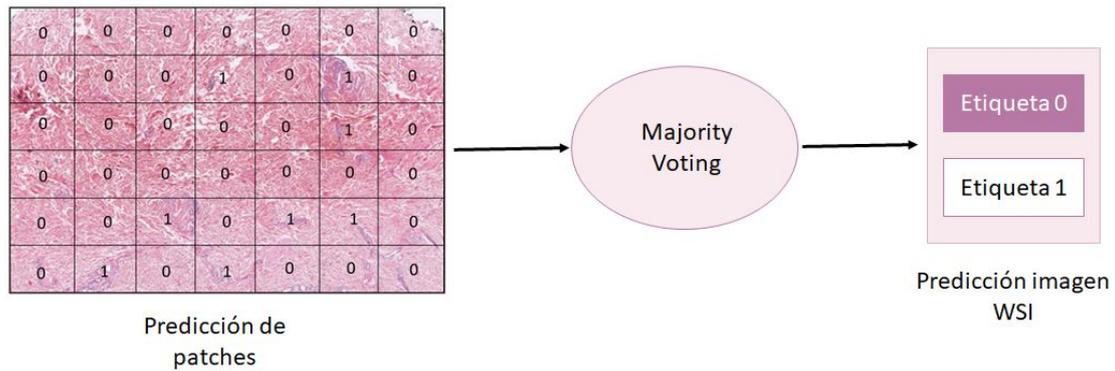


Figura 4.20: Técnica de *majority voting* para obtener la predicción a nivel de WSI.

4.4.4 Algoritmos de clasificación combinando modalidades

La integración de imágenes histopatológicas e información epigenética es útil para la extracción de valiosos biomarcadores asociados a diversos trastornos [72]. Las mejoras introducidas en el campo de la inteligencia artificial han permitido el desarrollo de algoritmos que facilitan la combinación de diferente información clínica. Los métodos de aprendizaje que combinan información de múltiples modalidades son, en principio, capaces de realizar una inferencia más sólida.

Los algoritmos basados en *deep learning* que combinan diferentes datos de entrada se denominan Redes Neuronales Híbridas (HNN). La combinación de modalidades pueden seguir los enfoques de fusión temprana (*early-fusion*) o tardía (*late-fusion*). En el enfoque de fusión temprana, los modelos de clasificación se construyen con los datos multimodales fusionados en una etapa temprana del modelo. Para que esto se realice, se desarrollan modelos individuales para cada tipo de datos, pero solo para las primeras capas, donde las diferentes entradas se separarán en grupos. Después de esto, los vectores de características de diferentes modalidades se combinarán por concatenación. Estos vectores de características concatenados serán la entrada de las redes del clasificador. En el enfoque de fusión tardía, se desarrollan diferentes modelos de clasificación para cada tipo de datos de entrada por separado y, en una etapa posterior, se combinan las características.

En el caso que nos ocupa, fusión de datos para la clasificación de lesiones melanocíticas spitzoides, las redes neuronales híbridas combinarán la imagen y los datos correspondientes al estudio de metilación para desarrollar un modelo entrenado con diferentes conjuntos de datos. Los datos de la imagen pueden ser procesados por una rama CNN, mientras que en paralelo las variables numéricas (datos de metilación) alimentarán una rama perceptrónica de múltiples capas, Figura 4.21.

4.5 Algoritmos para el pronóstico de neoplasias melanocíticas spitzoides

Como se ha comentado en la sección 3.1, la base de datos con la que se trabaja en este estudio, está conformada por pacientes con lesiones melanocíticas malignas, benignas y de potencial maligno incierto (PMI). Tal y como se ha explicado en la sección 2.1.2, la clasificación como tumores de potencial maligno incierto se hace para aquellos casos en el que no se puede establecer un diagnóstico claro como benignos o malignos pues presentan características ambiguas de ambos grupos. La predicción del pronóstico en este grupo de tumores es un desafío diagnóstico para los patólogos de forma regular. Por lo tanto, el desarrollo de algoritmos que permitan establecer alguna directriz en torno a la evolución de

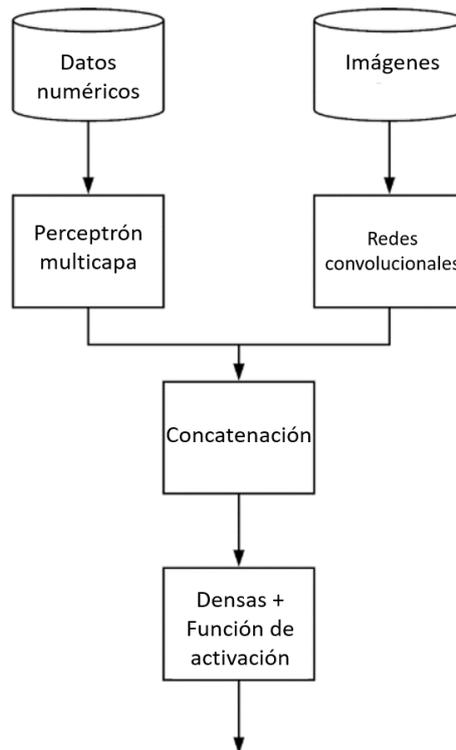


Figura 4.21: Arquitectura de una red híbrida para la fusión de datos.

estos pacientes es una tarea esencial, ya que en el caso de que el pronóstico sea maligno, se podrá llevar a cabo un seguimiento más exhaustivo actuando de forma temprana en el caso de surjan complicaciones. Es por ello que, en este apartado se lleva a cabo la implementación de algoritmos que permitan realizar el pronóstico de las lesiones melanocíticas de potencial maligno incierto en lesiones malignas o benignas.

4.5.1 Algoritmos para el pronóstico basados en datos de metilación.

El primer paso para llevar a cabo el pronóstico de estos tumores basado en datos de metilación consiste en la eliminación de las características consideradas no relevantes en el estudio estadístico realizado sobre las lesiones melanocíticas spitzoides malignas y benignas. Como se ha explicado en la sección 4.1.1, tras el análisis de los sitios CpG independientes entre sí y aquellos que tienen cierta dependencia con respecto a la clase, se obtienen un total de 62 sitios CpG discriminantes por paciente. Por lo tanto, de esta forma se obtienen 62 características discriminatorias en los pacientes con potencial maligno incierto y con ellas se procede a realizar el estudio basado en el pronóstico de estas lesiones. Remarcar que en este caso no se vuelve a realizar un análisis estadístico sino que únicamente se escogen las características relevantes tras la realización del análisis estadístico con lesiones benignas y malignas.

El pronóstico llevado a cabo sobre las lesiones melanocíticas de potencial maligno incierto está basado en una aproximación extraída de los algoritmos basados en redes prototípicas, *prototypical networks* (PN) [73]. Estas redes se basan en la idea de que existe un espacio latente en el que los puntos se agrupan alrededor de una única representación prototipo para cada clase. Para hacer esto, se aprende un mapeo no lineal de la entrada en un espacio latente usando una red neuronal y se toma el prototipo de

una clase como la media de su conjunto de soporte en el espacio de inclusión. La clasificación para un nuevo punto de consulta incrustado se realiza simplemente buscando el prototipo de clase más cercano.

En este caso, el mapeo no lineal entre los datos y el espacio latente se consigue con los algoritmos propuestos en la sección 4.2.1 con el que se consigue transformar los datos en características latentes de menor dimensión. Cabe destacar que al ser algoritmos no supervisados, los algoritmos de reducción de la dimensionalidad empleados en los datos de metilación han sido entrenados con todos los datos de la base de datos etiquetados como lesiones melanocíticas benignas y malignas. Tras trasladar los datos de los pacientes con melanoma y nevus melanocítico al espacio embebido (parte de *encoder*), se calcula el prototipado para cada clase haciendo uso de la siguiente fórmula:

$$\mathbf{p}_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\phi(x_i) \quad (4.18)$$

donde S_k denota el conjunto de ejemplos etiquetados con la clase k y f_ϕ la función que transforma los datos de entrada x_i a características del espacio latente.

A continuación, se calculan las características del espacio latente de los pacientes con potencial maligno incierto utilizando los modelos entrenados con los datos de lesiones melanocíticas malignas y benignas. Posteriormente, se calcula la distancia de las características de cada paciente diagnosticado como potencial maligno incierto en el espacio embebido al prototipado incrustado de cada clase. En este caso y acorde con los mejores resultados reportados en la literatura [73], se hecho uso de la distancia euclídea:

$$d(\mathbf{p}_k, f_\phi(\mathbf{q})) = \sqrt{(p_{k,i} - f_\phi(q_i))^2} \quad (4.19)$$

donde $i = 1 \dots N$, siendo N la dimensión del espacio latente y \mathbf{q} el vector de características inicial de los pacientes con potencial maligno incierto. Cabe destacar que la dimensión tanto de los centroides como de las características embebidas en el espacio latente de los pacientes con potencial maligno incierto es la misma. Por tanto, las distancias se computan en cada una de las dimensiones del espacio latente.

Posteriormente, se calcula la distribución sobre las clases (melanoma y benigno) para cada punto de consulta q_i mediante la función *softmax* de distancias a los prototipos en el espacio de incrustación:

$$prob(y = k | \mathbf{q}) = \frac{\exp(-d(\mathbf{p}_k, f_\phi(\mathbf{q})))}{\sum_{k'} \exp(-d(\mathbf{p}_{k'}, f_\phi(\mathbf{q})))} \quad (4.20)$$

En la Figura 4.22, se puede visualizar como actúa el algoritmo expuesto en esta sección. En esta imagen se representan los centroides de cada clase (benigna y maligna) incrustados en el espacio latente. Cuando llega un nuevo punto de consulta (paciente con potencial maligno incierto), este se lleva al espacio latente, y se observa a que prototipo se encuentra más próximo.

Por tanto, mediante este algoritmo se puede concluir para cada paciente diagnosticado como potencial maligno la probabilidad de que tenga un pronóstico como melanoma o como nevus melanocítico basado en la similitud de las características embebidas en el espacio latente. De igual forma, este método permite llevar a cabo la predicción de nuevas muestras que quieran ser diagnosticadas.

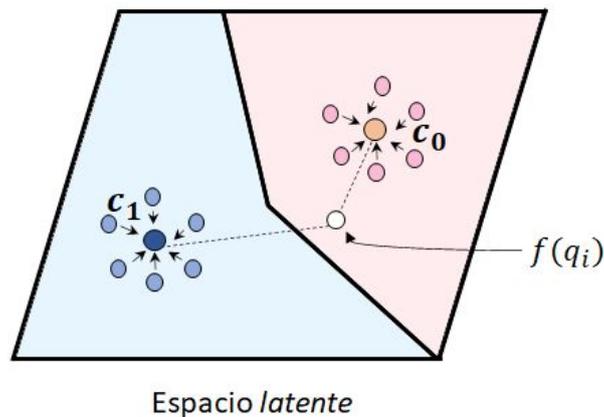


Figura 4.22: Prototipos c_k computados como la media de los ejemplos embebidos en el espacio latente para cada clase. Los puntos de consulta embebidos se clasifican a través de una *softmax* de distancias a los prototipos de clase. Elaboración propia.

4.5.2 Algoritmos para el pronóstico basados en imágenes histopatológicas

Como se ha explicado a lo largo del documento, para llevar a cabo el entrenamiento de redes neuronales convolucionales sobre las imágenes histopatológicas spitzoides es necesario llevar a cabo parcheado de la imagen. Por tanto, el primer paso para realizar el pronóstico de las imágenes con potencial maligno incierto consiste en parchear las imágenes WSI y seleccionar aquellos patches que serán utilizados para el pronóstico, tal y como se ha llevado a cabo en la sección 4.1.2 para las lesiones melanocíticas malignas y benignas.

Como se ha mencionado en secciones anteriores, los algoritmos basados en CNNs, utilizados para clasificar las imágenes histopatológicas, son algoritmos supervisados, por lo que como se ha explicado en la sección 4.3, requieren de una partición de datos para garantizar la robustez de los modelos desarrollados. Por tanto, si se tiene una partición basada en *K-fold* en este caso $K = 5$, se tienen 5 modelos entrenados, uno para cada set de validación. Con el objetivo de tener un único modelo que contenga toda la información de los pacientes malignos y benignos que componen la base de datos, se lleva a cabo un entrenamiento de la arquitectura de clasificación propuesta con todos los pacientes benignos y malignos. Este modelo es el que será utilizado para predecir el pronóstico de los pacientes con potencial maligno incierto.

Una vez que se tienen los *patches* que se quieren diagnosticar así como el modelo que se utilizará para realizar la predicción, se procede a la realización del pronóstico propiamente dicho. Para llevar a cabo el pronóstico de los pacientes con potencial maligno incierto se utilizará la misma idea que la explicada en el pronóstico basado en datos de metilación, redes prototípicas. En este caso, el mapeo no lineal entre los datos de entrada, imágenes, se lleva a cabo mediante la extracción de características del *base model*. Por tanto, las características embebidas en el espacio latente de los pacientes con lesiones melanocíticas malignas y benignas serán en este caso los mapas de características extraídos del *base model* y convertidos en un vector 1-D. Tras obtener las características de los pacientes con melanoma y benignos en el espacio latente, se utiliza la Ecuación (4.18) para computar los prototipos de cada clase. Posteriormente, se extraen las características de los pacientes con potencial maligno incierto mediante los modelos anteriormente entrenados. Una vez que se tienen las características de estos pacientes en el

espacio latente, al igual que en los datos de metilación, se calculan las distancias, Ecuación (4.19), y la probabilidad de pertenencia a la clase benigna y maligna para cada *patch* de cada paciente. Una vez que se tiene el pronóstico de todos los *patches* de cada paciente se utiliza la técnica de *majority voting* para realizar el pronóstico a nivel de paciente.

Experimentación y Resultados

Índice de contenidos

| | |
|--|----|
| 5.1 Experimentos realizados | 56 |
| 5.1.1 Modelos basados en datos de metilación | 56 |
| 5.1.2 Modelos basados en imágenes | 59 |
| 5.1.3 Modelos basados en fusión de datos | 60 |
| 5.2 Métricas utilizadas para la validación de los modelos | 61 |
| 5.3 Resultados y discusión | 63 |
| 5.3.1 Resultados sobre la base de datos de metilación | 63 |
| 5.3.2 Resultados sobre la base de datos de imágenes histológicas | 68 |
| 5.3.3 Resultados sobre fusión de datos | 72 |

5.1 Experimentos realizados

A continuación, se exponen los parámetros e hiperparámetros utilizados para cada uno de los modelos desarrollados, con el objetivo de llevar a cabo la clasificación de los tumores melanocíticos spitzoides. Con el fin de reducir la información proporcionada en esta sección, cabe destacar que únicamente se muestran los parámetros e hiperparámetros con los que se ha obtenido los mejores resultados. Sin embargo, durante el desarrollo de este proyecto, se ha llevado a cabo una rigurosa exploración en busca de la combinación óptima de hiperparámetros y arquitecturas.

5.1.1 Modelos basados en datos de metilación

En este apartado se muestran las diferentes combinaciones de los algoritmos de clasificación propuestos sobre los datos de metilación.

- *Autoencoder + MLP*

Como se ha explicado anteriormente, el objetivo de los algoritmos basados en *autoencoder* consiste en la reducción de la dimensionalidad de los datos para obtener así una representación comprimida que exponga las características más discriminatorias de los datos de entrada. En la Tabla 5.1, se muestra la arquitectura del *autoencoder* que proporciona los mejores resultados.

| Layer name | Output shape |
|-------------|--------------|
| Input_layer | 62 |
| Encoder_0 | 20 |
| Encoder_1 | 7 |
| Decoder_1 | 20 |
| Decoder_2 | 62 |

Tabla 5.1: Modelo del *autoencoder* propuesto para la reducción de la dimensionalidad de los datos de metilación. En esta tabla se expone el número de capas que componen la parte del *encoder* y el *decoder* así como el número de neuronas de cada capa.

La parte del *encoder* y *decoder* está compuesta por una serie de capas densas o capas completamente conectadas que utilizan la función ReLU como función de activación. Como se puede observar en la Tabla 5.1, la capa de entrada del *autoencoder* está compuesta por 62 neuronas, que son el número de sitios CpG discriminatorios obtenidos tras la realización del análisis estadístico. A esta capa se le une otra de 20 neuronas que conecta con la capa de 7 neuronas que representa al espacio latente. Posteriormente, en la parte del *decoder* el número de neuronas asciende hasta llegar a la dimensión de las características de entrada. Como inicializador de los pesos w que definen estas capas se ha utilizado la distribución conocida como *Glorot Uniform* que toma muestras de una distribución uniforme dentro de $[-limit, limit]$, donde $limit = \sqrt{6/(fan_{in} + fan_{out})}$, siendo fan_{in} el número de unidades de entrada y fan_{out} el número de unidades de salida.

Con respecto a la selección de la combinación de hiperparámetros para el entrenamiento del *autoencoder*, el optimizador de descenso de gradiente estocástico (SGD) con una tasa de aprendizaje de 0.1 expuso las mejores curvas de aprendizaje cuando el modelo se propagó hacia adelante y hacia atrás. El *autoencoder* se entrenó durante 900 épocas con un *batch-size* de 8 muestras, tratando de minimizar la función de pérdidas de error cuadrático medio.

Tras llevar a cabo el entrenamiento del *autoencoder*, extraer las características obtenidas en el espacio latente y realizar la partición de datos correspondiente, las características obtenidas a la salida del *encoder* se utilizaron como *input* del perceptrón multicapa implementado para llevar a cabo el problema de

clasificación binaria propuesto. De este modo, el perceptrón estaba formado por tres capas, la de entrada de 7 neuronas, dimensión del espacio latente de la arquitectura del autoencoder, una única capa oculta de 4 neuronas, y finalmente, las dos neuronas de salida, que se corresponden con las clases melanoma y benigno, respectivamente. La función de activación de las neuronas de las dos primeras capas fue ReLU, mientras que la de la última capa, fue la función *softmax*. El modelo se entrenó durante 300 épocas utilizando el optimizador SGD con una tasa de aprendizaje de 0.01 y un *batch-size* de 8. Cabe destacar que, durante el entrenamiento, la tasa de aprendizaje se redujo un factor igual a 0.9 en el caso de que durante 10 épocas las pérdidas se mantuviesen constantes. Como función de pérdidas se utilizó la función *weighted binary cross entropy* que pondera por un factor de penalización γ en función del número de muestras de cada clase con el objetivo de tener en cuenta el desbalanceo en el número de pacientes benignos y malignos.

- *Variational autoencoder + MLP*

En este caso, la arquitectura del *variational autoencoder* propuesta es la misma que la que se utilizó para el autoencoder convencional, capas densas compuestas por 62, 20 y 7 neuronas, Tabla 5.2. En este caso, cada capa densa está acompañada de la función sigmoide como función de activación. En cuanto a la combinación de hiperparámetros, los que proporcionaron los mejores resultados han sido los mismos que en la arquitectura del autoencoder, SGD como optimizador, 0.1 como *learning rate*, 900 épocas y un *batch-size* de 8 muestras. En este enfoque se trataron de minimizar la función de pérdidas de error cuadrático medio (pérdidas de reconstrucción) y la de *Kullback–Leibler divergence* (pérdidas de regularización).

El perceptrón multicapa utilizado tras la reducción de la dimensionalidad y la partición de datos correspondiente está formado por tres capas, la de entrada de 7 neuronas, una única capa oculta de 3, y finalmente, las dos neuronas de salida, que se corresponden con las clases melanoma y benigno, respectivamente. La activación de las neuronas de las dos primeras capas fue ReLU, mientras que la de la última capa, fue *softmax*. El modelo se entrenó durante 150 épocas utilizando el optimizador SGD con una tasa de aprendizaje de 0.001 y un *batch-size* de 8 muestras. Como función de pérdidas se utilizó la función *binary cross entropy* ponderada por un factor de penalización γ en función del número de muestras de cada clase.

| Layer name | Output shape | Connected to |
|-------------------------|--------------|--|
| Input_layer | 62 | N/A |
| Encoder_0 | 20 | Input_layer |
| μ _Layer | 7 | Encoder_0 |
| $\log(\sigma^2)$ _Layer | 7 | Encoder_0 |
| Sampling_Layer | 7 | [\mathmu_Layer ; \log(\sigma^2)_Layer] |
| Decoder_1 | 20 | Sampling_Layer |
| Decoder_0 | 62 | Decoder_1 |

Tabla 5.2: Modelo del *variational autoencoder* propuesto para para la reducción de los datos de metilación. En esta tabla se expone el número de capas que componen la parte del *encoder* y *decoder* así como el número de neuronas de cada capa.

- *Deep clustering*

Determinar los hiperparámetros por *cross-validation* en un *set* de validación en algoritmos de clustering no es posible ya que la base de datos al completo es el *input* del algoritmo. Acorde con los experimentos realizados anteriormente, la arquitectura del *autoencoder* propuesta para la inicialización de los centroides de los *cluster* ha sido la misma que aquella utilizada en el modelo que constaba de *autoencoder* +

MLP. Con respecto a la selección de la combinación de hiperparámetros del pre-entrenamiento del *autoencoder*, al igual que en el algoritmo anteriormente descrito, el optimizador de descenso de gradiente estocástico (SGD) con una tasa de aprendizaje de 0.1 obtuvo las mejores curvas de aprendizaje. El *autoencoder* se pre-entrenó durante 300 épocas con un *batch-size* de 8, tratando de minimizar la función de pérdida de error cuadrático medio (MSE). Con el objetivo de inicializar los centroides de los *cluster*, en este caso $k = 2$, el algoritmo de *clustering* basado en *k-means* se utilizó en el espacio de características latentes después de realizar el pre-entrenamiento del *autoencoder*. En concreto, se realizaron 80 ejecuciones automáticas del algoritmo *K-means* y se seleccionó la mejor solución.

Tras pre-entrenar el modelo de *autoencoder*, se agregó la capa de *clustering* propuesta a la salida del espacio latente del *autoencoder*, Tabla 5.3.

| Layer name | Output shape | Connected to |
|------------------|--------------|--------------|
| Input_layer | 62 | N/A |
| Encoder_0 | 20 | Input_layer |
| Encoder_1 | 7 | Encoder_0 |
| Decoder_1 | 20 | Encoder_1 |
| Clustering_layer | 2 | Encoder_1 |
| Decoder_0 | 62 | Decoder_1 |

Tabla 5.3: Arquitectura del modelo de *deep clustering* propuesto. Como se puede observar, este algoritmo está compuesto por un *autoencoder* (parte de *encoder* y *decoder*) al que se le ha unido en la salida del *encoder* una capa de *clustering* para llevar a cabo el proceso de clasificación no supervisado.

El enfoque de *deep clustering* propuesto está optimizado por Adadelta con una tasa de aprendizaje de 1 y un valor de ρ , el cual corresponde el *decay* del optimizador, de 0.95. El método propuesto fue entrenado durante 240 épocas utilizando un *batch-size* de 8 muestras. La distribución objetivo de la capa de *clustering* se actualizó cada 10 iteraciones. Como se ha comentado en la sección 4.4.2, en este caso, la función de pérdidas para llevar a cabo el entrenamiento del algoritmo está constituida por dos términos. El primer término corresponde a la pérdida del *autoencoder* cuyo objetivo es la obtención de la máxima similitud entre la salida del *autoencoder* y los datos de entrada. En este caso, se utilizó el error cuadrático medio. El segundo término de la función de pérdidas se corresponde a la pérdida KL, que se minimiza con el objetivo de llevar a cabo un refinamiento de los centroides que garantice un agrupamiento entre clases óptimo. Para la aplicación propuesta, el valor de β , que pondera los términos que componen la función de pérdidas, se estableció a 0.7.

Tras el desarrollo de los algoritmos de clasificación para los pacientes con tumores melanocíticos benignos y malignos, se desarrollaron los modelos para el pronóstico de los pacientes con potencial maligno incierto.

- Pronóstico potencial maligno incierto (PMI)

Para el pronóstico de los pacientes con potencial maligno incierto se ha hecho uso de la metodología explicada en la sección 4.5. Tras aplicar el análisis estadístico realizado sobre los pacientes de entrenamiento a estos pacientes, se utilizó el *autoencoder* del algoritmo *deep clustering* que se entrenó con los pacientes benignos y malignos para embeber las características de los pacientes con potencial maligno incierto en el espacio latente. Como se verá en la sección posterior, este método es el que consigue representar las características de entrada de una forma más discriminatoria. Posteriormente, se establecieron dos centroides en el espacio latente correspondientes a las clases melanoma y benigno, obtenidos con las características embebidas de ambos grupos, y se utilizó la distancia euclídea de las caracterís-

ticas embebidas de un nuevo paciente PMI a cada uno de estos centroides prototipo para predecir su pronóstico.

5.1.2 Modelos basados en imágenes

En este apartado se explican los algoritmos propuestos para la clasificación utilizando las imágenes histopatológicas.

■ Redes neuronales convolucionales

Como se ha explicado en la sección 4.4.3, para la clasificación de tumores melanocíticos benignos y malignos se ha hecho uso de redes neuronales convolucionales con la combinación de técnicas basadas en transferencia del conocimiento. La arquitectura utilizada como *base model* que mejor resultados ha proporcionado para la clasificación de lesiones melanocíticas spitzoides ha sido la VGG16, que tenía como entrada los *patches* de las WSI con un tamaño de $256 \times 256 \times 3$. En concreto, la estrategia óptima fue el re-entrenamiento de todos los bloques convolucionales que componen la VGG16 excepto el primero. Posteriormente, se implementó el *top model* que en este caso estaba constituido por una capa *flatten*, una capa densa de 128 neuronas + ReLU y una capa de dos neuronas con la función *softmax* como función de activación. Además, se ha hecho uso del módulo de atención explicado la sección 4.4.3 y se ha estudiado la influencia de la incorporación de este entre el *base model* y el *top model*, Figura 5.1.

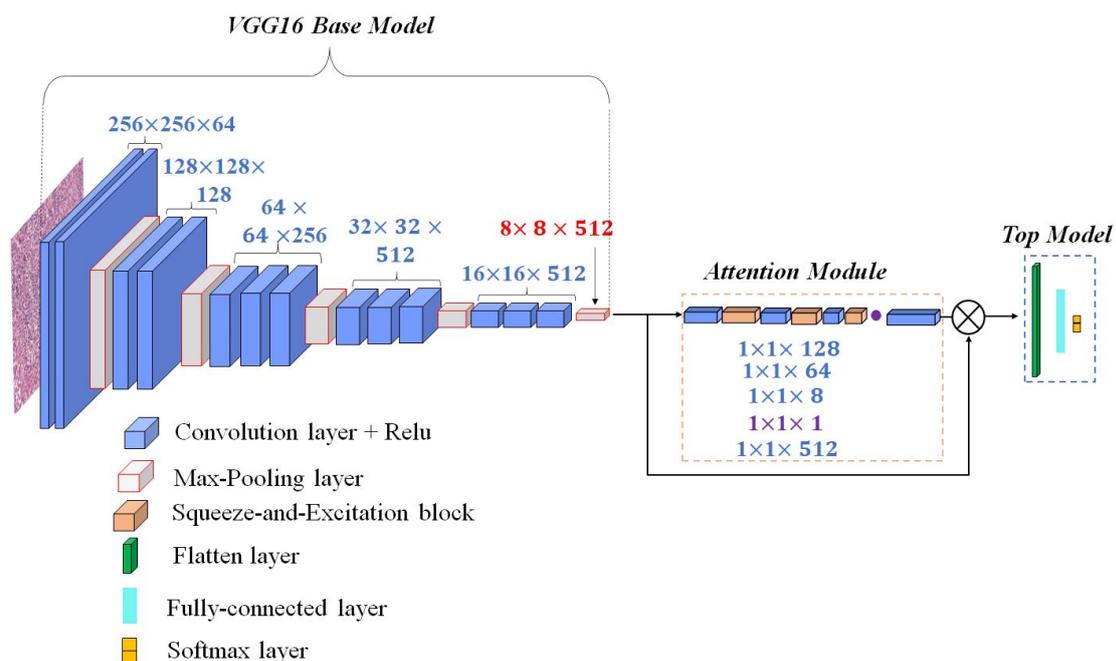


Figura 5.1: Arquitectura convolucional propuesta para la clasificación de lesiones melanocíticas spitzoides. Cabe destacar que en algunos casos el tamaño de los bloques no se corresponde con su dimensión real.

La arquitectura del módulo de atención que mejor resultados proporcionó estaba constituida por capas convolucionales de 128, 64 y 8 filtros para la reducción de la dimensionalidad de los mapas de características en la dimensión espacial con bloques *squeeze and excitation* y una capa convolucional de 1 filtro seguido por una sigmoide. Posteriormente, a esta capa se le une otra capa convolucional con el objetivo de obtener el mismo número de filtros que a la salida del *base model* implementado. La salida

del módulo de atención se ponderó con la salida del *base model* y esta fue la entrada del *top model* citado anteriormente.

Las arquitecturas propuestas se entrenaron durante 150 épocas utilizando Adadelta como optimizador con un *learning rate* de 0.001, un *batch-size* de 64 muestras y la *binary cross entropy* como función de pérdidas. Cabe destacar que también se ha estudiado la influencia del acondicionamiento de los datos explicado en la sección 4.2.2.

De la misma forma que se ha realizado con los datos de metilación, una vez que se tiene definido la arquitectura y los hiperparámetros que se utilizan para la distinción de lesiones melanocíticas benignas y malignas, se procede a la definición del algoritmo para el pronóstico de los pacientes con potencial maligno incierto.

- Pronóstico potencial maligno incierto (PMI)

Para el desarrollo del pronóstico de los pacientes con potencial maligno incierto se ha entrenado el modelo propuesto para la clasificación de lesiones melanocíticas benignas y malignas con todos los datos de la base de datos utilizando los mismos hiperparámetros. Posteriormente, se han extraído las características de los pacientes benignos y malignos tras el *base model + attention module*, es decir las características a la salida de la capa *flatten*, se han calculado los centroides de ambos grupos, se ha realizado una predicción de las características de los pacientes con potencial maligno incierto y por último se ha llevado a cabo un pronóstico de estos pacientes en función de las distancias a cada prototipo o centroide.

5.1.3 Modelos basados en fusión de datos

En este apartado se expone la arquitectura que ha sido utilizada para la clasificación de las lesiones melanocíticas spitzoides mediante la unión de los datos de metilación y de las imágenes. La idea de este enfoque es fusionar la información extraída de los *patches* de las imágenes histológicas y aquella obtenida mediante el análisis de metilación del ADN. Para ello, se ha utilizado la arquitectura mostrada en la Figura 5.2.

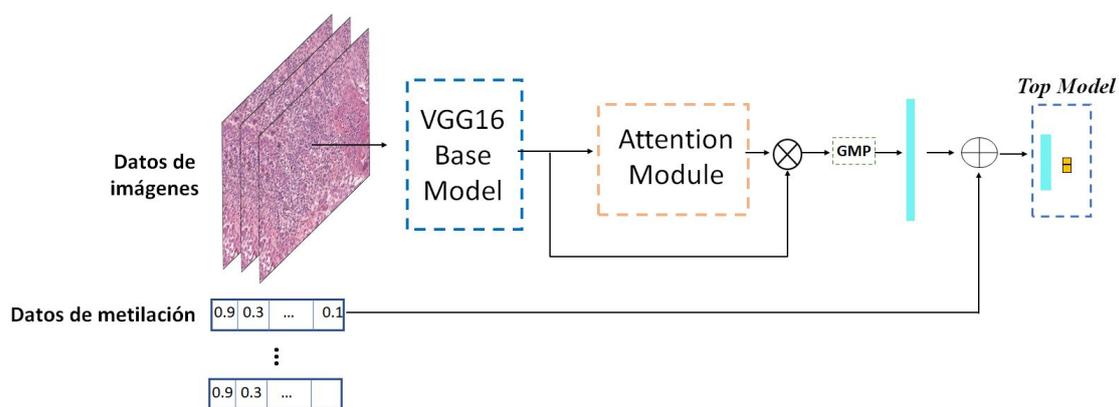


Figura 5.2: Arquitectura propuesta para la clasificación de las lesiones melanocíticas a nivel de paciente mediante la fusión de las características de las imágenes histopatológicas y los datos de metilación.

Los *input* de esta arquitectura son todos los *patches* de un paciente, cuyo número varía entre pacientes, así como el vector correspondiente a los datos de metilación. Una vez que se tienen los datos de entrada, se utiliza el mismo *base model* y *attention module* que se utilizó en la sección anterior para la

extracción automática de características de las imágenes. Posteriormente, se aplica la función de *global max pooling* (GMP) sobre los mapas de características extraídos, generando un vector de dimensiones $N_{patches} \times N_{filtros}$, siendo $N_{patches}$ el número de *patches* de cada *batch-size* y $N_{filtros}$ el número de filtros de los mapas de características obtenidos tras *base model + module attention*. Cabe destacar que en este caso para cada *patch* de cada paciente se le ha asignado el vector de metilación correspondiente a ese paciente. Tras realizar el *global max-pooling* sobre los mapas de características, se ha utilizado una capa densa de 62 neuronas para llevar las características procedentes de las imágenes a la misma dimensión que el vector de metilación entrante (62 sitios CpG). Posteriormente, se lleva a cabo un *skip connection*¹ entre las características de las imágenes y los datos de metilación. Por último, estas características concatenadas pasan al *top model*. Este está conformado por una capa densa de 16 neuronas y una *softmax* con dos neuronas para llevar a cabo la clasificación propuesta.

En cuanto a los hiperparámetros, en este caso los que aportaron mejores resultados fueron el optimizador SGD, con un *learning rate*=0.01, un *batch-size* de 64 y un número total de épocas igual a 150. Se utilizó la función *binary cross-entropy* como función de pérdidas.

5.2 Métricas utilizadas para la validación de los modelos

Una vez se han entrenado todos los modelos desarrollados en este proyecto, se procede a realizar una evaluación de los resultados obtenidos en la etapa de validación con el fin de comprobar la validez de los modelos implementados y poder compararlos con los propuestos en el estado del arte. Sin embargo, la comparación de los resultados obtenidos con los propuestos en el estado del arte es muy complicada de llevar a cabo debido a diversos factores entre los que se encuentran, la modalidad de los datos utilizados así como la finalidad del estudio propuesto.

A continuación, se definen y se explican las medidas que se utilizarán como indicadores para la evaluación de los modelos de clasificación binaria (ver Figura 5.3).

| | | Estimación del modelo | |
|----------|----------|-------------------------|-------------------------|
| | | Melanoma | Benigno |
| Realidad | Melanoma | Verdadero Positivo (VP) | Falso Negativo (FN) |
| | Benigno | Falso Positivo (FP) | Verdadero Negativo (VN) |

Figura 5.3: Visualización de los cuatro tipos de soluciones en función de la estimación del modelo y la etiqueta de cada muestra.

Teniendo en cuenta los cuatro resultados existentes (VP, VN, FP y FN), se exponen las diferentes métricas utilizadas:

- **Sensibilidad:** define la efectividad de un clasificador en la identificación de las etiquetas positivas, en este caso, pacientes con melanoma.

$$Sensibilidad = \frac{VP}{VP + FN} \quad (5.1)$$

¹El término *skip connection* hace referencia a la concatenación de dos vectores.

- **Especificidad:** define la efectividad de un clasificador en la identificación de las etiquetas negativas, en este caso los pacientes con neoplasias benignas.

$$Especificidad = \frac{VN}{VN + FP} \quad (5.2)$$

- **Valor predictivo positivo (VPP) o precisión:** eficiencia real de una prueba diagnóstica cuyo resultado es positivo. Mide la probabilidad de que un paciente, en este caso diagnosticado como melanoma, realmente tenga melanoma.

$$VPP = \frac{VP}{VP + FP} \quad (5.3)$$

- **Valor predictivo negativo (VPN):** eficiencia real de una prueba diagnóstica cuyo resultado es negativo. Mide la probabilidad de que un paciente, en este caso diagnosticado como benigno, realmente presente una neoplasia benigna.

$$VPN = \frac{VN}{VN + FN} \quad (5.4)$$

- **F1score:** determina la exactitud de un test considerando la media entre la sensibilidad y la precisión.

$$F1score = \frac{2 \cdot VPP \cdot Sensibilidad}{VPP + Sensibilidad} \quad (5.5)$$

- **Accuracy (ACC):** determina la exactitud con la que un clasificador es capaz de etiquetar correctamente las muestras. En el caso de los algoritmos supervisados se encuentra definida por la siguiente fórmula:

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \quad (5.6)$$

En el caso de los algoritmos no supervisados, se define la *unsupervised clustering accuracy (ACC)* [43], cuyo cálculo sigue la siguiente ecuación:

$$ACC = \max \frac{\sum_{i=1}^n 1\{y_i = m(c_i)\}}{n} \quad (5.7)$$

donde y_i es la etiqueta, c_i es la asignación de cluster generada por el algoritmo, y m es una función de mapeo que se extiende sobre todos los mapeos posibles entre las asignaciones y las etiquetas. Esta métrica encuentra la mejor coincidencia entre las asignaciones de clúster y la etiqueta real. La función de mapeo óptima puede ser eficientemente calculada por el algoritmo *Hungarian* [74].

- **Area Under Curve (AUC):** Mide el área bajo la Curva ROC (del inglés, Receiver Operating Characteristic). La Curva ROC es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según varía el umbral de discriminación (valor a partir del cual se decide que un caso tiene melanoma). Para ello, se representa gráficamente la tasa de verdaderos positivos (*True Positive Rate*) contra la tasa de falsos positivos (*False Positive Rate*). Cuanto mayor sea el área bajo la curva, mayor será la capacidad discriminatoria del clasificador.

- **Ratio de error (%)**: mide el porcentaje de error que existe en el modelo que se está evaluando. Se calcula con la siguiente fórmula:

$$Error\ rate\ (\%) = (1 - ACC) \cdot 100 \quad (5.8)$$

- **Predicción de clase**: estas métricas miden el número de pacientes con neoplasias malignas que son clasificados como neoplasias benignas y viceversa. Estas métricas son designadas como **Cáncer** → **Sano** y **Sano** → **Cáncer**, respectivamente.

5.3 Resultados y discusión

A continuación se presentan y discuten los resultados obtenidos para cada modalidad de la base de datos de lesiones melanocíticas spitzoides utilizada.

5.3.1 Resultados sobre la base de datos de metilación

En esta sección, se presentan los resultados obtenidos utilizando los datos de metilación para cada uno de los modelos propuestos, tanto para la clasificación de los tumores melanocíticos spitzoides (benignos y malignos) como para el pronóstico de los tumores de potencial maligno incierto.

Respecto a los algoritmos de reducción de la dimensionalidad propuestos, *autoencoder* (AE) y *variational autoencoder* (VAE), en la Figura 5.4, se pueden observar las pérdidas obtenidas durante el entrenamiento de estos algoritmos no supervisados.

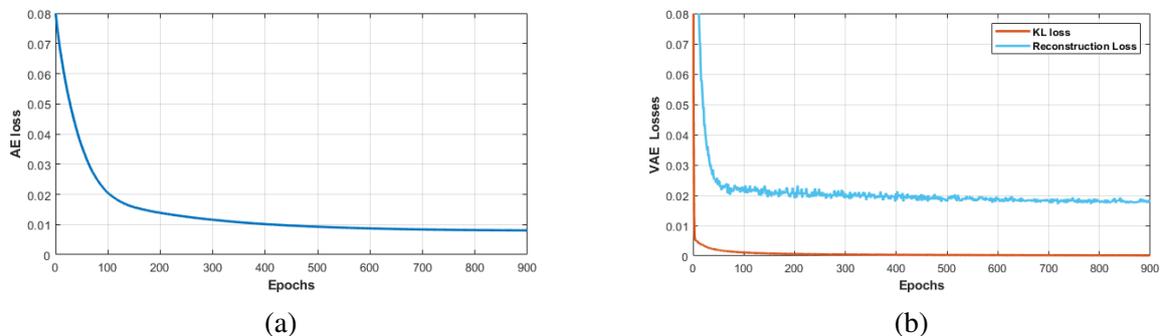


Figura 5.4: Pérdidas de los modelos de reducción de la dimensionalidad propuestos. (a) Pérdidas de reconstrucción obtenidas con el algoritmo de autoencoder. (b) Pérdidas de reconstrucción y de regularización (KL) obtenidas por el autoencoder variacional.

Como se puede observar en la Figura 5.4 (b), la pérdida del algoritmo basado en *variational autoencoder* está compuesta por dos términos, el término derivado de la reconstrucción de los datos (*Reconstruction loss*) y el término *Kullback–Leibler divergence* (KL Loss) que mide la diferencia entre la distribución normal con los parámetros μ y σ predichos por el algoritmo y la distribución normal estándar. Sin embargo, tal y como se muestra en la Figura 5.4 (a), la función de pérdidas que se minimiza en el algoritmo basado en *autoencoder* convencional solo está formada por un término que corresponde a las pérdidas relacionadas con la reconstrucción de los datos. Las pérdidas de reconstrucción son minimizadas cuando la diferencia entre la representación de los datos obtenida a la salida del *decoder* y la de los datos de entrada se aproxima a cero, lo que significa que la representación obtenida en el espacio latente (salida

del *encoder*) explica de forma correcta los datos de entrada. Sin embargo, las pérdidas referentes a KL *divergence* son minimizadas cuando $\mu = 0$ y $\sigma = 1$. Por tanto, el término de divergencia KL penaliza a la red por codificar observaciones a variables μ y σ que difieren significativamente de los parámetros de una distribución normal estándar. Como se puede observar en la Figura 5.4 (b), la pérdida correspondiente al término KL es prácticamente 0, lo que significa que los datos con los que se está entrenando el modelo no están dispersos en el espacio y rápidamente se pueden modelar como una distribución normal estándar.

En la Figura 5.5, se comparan las pérdidas obtenidas por la arquitectura basada en *autoencoder* y la suma ponderada de las pérdidas del algoritmo basado en *variational autoencoder*.

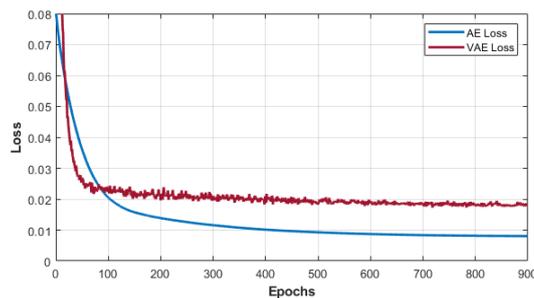


Figura 5.5: Comparación de las pérdidas obtenidas por el modelo basado en *autoencoder* (AE) y aquel basado en *variational autoencoder* (VAE).

Como se puede observar en la Figura 5.5, las pérdidas proporcionadas por el *autoencoder* son menores que aquellas obtenidas con el autoencoder variacional. Por tanto, en este caso, el *autoencoder* es capaz de llevar a cabo una mejor representación de los datos de entrada. En la Figura 5.6, se representa un *scatter plot* de las dos primeras variables, distribuidas por clases, en diferentes espacios de datos. Como se puede observar en la Figura 5.6 (b), que representa las dos primeras características del espacio latente proporcionado por el *autoencoder*, existe una cierta ordenación de los datos con respecto a la Figura 5.6 (a), espacio inicial de los datos. Es por ello que, se puede deducir que en este caso la representación de datos del *autoencoder* es óptima ya que permite una reordenación más distintiva de los mismos en sus diferentes clases. En el caso de la la Figura 5.6 (c), que representa la distribución de datos dada por el autoencoder variacional para las dos primeras dimensiones del espacio latente, se puede observar como los datos intentan concentrarse siguiendo una distribución con $\mu = 0$ y $\sigma = 1$, es decir, intentan centrarse alrededor del 0. Sin embargo, tal y como se puede deducir en base a estas gráficas, en este caso el autoencoder variacional no aporta un valor añadido al problema propuesto, ya que en este caso los datos se encuentran bastante concentrados en el espacio latente. En la Figura 5.6 (d), se representan los las dos primeras componentes del espacio latente de autoencoder utilizado en el algoritmo de *deep clustering* una vez que este ha sido unido a la capa de *clustering* propuesta. Como se puede observar en esta figura, el espacio latente del algoritmo de *deep clustering* propuesto es el que permite una distribución de los datos más distintiva.

A continuación, en las Tablas 5.4 y 5.5 se muestran los resultados de todos los algoritmos de clasificación propuestos en este proyecto. Como se ha comentado en secciones anteriores, los dos primeros algoritmos, AE + MLP y VAE + MLP, están basados en métodos supervisados ya que aunque la primera fase de reducción de la dimensionalidad es un método no supervisado, el posterior clasificador, en este caso perceptrón multicapa (MLP), es un algoritmo de aprendizaje supervisado. Además, en estas tablas se muestran los resultados de dos enfoques totalmente no supervisados. En primer lugar, se muestran los

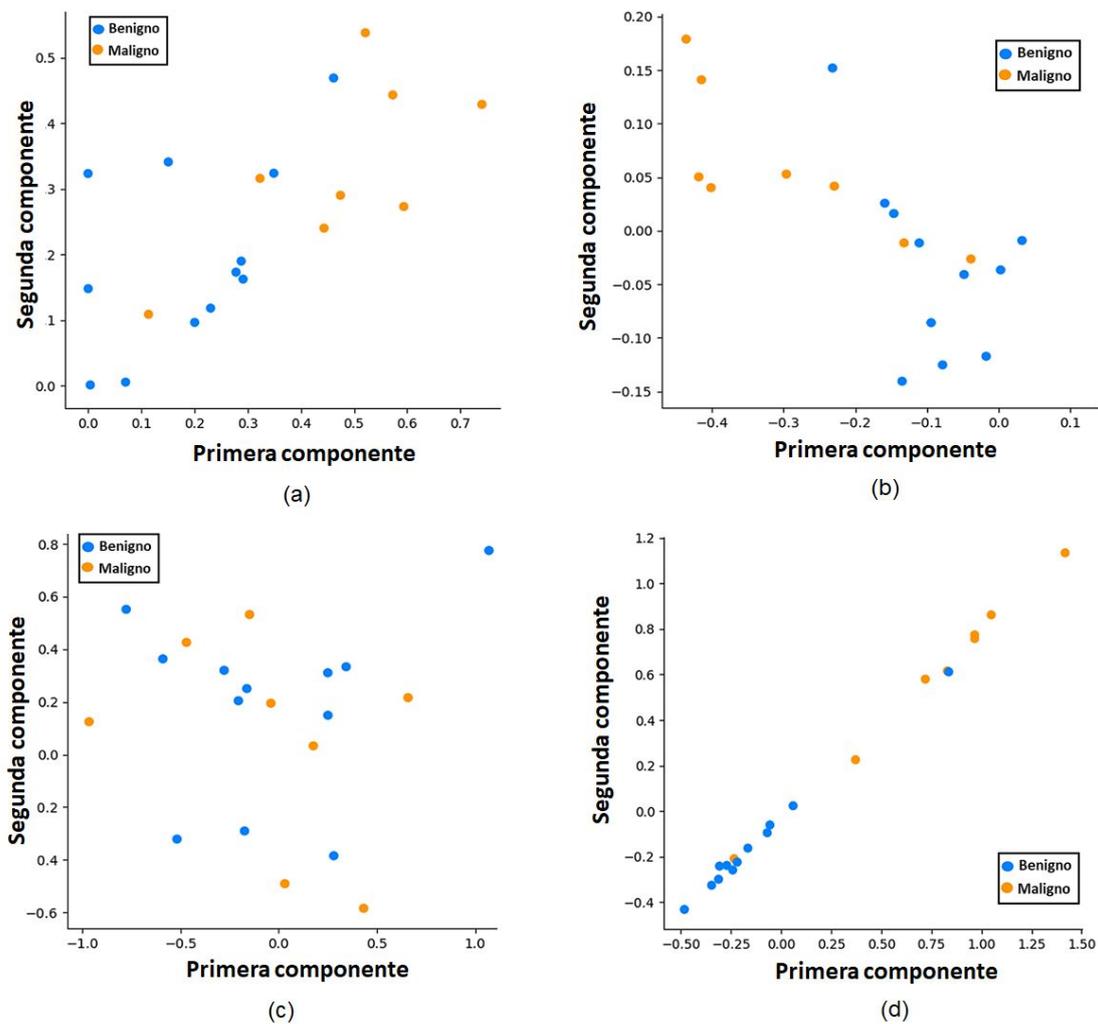


Figura 5.6: Representación de las dos primeras variables en distintos espacios de datos. (a) Visualización de las dos primeras variables de los datos de entrada iniciales. (b) Visualización de las dos primeras variables en el espacio latente del algoritmo basado en autoencoder. (c) Visualización de las dos primeras variables en el espacio latente del autoencoder variacional. (d) Visualización de las dos primeras características del espacio latente tras el entrenamiento del algoritmo basado en *deep clustering*.

resultados que obtiene el clasificador *K-means* cuando se aplica en el espacio latente del autoencoder. Cabe destacar que los centroides obtenidos mediante este clasificador se utilizarán como inicializador de los centroides del enfoque de *deep clustering* desarrollado. En segundo lugar, se exponen los resultados del método de *deep clustering* en que se lleva a cabo una clasificación no supervisada al tiempo que optimiza la reducción de la dimensionalidad. Además, en la Figura 5.7, se representa la curva ROC para visualizar las diferencias entre los métodos propuestos.

Como era de esperar, el algoritmo basado en AE+MLP obtiene en promedio mejor resultado que el mostrado por el algoritmo VAE+MLP. Como se ha observado en la Figura 5.6, el algoritmo basado en *variational autoencoder* no muestra en el espacio latente una distinción clara entre benigno y maligno, por lo que el perceptrón multicapa no logra diferenciar en este caso entre estas dos clases, clasificando la mayoría de ejemplos como benignos. Respecto a los algoritmos no supervisados utilizados en este estudio, los resultados obtenidos por el clasificador *K-means* aplicado en el espacio latente de caracte-

| | AE + MLP | VAE + MLP | AE + K-means | Deep clustering |
|-----|----------|-----------|--------------|-----------------|
| SN | 0.75 | 0.25 | 0.75 | 0.88 |
| SPC | 0.92 | 1 | 0.92 | 0.92 |
| PPV | 0.86 | 1 | 0.86 | 0.88 |
| NPV | 0.85 | 0.67 | 0.85 | 0.92 |
| FIS | 0.8 | 0.40 | 0.80 | 0.88 |
| ACC | 0.85 | 0.65 | 0.85 | 0.90 |
| AUC | 0.90 | 0.70 | 0.91 | 0.90 |

Tabla 5.4: Resultados obtenidos para los algoritmos de clasificación propuestos. Cabe destacar que para el caso de los algoritmos supervisados esta tabla muestran los resultados obtenidos en validación. Sin embargo, en el caso de los algoritmos no supervisados, esta tabla muestra los resultados obtenidos tras el entrenamiento de los modelos.

| Method | Ratio de error(%) | Cáncer → Sano | Sano → Cáncer |
|-----------------|-------------------|---------------|---------------|
| AE +MLP | 15.00 | 2 | 1 |
| VAE + MLP | 35.00 | 6 | 0 |
| AE + K-means | 15.00 | 1 | 2 |
| Deep Clustering | 10.00 | 1 | 1 |

Tabla 5.5: Ratio de error y número de ejemplos en los que los diferentes clasificadores dan una respuesta errónea.

rísticas del *autoencoder* obtiene resultados muy parecidos al perceptrón multicapa (MLP) que utiliza el mismo espacio latente, presentando este último un área bajo la curva relativamente menor. Sin embargo, la principal diferencia entre estos dos algoritmos reside en que debido a que el algoritmo *K-means* es un método no supervisado, no necesita una partición de datos en entrenamiento y validación para llevar a cabo el aprendizaje del modelo y es por ello que, el input de este algoritmo es el conjunto de todas las características del espacio latente. Por el contrario, el clasificador perceptrón multicapa necesita un proceso de entrenamiento para posteriormente obtener los resultados en un *set* distinto a aquel con el que ha sido entrenado. Es por ello, que en este caso para obtener la predicción de todos los pacientes bajo estudio es necesario llevar a cabo el algoritmo de *cross-validation* expuesto anteriormente, en el que en cada iteración solo se utiliza un porcentaje de pacientes para llevar a cabo el entrenamiento del modelo. Por último, cabe destacar que el algoritmo basado en *deep clustering* es el que logra los mejores resultados de clasificación. Estos resultados verifican lo que se mostraba en la Figura 5.6, el entrenamiento *end-to-end* en el que se minimizan las pérdidas de reconstrucción de los datos mientras se lleva a cabo un refinamiento de los centroides de cada grupo permite llevar a cabo una mejor distinción entre benigno y maligno.

En cuanto al pronóstico de los pacientes con potencial maligno incierto (PMI), en la Tabla 5.6, se pueden observar los resultados obtenidos para la clase benigno y maligno. En esta tabla, se muestra el número de pacientes que se designan automáticamente en cada clase, la probabilidad que representa cada clase con respecto al número total de pacientes y la media y desviación típica del parámetro denominado confianza, que indica la probabilidad con la que el modelo propuesto ha clasificado las muestras como malignas o benignas. Además, en la Figura 5.8 se puede observar una representación de las diferentes características que componen el espacio latente de características de los pacientes con potencial maligno incierto, en función de si han sido predichos como malignos o benignos por el algoritmo. De esta forma, se puede observar la disposición de los pacientes predichos de cada clase con respecto al centroide de su clase cuando se enfrentan las diferentes componentes de los centroides y características del espacio latente.

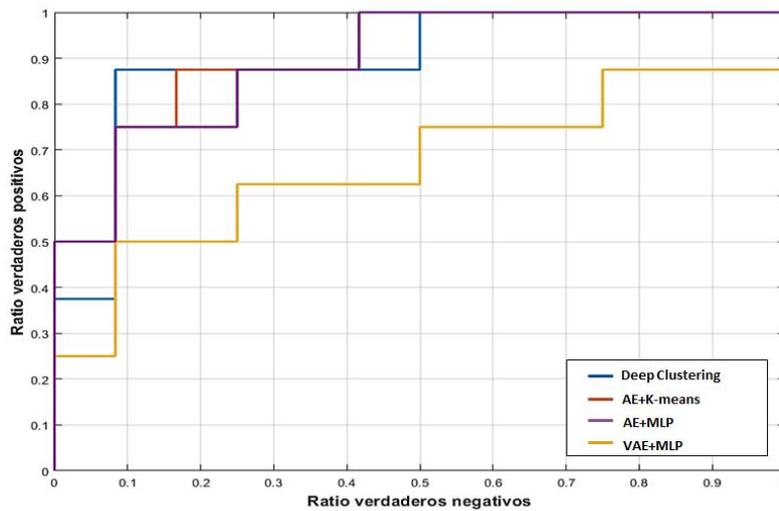


Figura 5.7: Curvas ROC correspondientes a los resultados alcanzados por los diferentes métodos propuestos.

| Pronóstico | Nº pacientes | Probabilidad Total (%) | Confianza ($\mu \pm \sigma$) |
|--------------------|--------------|------------------------|--------------------------------|
| Melanoma spitzoide | 7 | 36.84 | 0,65 \pm 0,0511 |
| Benigno | 12 | 63.16 | 0,64 \pm 0,0682 |

Tabla 5.6: Pronóstico obtenido para los pacientes con potencial maligno incierto (PMI).

Como se puede observar en la Tabla 5.6, la mayor parte de los pacientes designados como potencial maligno incierto tienen un pronóstico de benignos (63.16%), lo cual cobra sentido ya que como se ha comentado anteriormente, cuando los patólogos detectan en una muestra una malignidad representativa tienden a diagnosticarla como maligna antes de clasificarla como PMI. Con respecto a la confianza aportada por el algoritmo para la predicción de cada una de las clases, cabe destacar que esta no es muy alta y se sitúa en torno al 65%. Este resultado se puede contrastar con aquel visualizado en la Figura 5.8, donde se representa, en gráficos 2D, la distribución de los pacientes PMI con respecto al centroide de cada clase. Como se puede observar, en algunas dimensiones del espacio latente, los pacientes predichos como una clase no se encuentran concentrados alrededor del centroide de esa clase sino que se encuentran expandidos en el espacio. Este hecho explica que el porcentaje de confianza no sea excesivamente alto para cada una de las clases. Cabe destacar que el desarrollo de este tipo de algoritmos, centrados en el pronóstico de lesiones melanocíticas de potencial maligno incierto, podría orientar a los patólogos con el objetivo de realizar un seguimiento más exhaustivo a los pacientes que el algoritmo pronostica como melanomas. De esta forma, se podría llevar a cabo un diagnóstico temprano de este tipo de tumores.

Con respecto a la comparación con el estado del arte, como se ha comentado anteriormente, esta resulta complicada debido a que la mayoría de bases de datos no son públicas y por tanto, no se puede llevar a cabo una comparación directa. En este caso, aunque varios estudios se han centrado en el desarrollo de algoritmos basados en *deep learning*, como *autoencoder* y *autoencoder* variacionales, para la predicción de distintos tipos de tumores como cáncer de pecho [40], [42] o carcinomas [41], ningún algoritmo basados en inteligencia artificial se ha propuesto para la distinción de tumores melanocíticos spitzoides malignos y benignos basados en datos de metilación. Esto puede deberse a la complejidad que subyace en estos tipos de tumores. Por tanto, se puede deducir que el diagnóstico de los tumores melanocíticos spitzoides es un campo poco explorado y por tanto, los algoritmos desarrollados y los resultados obteni-

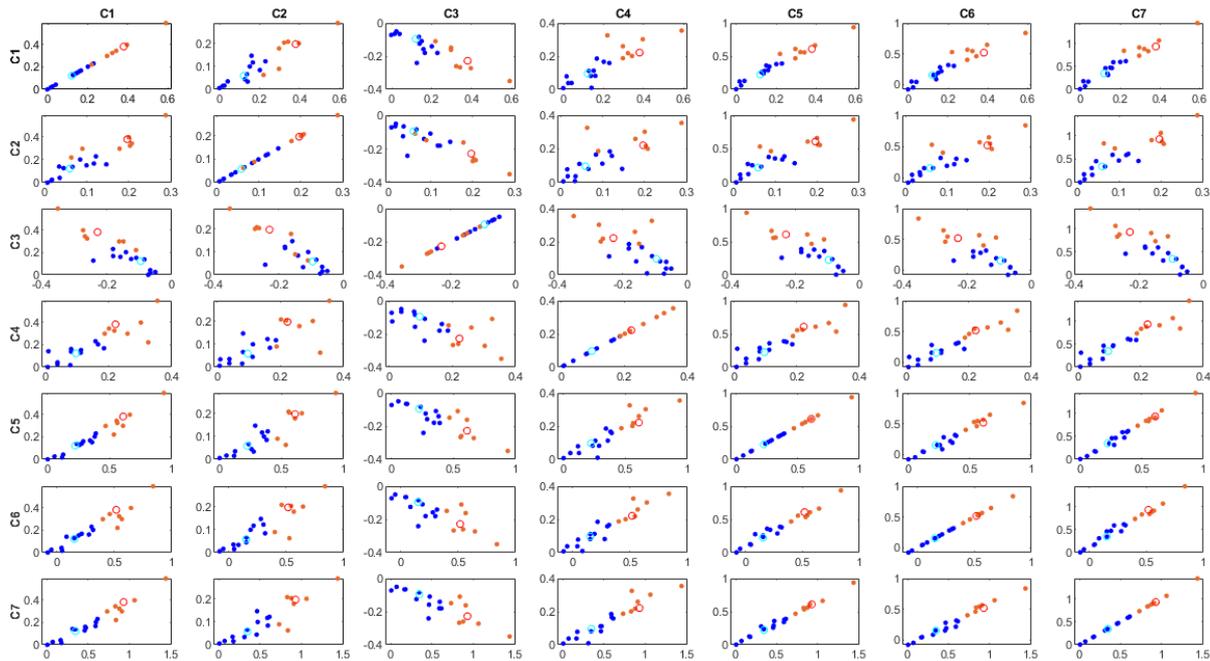


Figura 5.8: Representación de las características del espacio latente de los pacientes con potencial maligno incierto. Leyenda: el color azul representa aquellos pacientes que han sido predichos como benignos mientras que el color naranja representa aquellos pacientes que han sido predichos como malignos. Mediante los dos círculos de color cian y rojo, se representan los centroides de la clase benigna y maligna respectivamente. El acrónimo C hace referencia al número de componente del centroide y de las características del espacio latente que se está representando.

dos suponen un gran avance para la investigación de este tipo de tumores. Por otro lado, es importante remarcar que como se ha mencionado anteriormente, se han desarrollado diversos algoritmos basados en inteligencia artificial que utilizan los datos de metilación para el diagnóstico de diferentes tipos de cáncer. La mayoría de estos algoritmos proponen diferentes métodos para la reducción de la dimensionalidad (*autoencoder*, *autoencoder variacionales*) con la subsecuente clasificación de los datos basada en métodos supervisados y no supervisados. Sin embargo, ningún algoritmo de clasificación en base a metilación del DNA desarrollado hasta el momento ha sido capaz de llevar a cabo una optimización de las características representadas en el espacio latente mientras que realiza el proceso de clasificación. Como se ha podido comprobar en este tipo de cáncer, estos de algoritmos mejoran los resultados obtenidos, por lo que su aplicación en el campo de la metilación del ADN puede suponer un gran avance en el diagnóstico de esta enfermedad.

5.3.2 Resultados sobre la base de datos de imágenes histológicas

En este apartado, se presentan los resultados de clasificación obtenidos tras la extracción automática de características mediante algoritmos basados en redes convolucionales aplicados a las imágenes histopatológicas spitzoides. Como se ha comentado en la sección 4.4.3, la predicción llevada a cabo por este algoritmo es a nivel de *patch*. Por tanto, para obtener las métricas que aparecen en la Tabla 5.7, que muestran los resultados por paciente, se ha hecho uso de la técnica de *majority voting* sobre todos los parches de cada paciente. En concreto, tras la realización de experimentos con diferentes *base model*, VGG19, Inception, Res-Net, en las Tablas 5.7 y 5.8 se muestran los resultados obtenidos tras realizar *fine-tuning* sobre la VGG16 puesto que es la arquitectura que mejor comportamiento ha mostrado. Por

| | Sin acondicionamiento | | Con acondicionamiento | |
|-----|-----------------------|-------------|-----------------------|-------------|
| | BM+TM | BM+AM+TM | BM+TM | BM+AM+ TM |
| SN | 0.63 | 0.88 | 0.75 | 0.88 |
| SPC | 0.75 | 0.5 | 0.67 | 0.75 |
| PPV | 0.63 | 0.54 | 0.60 | 0.70 |
| NPV | 0.75 | 0.86 | 0.80 | 0.90 |
| FIS | 0.63 | 0.67 | 0.67 | 0.78 |
| AUC | 0.78 | 0.79 | 0.76 | 0.83 |
| ACC | 0.7 | 0.65 | 0.70 | 0.80 |

Tabla 5.7: Resultados de clasificación sobre las imágenes histopatológicas obtenidos con los algoritmos propuestos. BM:base model; AM:módulo de atención; TM: top model.

tanto, la arquitectura compuesta por la VGG16 más el *top model* que se ha definido en la sección 5.1, se refleja en esta tabla como BM+TM. La unión del módulo de atención al algoritmo anteriormente mencionado se especifica como BM+AM+TM. Además, en esta tabla se establece una comparativa entre los resultados obtenidos por cada uno de estos modelos sin y con el acondicionamiento de datos explicado en la sección 4.2.2, SA y CA, respectivamente.

| Method | Ratio de error(%) | Cáncer → Sano | Sano → Cáncer |
|------------------|-------------------|---------------|---------------|
| BM + TM (SA) | 30.00 | 4 | 2 |
| BM + AM +TM (SA) | 35.00 | 3 | 4 |
| BM + TM (CA) | 30.00 | 2 | 4 |
| BM + AM+ TM (CA) | 20.00 | 1 | 3 |

Tabla 5.8: Ratio de error y número de ejemplos en los que los diferentes clasificadores dan una respuesta errónea.

Además, en la Figura 5.9, se ilustra la curva ROC de cada uno de los métodos para que se pueda observar con mayor detalle las diferencias existentes. Cabe destacar que en este caso, para calcular la probabilidad de pertenencia a la clase melanoma, que es lo que se refleja en esta curva, se ha hecho uso de la siguiente fórmula para cada paciente:

$$P_K = \frac{c^K}{\sum_i^k c^i} \quad (5.9)$$

siendo K la clase melanoma y k el número de clases en total, en este caso 2.

Como se puede observar en los resultados mostrados, el mejor resultado de clasificación se obtiene cuando se emplea el *base model* de la VGG16 y el *top model* junto con el módulo de atención. Además, se demuestra que los resultados proporcionados por este algoritmo son mejores cuando previamente se realiza el acondicionamiento de los datos, el cual consiste en balancear las distintas clases en la fase de entrenamiento y realizar *data augmentation*. Por tanto, se puede concluir que tanto el balanceo de clases como el refinamiento de características son pasos clave para llevar a cabo la clasificación de los tumores melanocíticos spitzoides.

Con el objetivo de comprobar que el módulo de atención unido al *base model* empleado permite discernir los detalles más finos de la imagen, se calculan los mapas de activación de clase (CAMs). Los mapas de activación de clase son mapas de calor que resaltan las regiones relevantes en las que el modelo propuesto está prestando atención para determinar la clase de cada parche histopatológico. En la Figura 5.10, se exponen los CAMs relativos a muestras con neoplasias melanocíticas benignas y malignas aleatorias

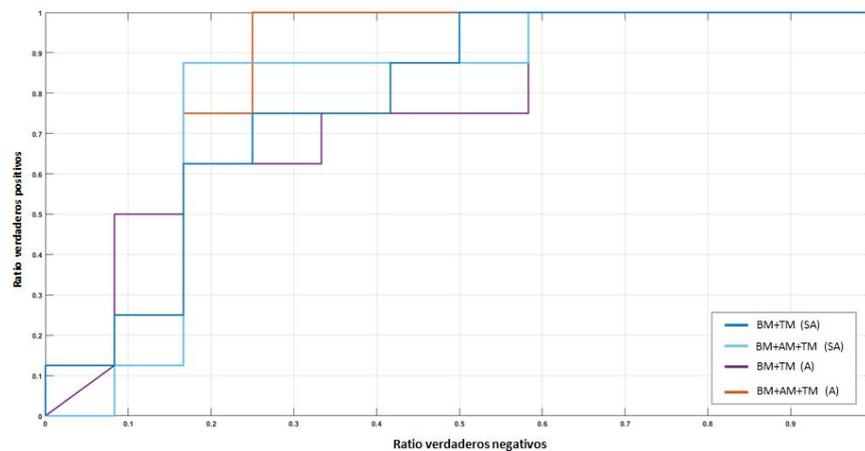


Figura 5.9: Curva ROC de los distintos métodos empleados para la clasificación de lesiones melanocíticas.

para dilucidar las diferencias entre las zonas que las arquitecturas BM y BM+AM tienen en cuenta para discernir entre clases.

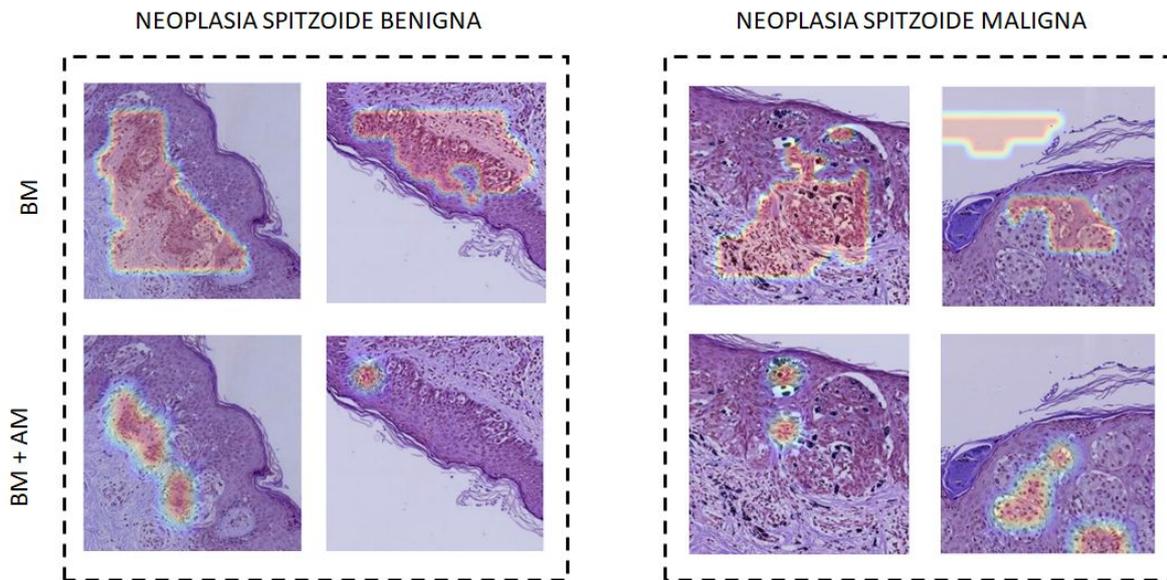


Figura 5.10: Mapas de activación de clase para muestras benignas y malignas aleatorias obtenidas por los modelos BM y BM+AM. Cada columna representa los *patches* de un paciente elegido al azar mientras que cada fila representa los resultados obtenidos por cada uno de los modelos implementados.

Como se puede observar en la Figura 5.10, el modelo basado en BM se fija en zonas amplias mientras que tras el módulo de atención estas zonas se reducen, centrándose en detalles más específicos. Además, como se puede observar en estos parches, con el uso del módulo de atención se eliminan las zonas relativas al fondo de los *patches* que no aportan información relevante para la clasificación de lesiones melanocíticas benignas y malignas. Por tanto, este refinamiento de características como se ha podido comprobar con la visualización de resultados, ayuda a obtener una clasificación más precisa en base a las imágenes histopatológicas.

En la Tabla 5.9, se muestran los resultados obtenidos para el pronóstico de las lesiones melanocíticas de potencial maligno incierto mediante el análisis de imágenes.

| Pronóstico | Nº pacientes | Probabilidad Total (%) | Confianza ($\mu \pm \sigma$) |
|--------------------|--------------|------------------------|--------------------------------|
| Melanoma spitzoide | 9 | 47.37 | $0,80 \pm 0,18$ |
| Benigno | 10 | 52.63 | $0,73 \pm 0,15$ |

Tabla 5.9: Pronóstico obtenido para los pacientes con potencial maligno incierto (PMI).

Como se puede observar en la Tabla 5.9, la confianza aportada por el modelo basado en imágenes es mayor que la que anteriormente se tenía con los datos de metilación, lo que se puede observar en una concentración de puntos alrededor de los centroides malignos y benignos, Figura 5.11. Para obtener este gráfico, dado que se partían de numerosas características embebidas, todas las características a la salida de la capa *flatten*, se ha utilizado el algoritmo *t-distributed stochastic neighbour embedding* (t-SNE) para representar los datos en un espacio bidimensional [75].

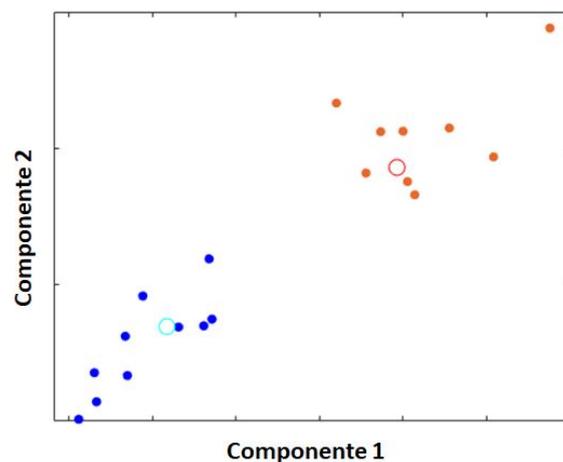


Figura 5.11: Visualización del pronóstico basado en modelos de clasificación de imágenes utilizando las primeras componentes del método *t-SNE*. Leyenda: el color azul representa aquellos pacientes que han sido predichos como benignos mientras que el color naranja representa aquellos pacientes que han sido predichos como malignos. Mediante los dos círculos de color cian y rojo, se representan los centroides de la clase benigna y maligna respectivamente.

En la Tabla 5.10, aparece una comparación de los modelos que aportan los mejores resultados tanto para la metilación del ADN (DM) como para el análisis de imágenes histopatológicas (IH). Tras observar estos resultados, se puede concluir que los los modelos que están entrenados con los datos de metilación logran obtener un resultado más satisfactorio que el estudio de las imágenes histopatológicas. Por tanto, aunque el pronóstico llevado a cabo por los métodos basados en imágenes presenten una mayor confianza son más engañosos puesto que los modelos de clasificación binaria (benigno y maligno) de los que provienen obtienen peores resultados.

Por tanto, cabe destacar que tal y como se exponía en secciones anteriores, el estudio de nuevas tecnologías de diagnóstico que complementen al estudio de imágenes histopatológicas permite mejorar el diagnóstico de las lesiones melanocíticas spitzoides. En este estudio, se puede concluir en primera instancia que los datos basados en datos de metilación permiten realizar un diagnóstico más efectivo de las lesiones melanocíticas spitzoides.

| | DM | IH |
|------------|-------------|-------------|
| SN | 0.88 | 0.88 |
| SPC | 0.92 | 0.75 |
| PPV | 0.88 | 0.70 |
| NPV | 0.92 | 0.90 |
| FIS | 0.88 | 0.78 |
| ACC | 0.90 | 0.83 |
| AUC | 0.90 | 0.80 |

Tabla 5.10: Comparación de resultados obtenidos mediante los datos de metilación (DM) y los datos procedentes de imágenes histopatológicas (IH).

Como ocurre en el caso de metilación del ADN, aunque se han desarrollado diferentes algoritmos basados en redes convolucionales y aplicados a imágenes histopatológicas para llevar a cabo el diagnóstico de diferentes tipos de cáncer como el cáncer de pecho [56]-[58], el cáncer de pulmón [61]-[63], el cáncer de próstata [59], [60] o el melanoma común [55] hasta el momento, no se ha desarrollado ningún algoritmo basado en *deep learning* que lleve a cabo la distinción de tumor melanocítico spitzoide basado en la extracción automática de características de estas imágenes. El único estudio llevado a cabo sobre imágenes histopatológicas spitzoides es el presentado en [76]. En este estudio, los autores proponen un algoritmo basado en *Random Forest* para distinguir entre lesiones melanocíticas benignas (45 pacientes) y tumores melanocíticos de potencial maligno incierto (7 pacientes). En este caso, la entrada al algoritmo no son las imágenes histopatológicas sino características extraídas manualmente tales como el tamaño del *spitz*, el índice mitótico o la existencia de una mitosis atípica. Por tanto, aunque este algoritmo obtiene un *accuracy* del 97%, no es capaz de llevar a cabo una extracción automática de características sobre las imágenes histopatológicas ni distinguir entre lesiones melanocíticas malignas y benignas, lo que es realmente importante para el desarrollo de un sistema de ayuda al diagnóstico. Por tanto, se puede concluir que en este proyecto se propone el primer algoritmo que lleva a cabo el estudio de lesiones melanocíticas spitzoides mediante imágenes histopatológicas, lo que abre un gran camino a la investigación de esta patología.

5.3.3 Resultados sobre fusión de datos

En este apartado, se presentan los resultados de clasificación obtenidos tras la extracción automática de características mediante algoritmos basados en redes convolucionales aplicados a las imágenes histopatológicas spitzoides y la fusión con los datos correspondientes a la metilación del ADN (FD). En concreto, tras la realización de experimentos con la concatenación de características a diferentes niveles, en las Tablas 5.11 y 5.12 se muestran los mejores resultados obtenidos. Además en la Figura 5.12 se muestra la curva ROC obtenida.

| Method | SN | SPC | PPV | NPV | FIS | ACC | AUC |
|--------|------|------|------|------|------|------|------|
| FD | 0.75 | 0.92 | 0.86 | 0.85 | 0.80 | 0.85 | 0.88 |

Tabla 5.11: Resultados obtenidos para los algoritmos de clasificación fusionando los datos (FD).

Como se puede observar, los resultados obtenidos tras la fusión de datos superan a los obtenidos cuando se procedía a la clasificación teniendo en cuenta únicamente las características extraídas de las imágenes. Sin embargo, cabe destacar que, estos resultados son inferiores a aquellos alcanzados en base únicamente

| Method | Ratio de error(%) | Cáncer → Sano | Sano → Cáncer |
|--------|-------------------|---------------|---------------|
| FD | 15.00 | 2 | 1 |

Tabla 5.12: Ratio de error y número de ejemplos en los que el clasificador basado en la fusión de datos (FD) da una respuesta errónea.

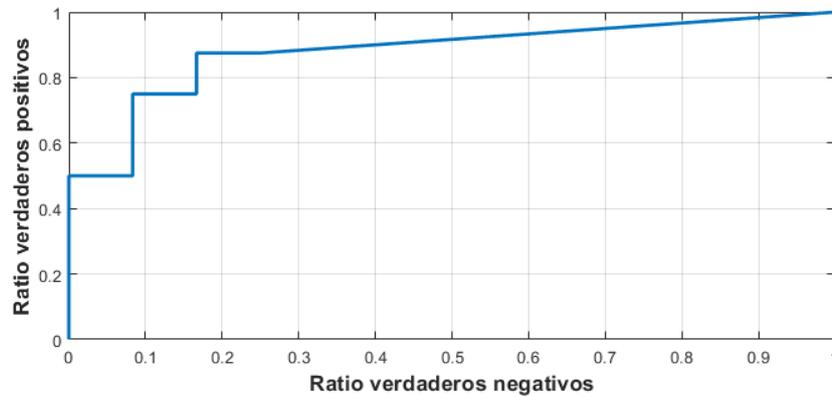


Figura 5.12: Curva ROC de los distintos métodos empleados para la clasificación de lesiones melanocíticas.

a la metilación del ADN. Por tanto, se puede concluir que en este caso, las características extraídas de los datos de metilación aportan valiosa información al análisis de las imágenes histológicas permitiendo obtener resultados más satisfactorios que cuando se procede al análisis de las imágenes por separado. En este caso, se puede observar como las características de las imágenes histológicas no aportan de interés a los datos de metilación empeorando la clasificación a cuando estos actúan por separado. Esto puede deberse a que el número de muestras que se dispone es relativamente bajo.

Por tanto, tras el desarrollo de diferentes algoritmos basados en el análisis de imágenes, de datos de metilación y de fusión de ambos datos, se puede concluir que son los datos de metilación los que aportan una información más relevante para la clasificación de las lesiones melanocíticas spitzoides.

Capítulo 6

Conclusiones y líneas futuras

Índice de contenidos

| | |
|------------------------------|----|
| 6.1 Conclusiones | 76 |
| 6.2 Líneas futuras | 78 |

6.1 Conclusiones

Las neoplasias melanocíticas spitzoides son lesiones cutáneas poco frecuentes que suponen un desafío diagnóstico. Dentro de este grupo de lesiones se encuentran los nevus spitzoides, que son totalmente benignos, los melanomas spitzoides y los tumores atípicos de Spitz. Estos últimos son los más desafiantes ya que presentan características superpuestas con tumores spitzoides benignos y malignos. Actualmente, no se ha desarrollado ningún algoritmo automático que lleve a cabo una clasificación entre tumores spitzoides benignos y malignos así como una predicción del pronóstico de los pacientes con potencial maligno incierto (tumores atípicos de Spitz). Es por ello que, la principal motivación de este TFM reside en el desarrollo de algoritmos basados en *deep learning* para la clasificación y el pronóstico de este tipo de tumores que sirvan como un sistema de ayuda al diagnóstico.

Para el desarrollo de los algoritmos de clasificación, se ha contado con una base de datos que contenía información relativa al estudio de metilación del ADN así como imágenes histopatológicas extraídas de la sección tumoral de interés. Aunque el objetivo de la utilización de estos datos era el mismo, desarrollo de sistemas de clasificación y pronóstico, los métodos desarrollados para cada modalidad de datos difieren puesto que las características de estos son distintas.

En lo relativo a los datos de metilación de ADN, se ha contado con el nivel de metilación de las islas CpG, localizadas en la región promotora de la mayoría de los genes. Tras un estudio del estado del arte, se ha podido comprobar que un nivel de metilación alto en ciertos promotores localizados en los genes supresores de tumores puede derivar en el desarrollo de diferentes tipos de cáncer. Por tanto, se ha estudiado por primera vez si el análisis del nivel de metilación en las lesiones melanocíticas spitzoides permite diferenciar un tumor benigno de uno maligno. Para llevar a cabo la clasificación de las lesiones malignas y benignas en base a los datos de metilación, primeramente, se ha llevado a cabo un análisis estadístico que permite preseleccionar los sitios CpG que pueden ser discriminatorios para la distinción de las lesiones melanocíticas. De esta forma, se reduce en gran medida el número de sitios CpG a analizar. Dado que la dimensión de los datos de metilación es relativamente grande debido al desarrollo de nuevas técnicas de análisis genómico que permiten obtener la metilación del genoma a nivel molecular, y el número de pacientes que se suele tener para estudiar esta modalidad de datos es relativamente pequeña, puesto que es una técnica cara, es necesario implementar técnicas para la reducción de la dimensionalidad de los datos. En este contexto, se han desarrollado diversas técnicas basadas en *deep learning* que permiten llevar a cabo una efectiva representación latente de las características para su posterior clasificación en base a algoritmos supervisados y no supervisados. Tras la reducción de la dimensionalidad de los datos, se realiza una partición de datos basada en la técnica *K-fold cross validation*. Esta técnica permite que los modelos de clasificación supervisados puedan ser entrenados y validados con diferentes muestras. Además, se ha desarrollado por primera vez en el estudio de datos de metilación aplicado al diagnóstico del cáncer, un algoritmo que permite reducir la dimensionalidad de los datos al mismo tiempo que lleva a cabo una clasificación de los mismos. Este tipo de metodología es completamente no supervisada, por lo que no necesita una partición de datos, y ha demostrado ser la más efectiva en el diagnóstico del melanoma spitzoide mejorando los resultados obtenidos cuando la reducción de la dimensionalidad y la clasificación se hacen por separado. Tras el desarrollo de los algoritmos de clasificación de las lesiones melanocíticas malignas y benignas, se han utilizado métodos basados en *prototypical network* para el pronóstico de los pacientes con potencial maligno incierto. De esta forma, se le pueden ofrecer unas directrices a los patólogos de cómo tienen que enfrentarse a los

pacientes que tienen este tipo de lesiones en función de si el algoritmo encuentra más similitudes entre el PMI a el prototipo de lesión maligna benigna o entre el PMI y el prototipo de muestra benigna.

En cuanto al análisis de imágenes histopatológicas, que es el estándar de oro para la clasificación automática de distintos tipos de cáncer, se han empleado distintos algoritmos que permiten mejorar el análisis y la clasificación basada en este tipo de imágenes. En primer lugar, dado que las imágenes histológicas con las que se cuenta son demasiado grandes y tienen una carga computacional elevada, se han generado *patches* de la imágenes completas. Estas sub-imágenes tienen una dimensión menor que las imágenes completas por lo que pueden ser utilizadas por los algoritmos de clasificación. Además, tras la generación de los *patches* se ha desarrollado un algoritmo para la selección automática de los *patches* relevantes, eliminando aquellos que no contiene tejido a analizar y que no representan la zona tumoral de interés. Tras la selección de estos *patches*, debido al bajo número de pacientes con los que se cuenta en este estudio, se ha llevado a cabo una técnica *K-fold cross validation* ponderada que ha permitido obtener la predicción de todos los pacientes bajo estudio. Tras el desarrollo de esta partición, se han desarrollado técnicas que permiten llevar a cabo el balanceo de las diferentes clases ya que debido a la rareza que presentan estas lesiones los tumores difieren considerablemente en tamaño. Para ello, se han elegido muestras al azar de la clase mayoritaria y se han generando muestras sintéticas de la clase minoritaria. Posteriormente, se han aumentado todos los datos de manera artificial usando la técnica conocida como *data augmentation*. Una vez realizado este pre-procesado y acondicionamiento de los datos, se ha implementado el proceso de clasificación propiamente dicho. Para ello, se han utilizado redes neuronales convolucionales empleando la técnica de *fine-tuning* sobre arquitecturas pre-entrenadas en grandes conjuntos de imágenes. Estas técnicas han demostrado ser las más efectivas cuando se trabaja con problemas desafiantes y de los cuales se tiene un número reducido de datos. Además, se han implementado módulos de atención que permiten refinar las características extraídas mejorando así la clasificación en este tipo de lesiones. Al igual que con los datos de metilación, tras el desarrollo de los modelos de clasificación se han creado modelos para el pronóstico de este tipo de lesiones.

Tras el desarrollo de algoritmos basados en datos de metilación e imágenes histopatológicas y la obtención de resultados por ambas modalidades, se ha podido comprobar que la utilización de nuevos enfoques, como la metilación del ADN, para el diagnóstico de lesiones desafiantes permite aumentar la precisión de la clasificación. Este hecho se ve reflejado cuando se comparan los resultados obtenidos mediante las imágenes histopatológicas y los datos de metilación, siendo estos últimos más altos. Por tanto, en primera instancia se puede concluir que en este tipo de lesiones el análisis de metilación es la técnica de diagnóstico más eficaz.

Además, se ha implementado un enfoque que permite llevar a cabo una clasificación en base a la fusión de distintas modalidades de datos, en este caso, de los datos de metilación del ADN e imágenes histopatológicas y se ha podido comprobar que aunque mediante esta metodología se han superado los datos obtenidos mediante el análisis de las imágenes, la fusión de datos no supera los resultados obtenidos analizando únicamente los datos de metilación.

En definitiva, se puede concluir que se ha logrado diseñar y desarrollar un sistema automático de clasificación de tumores melanocíticos spitzoides. Dado que no existe ningún estudio en la literatura que lleve a cabo el diagnóstico de las lesiones melanocíticas spitzoides, los métodos desarrollados en este proyecto no se han podido comparar con ningún método del estado del arte. Sin embargo, los algoritmos y los resultados obtenidos en este proyecto suponen un gran avance en el estudio de estas lesiones melanocíticas y por ello, ya se está trabajando en la publicación de los mismos.

6.2 Líneas futuras

Como se ha comentado a lo largo de este proyecto, este trabajo se enmarca dentro del proyecto europeo CLARIFY por lo que las líneas futuras de este trabajo se ligan a sus objetivos finales. CLARIFY promueve el desarrollo de un entorno robusto de diagnóstico digital automatizado basado en inteligencia artificial. Para ello, propone la implementación de algoritmos de datos en la nube que faciliten la interpretación de las imágenes whole-slide (WSI) y que permitan integrar la información aportada por los estudios moleculares.

Dado que ahora mismo nos encontramos en la fase temprana de este proyecto, no se han podido adquirir suficientes muestras ya que la adquisición tanto de datos de metilación como imágenes histopatológicas es un proceso costoso y que requiere bastante tiempo. Es por ello que, el número de muestras utilizado en este proyecto es relativamente bajo. Por tanto, la principal línea futura de este trabajo es la validación externa de los algoritmos aquí propuestos con nuevas muestras que permitan comprobar su robustez.

Además, como se ha nombrado anteriormente, el proyecto CLARIFY promueve el desarrollo de sistemas automáticos de diagnóstico en la nube. Por ello, líneas futuras de este trabajo estarán centradas en el desarrollo de modelos embebidos en la nube. Para ello, se pretende desarrollar una aplicación web que integre los algoritmos implementados en el presente trabajo y que permita la predicción automática de nuevas imágenes subidas a la nube por el patólogo.

En definitiva, las líneas futuras de este trabajo se centran en el desarrollo de un sistema de clasificación automatizado alojado en la nube que sirva como sistema de ayuda al diagnóstico para cualquier patólogo independientemente de su localización.

Bibliografía

- [1] Z. Apalla, A. Lallas, E. Sotiriou, E. Lazaridou y D. Ioannides, “Epidemiological trends in skin cancer”, *Dermatology practical & conceptual*, vol. 7, n.º 2, pág. 1, 2017.
- [2] U. Leiter, T. Eigentler y C. Garbe, “Epidemiology of skin cancer”, en *Sunlight, vitamin D and skin cancer*, Springer, 2014, págs. 120-140.
- [3] J. Jaworek-Korjakowska y P. Kłeczek, “Automatic classification of specific melanocytic lesions using artificial intelligence”, *BioMed research international*, vol. 2016, 2016.
- [4] M. S. Soengas y S. W. Lowe, “Apoptosis and melanoma chemoresistance”, *Oncogene*, vol. 22, n.º 20, págs. 3138-3151, 2003.
- [5] C. Monteagudo, D. Ramos, A. Pellín-Carcelén, R. Gil, R. C. Callaghan, J. M. Martín, V. Alonso, A. Murgui, L. Navarro, S. Calabuig y col., “CCL27–CCR10 and CXCL12–CXCR4 chemokine ligand-receptor mRNA expression ratio: new predictive factors of tumor progression in cutaneous malignant melanoma”, *Clinical & experimental metastasis*, vol. 29, n.º 6, págs. 625-637, 2012.
- [6] H. Xu, R. Berendt, N. Jha y M. Mandal, “Automatic measurement of melanoma depth of invasion in skin histopathological images”, *Micron*, vol. 97, págs. 56-67, 2017.
- [7] T. Wiesner, H. Kutzner, L. Cerroni, M. C. Mihm Jr, K. J. Busam y R. Murali, “Genomic aberrations in spitzoid melanocytic tumours and their implications for diagnosis, prognosis and therapy”, *Pathology*, vol. 48, n.º 2, págs. 113-131, 2016.
- [8] M. C. González, J. T. Pérez y M. D. S. A. Rojas, “Tumores melanocíticos. Clasificación. Formas benignas. Melanoma. Etiopatogenia. Clínica. Diagnóstico. Aspectos terapéuticos”, *Medicine: Programa de Formación Médica Continuada Acreditado*, vol. 9, n.º 48, págs. 3123-3129, 2006.
- [9] B. C. Bastian, “The molecular pathology of melanoma: an integrated taxonomy of melanocytic neoplasia”, 2014.

- [10] Instituto Nacional de Cáncer, <https://www.cancer.gov/espanol/publicaciones/diccionario/def/melanocito>.
- [11] T. Wiesner, H. Kutzner, L. Cerroni, M. C. Mihm Jr, K. J. Busam y R. Murali, “Genomic aberrations in spitzoid melanocytic tumours and their implications for diagnosis, prognosis and therapy”, *Pathology*, vol. 48, n.º 2, págs. 113-131, 2016.
- [12] R. L. Barnhill, “The Spitzoid lesion: rethinking Spitz tumors, atypical variants, ‘Spitzoid melanoma’ and risk assessment”, *Modern pathology*, vol. 19, n.º 2, S21-S33, 2006.
- [13] A. Lallas, A. Kyrgidis, G. Ferrara, H. Kittler, Z. Apalla, F. Castagnetti, C. Longo, E. Moscarella, S. Piana, I. Zalaudek y col., “Atypical Spitz tumours and sentinel lymph node biopsy: a systematic review”, *The lancet oncology*, vol. 15, n.º 4, e178-e183, 2014.
- [14] J. G. García Pardo, “Diseño y desarrollo de un sistema automático de clasificación de estructuras glandulares en imágenes histológicas de próstata”, 2018.
- [15] R. Murali, R. N. Sharma, J. F. Thompson, J. R. Stretch, C. S. Lee, S. W. McCarthy y R. A. Scolyer, “Sentinel lymph node biopsy in histologically ambiguous melanocytic tumors with spitzoid features (so-called atypical spitzoid tumors)”, *Annals of Surgical Oncology*, vol. 15, n.º 1, págs. 302-309, 2008.
- [16] P. Gerami, S. S. Jewell, L. E. Morrison, B. Blondin, J. Schulz, T. Ruffalo, P. Matushek IV, M. Legator, K. Jacobson, S. R. Dalton y col., “Fluorescence in situ hybridization (FISH) as an ancillary diagnostic tool in the diagnosis of melanoma”, *The American journal of surgical pathology*, vol. 33, n.º 8, págs. 1146-1156, 2009.
- [17] K. Conway, S. N. Edmiston, J. S. Parker, P. F. Kuan, Y.-H. Tsai, P. A. Groben, D. C. Zedek, G. A. Scott, E. A. Parrish, H. Hao y col., “Identification of a Robust Methylation Classifier for Cutaneous Melanoma Diagnosis”, *Journal of Investigative Dermatology*, vol. 139, n.º 6, págs. 1349-1361, 2019.
- [18] B. Sánchez-Sendra, C. Martínez-Ciarpaglini, J. F. González-Muñoz, A. Murgui, L. Terrádez y C. Monteagudo, “Downregulation of intratumoral expression of miR-205, miR-200c and miR-125b in primary human cutaneous melanomas predicts shorter survival”, *Scientific reports*, vol. 8, n.º 1, págs. 1-14, 2018.
- [19] D. Pradhan, G. Jour, D. Milton, V. Vasudevaraja, M. T. Tetzlaff, P. Nagarajan, J. L. Curry, D. Ivan, L. Long, Y. Ding y col., “Aberrant DNA Methylation Predicts Melanoma-Specific Survival in Patients with Acral Melanoma”, *Cancers*, vol. 11, n.º 12, pág. 2031, 2019.
- [20] M. Kulis y M. Esteller, “DNA methylation and cancer”, en *Advances in genetics*, vol. 70, Elsevier, 2010, págs. 27-56.

- [21] L. Zhang, J. Meng, H. Liu e Y. Huang, “A nonparametric Bayesian approach for clustering bisulfate-based DNA methylation profiles”, *BMC genomics*, vol. 13, n.º S6, S20, 2012.
- [22] M. L. Mansego, F. I. Milagro, J. Campión y J. A. Martínez, “Techniques of DNA methylation analysis with nutritional applications”, *Lifestyle Genomics*, vol. 6, n.º 2, págs. 83-96, 2013.
- [23] H. Gu, Z. D. Smith, C. Bock, P. Boyle, A. Gnirke y A. Meissner, “Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling”, *Nature protocols*, vol. 6, n.º 4, pág. 468, 2011.
- [24] A. Chatterjee, E. J. Rodger, P. A. Stockwell, R. J. Weeks e I. M. Morison, “Technical considerations for reduced representation bisulfite sequencing with multiplexed libraries”, *BioMed Research International*, vol. 2012, 2012.
- [25] S. R. Choudhury y B. A. Walker, “A Rapid and Robust Protocol for Reduced Representation Bisulfite Sequencing in Multiple Myeloma”, en *Multiple Myeloma*, Springer, 2018, págs. 179-191.
- [26] C. Schinke, Y. Mo, Y. Yu, K. Amiri, J. Sosman, J. Greally y A. Verma, “Aberrant DNA methylation in malignant melanoma”, *Melanoma research*, vol. 20, n.º 4, pág. 253, 2010.
- [27] J. M. Estanyola y O. Bachsb, “Nuevos métodos de diagnóstico molecular”.
- [28] G. Micevic, N. Theodosakis y M. Bosenberg, “Aberrant DNA methylation in melanoma: biomarker and therapeutic opportunities”, *Clinical epigenetics*, vol. 9, n.º 1, pág. 34, 2017.
- [29] S.-G. Jin, W. Xiong, X. Wu, L. Yang y G. P. Pfeifer, “The DNA methylation landscape of human melanoma”, *Genomics*, vol. 106, n.º 6, págs. 322-330, 2015.
- [30] MicroDraw, <http://microdraw.pasteur.fr/>.
- [31] Python, <https://www.python.org/doc/essays/blurb/>.
- [32] NVIDIA cuDNN, <https://developer.nvidia.com/cudnn>.
- [33] MobaXterm, <https://mobaxterm.mobatek.net/>.
- [34] MathWorks, https://es.mathworks.com/help/matlab/learn_matlab/product-description.html.
- [35] J. Martorell-Marugán, S. Tabik, Y. Benhammou, C. del Val, I. Zwir, F. Herrera y P. Carmona-Sáez, “Deep Learning in Omics Data Analysis and Precision Medicine”, en *Computational Biology [Internet]*, Codon Publications, 2019.

- [36] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson y col., “High density DNA methylation array with single CpG site resolution”, *Genomics*, vol. 98, n.º 4, págs. 288-295, 2011.
- [37] P. W. Laird, “Principles and challenges of genome-wide DNA methylation analysis”, *Nature Reviews Genetics*, vol. 11, n.º 3, págs. 191-203, 2010.
- [38] N Yuvaraj y P Vivekanandan, “An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data”, en *2013 International Conference on Information Communication and Embedded Systems (Icices)*, IEEE, 2013, págs. 761-768.
- [39] N. Jazayeri y H. Sajedi, “Breast cancer diagnosis based on genomic data and extreme learning machine”, *SN Applied Sciences*, vol. 2, n.º 1, pág. 3, 2020.
- [40] Z. Si, H. Yu y Z. Ma, “Learning deep features for dna methylation data analysis”, *IEEE Access*, vol. 4, págs. 2732-2737, 2016.
- [41] M. Khwaja, M. Kalofonou y C. Toumazou, “A Deep Autoencoder System for Differentiation of Cancer Types Based on DNA Methylation State”, *arXiv preprint arXiv:1810.01243*, 2018.
- [42] A. J. Titus, O. M. Wilkins, C. A. Bobak y B. C. Christensen, “Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction”, *bioRxiv*, pág. 433 763, 2018.
- [43] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui y J. Long, “A survey of clustering with deep learning: From the perspective of network architecture”, *IEEE Access*, vol. 6, págs. 39 501-39 514, 2018.
- [44] J. Xie, R. Girshick y A. Farhadi, “Unsupervised deep embedding for clustering analysis”, en *International conference on machine learning*, 2016, págs. 478-487.
- [45] D. Foster, *Generative deep learning: teaching machines to paint, write, compose, and play*. O’Reilly Media, 2019.
- [46] J. Rocca, *Understanding Variational Autoencoders (VAEs)*, <https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73>.
- [47] N. V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique”, *Journal of artificial intelligence research*, vol. 16, págs. 321-357, 2002.
- [48] D. Svozil, V. Kvasnicka y J. Pospichal, “Introduction to multi-layer feed-forward neural networks”, *Chemometrics and intelligent laboratory systems*, vol. 39, n.º 1, págs. 43-62, 1997.
- [49] J. R. Hershey, Z. Chen, J. Le Roux y S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation”, en *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, págs. 31-35.

- [50] X. Guo, X. Liu, E. Zhu y J. Yin, “Deep clustering with convolutional autoencoders”, en *International conference on neural information processing*, Springer, 2017, págs. 373-382.
- [51] X. Guo, E. Zhu, X. Liu y J. Yin, “Deep embedded clustering with data augmentation”, en *Asian conference on machine learning*, 2018, págs. 550-565.
- [52] J. Enguehard, P. O’Halloran y A. Gholipour, “Semi-supervised learning with deep embedded clustering for image classification and segmentation”, *IEEE Access*, vol. 7, págs. 11 093-11 104, 2019.
- [53] B. H. Prasetio, H. Tamura y K. Tanno, “A Deep Time-delay Embedded Algorithm for Unsupervised Stress Speech Clustering”, en *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, IEEE, 2019, págs. 1193-1198.
- [54] T. Tian, J. Wan, Q. Song y Z. Wei, “Clustering single-cell RNA-seq data with a model-based deep learning approach”, *Nature Machine Intelligence*, vol. 1, n.º 4, págs. 191-198, 2019.
- [55] A. Hekler, J. S. Utikal, A. H. Enk, C. Berking, J. Klode, D. Schadendorf, P. Jansen, C. Franklin, T. Holland-Letz, D. Krahl y col., “Pathologist-level classification of histopathological melanoma images with deep neural networks”, *European Journal of Cancer*, vol. 115, págs. 79-83, 2019.
- [56] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol y col., “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer”, *Jama*, vol. 318, n.º 22, págs. 2199-2210, 2017.
- [57] A. Rakhlin, A. Shvets, V. Iglovikov y A. A. Kalinin, “Deep convolutional neural networks for breast cancer histology image analysis”, en *International Conference Image Analysis and Recognition*, Springer, 2018, págs. 737-744.
- [58] Y. Liu, T. Kohlberger, M. Norouzi, G. E. Dahl, J. L. Smith, A. Mohtashamian, N. Olson, L. H. Peng, J. D. Hipp y M. C. Stumpe, “Artificial intelligence–based breast cancer nodal metastasis detection: Insights into the black box for pathologists”, *Archives of pathology & laboratory medicine*, vol. 143, n.º 7, págs. 859-868, 2019.
- [59] M. A. S. R. M. V. N. Julio Silva-Rodríguez Adrián Colomer, “Going Deeper through the Gleason Scoring Scale: An Automatic end-to-end System for Histology Prostate Grading and Cribriform Pattern Detection”, *Computer Methods and Programs in Biomedicine*, 2020.
- [60] 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, 2019.
- [61] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin y M. Snyder, “Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features”, *Nature communications*, vol. 7, pág. 12 474, 2016.

- [62] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian y A. Tsirigos, “Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning”, *Nature medicine*, vol. 24, n.º 10, págs. 1559-1567, 2018.
- [63] X. Wang, A. Janowczyk, Y. Zhou, R. Thawani, P. Fu, K. Schalper, V. Velcheti y A. Madabhushi, “Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images”, *Scientific reports*, vol. 7, n.º 1, págs. 1-10, 2017.
- [64] M. Khoshdeli, R. Cong y B. Parvin, “Detection of nuclei in H&E stained sections using convolutional neural networks”, en *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, IEEE, 2017, págs. 105-108.
- [65] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway y J. Liang, “Convolutional neural networks for medical image analysis: Full training or fine tuning?”, *IEEE transactions on medical imaging*, vol. 35, n.º 5, págs. 1299-1312, 2016.
- [66] A. Krizhevsky, I. Sutskever y G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, en *Advances in neural information processing systems*, 2012, págs. 1097-1105.
- [67] K. Simonyan y A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, *arXiv preprint arXiv:1409.1556*, 2014.
- [68] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke y A. Rabinovich, “Going deeper with convolutions”, en *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, págs. 1-9.
- [69] K. He, X. Zhang, S. Ren y J. Sun, “Deep residual learning for image recognition”, en *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, págs. 770-778.
- [70] Z. Zhao, K. Zhang, X. Hao, J. Tian, M. C. H. Chua, L. Chen y X. Xu, “BiRA-Net: Bilinear Attention Net for Diabetic Retinopathy Grading”, en *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, págs. 1385-1389.
- [71] J. Hu, L. Shen y G. Sun, “Squeeze-and-excitation networks”, en *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, págs. 7132-7141.
- [72] Y. Bai, P. Zille, W. Hu, V. D. Calhoun e Y.-P. Wang, “Biomarker Identification Through Integrating fMRI and Epigenetics”, *IEEE Transactions on Biomedical Engineering*, 2019.
- [73] J. Snell, K. Swersky y R. Zemel, “Prototypical networks for few-shot learning”, en *Advances in neural information processing systems*, 2017, págs. 4077-4087.

- [74] I. Khoury, R. M. El-Mawas, O. El-Rawas, E. F. Mounayar y H. Artail, “An efficient web page change detection system based on an optimized Hungarian algorithm”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, n.º 5, págs. 599-613, 2007.
- [75] L. v. d. Maaten y G. Hinton, “Visualizing data using t-SNE”, *Journal of machine learning research*, vol. 9, n.º Nov, págs. 2579-2605, 2008.
- [76] A. Belaala, L. S. Terrissa, N. Zerhouni y C. Devalland, “Spitzoid Lesions Diagnosis based on GA feature selection and Random Forest”,

Parte II

Presupuesto.

Capítulo 1

Presupuesto

Índice de contenidos

| | |
|--------------------------------------|----|
| 1.1 Objetivo | 90 |
| 1.2 Presupuestos parciales | 90 |
| 1.2.1 Costes de personal | 90 |
| 1.2.2 Costes de hardware | 90 |
| 1.2.3 Costes de software | 92 |
| 1.3 Presupuesto Total | 92 |

1.1 Objetivo

El objetivo de esta sección es aportar una valoración económica de este proyecto basado en el desarrollo de sistemas de automáticos de clasificación de tumores melanocíticos spitzoides.

1.2 Presupuestos parciales

Con el objetivo de cuantificar los presupuestos parciales, estos se han desglosado en: (i) coste de personal, (ii) coste de *software* y (iii) coste de *hardware*. Este desglose ofrece la posibilidad de realizar de forma rápida un recálculo del coste total si alguna condición varía.

1.2.1 Costes de personal

En este apartado se tienen en cuenta los recursos humanos necesarios para el desarrollo del presente proyecto.

En la Tabla 1.1 se desglosan los costes de mano de obra, mostrando la remuneración de cada uno de los participantes en base al tiempo dedicado al proyecto. Concretamente, en este proyecto han participado:

- D. José Manuel Mossi García, profesor titular de universidad y tutor del presente trabajo.
- D. Adrián Colomer Granero, doctor en telecomunicación y encargado de cotutorizar el proyecto.
- D^a Rocío del Amor del Amor, estudiante del Máster de Ingeniería Biomédica y autora del proyecto.

| Descripción | Uds. | Cantidad | Precio unitario (€/h) | Coste imputable (€) |
|------------------------|------|----------|-----------------------|---------------------|
| Titular de universidad | h | 35 | 29,50 | 1.032,50 |
| Doctor | h | 38 | 20,50 | 779,00 |
| Estudiante | h | 600 | 13,50 | 8.280,00 |
| TOTAL | | | | 10.091,50 € |

Tabla 1.1: Descomposición de los costes de personal.

1.2.2 Costes de hardware

Como se ha explicado en secciones anteriores, la mayor parte del proyecto se ha desarrollado utilizando un único ordenador. Sin embargo, dado que el entrenamiento de los algoritmos desarrollados requiere una potencia computacional elevada, este se ha llevado a cabo en uno de los servidores de computación del grupo de investigación CVB Lab.

Se han utilizado tres enfoques diferentes, uno para los datos de metilación, otro para las imágenes histopatológicas y un último enfoque que combina las características extraídas de las imágenes histopatológicas con los datos de metilación. Puesto que se han realizado varios experimentos hasta encontrar los mejores hiperparámetros para cada uno de los enfoques citados, se puede concluir que se ha hecho uso de este servidor aproximadamente tres meses. Además, durante este tiempo se ha hecho uso de un servidor de datos para el almacenamiento de todas las imágenes utilizadas.

En la Tabla 1.2 se calcula el coste que ha supuesto el uso del equipo personal, mientras que en la Tabla 1.3 se muestra el coste del uso de cada uno de los componentes de los servidores (de computación y de datos) pertenecientes a CVBLab. Finalmente, en la Tabla 1.4 se calcula el coste total en concepto del *hardware* utilizado.

| Descripción | Cantidad | Coste unitario (€) sin IVA | Periodo de amortización (meses) | Intervalo amortizado (meses) | Coste imputable (€) sin IVA |
|--|----------|----------------------------|---------------------------------|------------------------------|-----------------------------|
| ASUS M51AC-IT012S intel Core i7-4770, 3.4 GHz, 8 GB RAM, 32 GB, nVIDIA GeForce GTX 760" | 1 | 882,15 | 72 | 8 | 98,02 |
| TOTAL | | | | | 98,02 € |

Tabla 1.2: Coste del uso del equipo personal.

| Descripción | Cantidad | Coste unitario (€) sin IVA | Periodo de amortización (meses) | Intervalo amortizado (meses) | Coste imputable (€) sin IVA |
|--------------------------------|----------|----------------------------|---------------------------------|------------------------------|-----------------------------|
| procesador Intel i7 @4.20GHz | 1 u | 344,00 | 48 | 3 | 21,50 |
| Tarjeta gráfica NVIDIA Titan V | 1 u | 3300,00 | 48 | 3 | 206,25 |
| NAS Synology DS918 | 1 u | 77,00 | 48 | 3 | 4,81 |
| TOTAL | | | | | 232,56 € |

Tabla 1.3: Coste del uso de los servidor del grupo CVBLAB.

| Descripción | Coste imputable (€) sin IVA |
|--|-----------------------------|
| ASUS M51AC-IT012S intel Core i7-4770, 3.4 GHz, 8 GB RAM, 32 GB, nVIDIA GeForce GTX 760" | 98,02 |
| Servidor computación CVBLab | 232,56 |
| TOTAL | |
| 330,58 € | |

Tabla 1.4: Descomposición de los costes de hardware.

1.2.3 Costes de software

A continuación se exponen los costes necesarios para la obtención de las licencias de software imprescindibles para el desarrollo del proyecto. El ordenador utilizado incluye la licencia del sistema operativo Windows 10, así como sus posteriores actualizaciones. Por otro lado, la redacción del documento se ha llevado a cabo mediante el editor de texto online *www.overleaf.com*, de uso gratuito. Las librerías *keras* utilizadas también son gratuitas debido a que son de código abierto. Si que ha supuesto un coste la utilización de la versión de Matlab empleada para el desarrollo de este proyecto. En la Tabla 1.5 se puede observar el coste del software utilizado.

| Descripción | Cantidad | Coste (€) sin IVA | Periodo de amortización (meses) | Intervalo de uso (meses) | Coste imputable (€) sin IVA |
|-----------------|----------|-------------------|---------------------------------|--------------------------|-----------------------------|
| Overleaf | 1 | 0 | 12 | 8 | 0 |
| Matlab R2018a | 1 | 800 | 12 | 8 | 533,30 |
| Librerías Keras | 1 | 0 | 12 | 3 | 0 |
| TOTAL | | | | | 533,30 € |

Tabla 1.5: Descomposición de los costes de software. Todos los precios corresponden a las licencias de uso Académico.

1.3 Presupuesto Total

Una vez calculado el coste relativo al personal, hardware y software se computa el presupuesto de Ejecución de Material (PEM) que es la suma de los costes anteriormente citados. Esta operación puede verse ejemplificada en la Tabla 1.6.

| Descripción | Coste (€) |
|------------------------|------------------|
| Costes de mano de obra | 10.0091,50 |
| Costes de hardware | 330,58 |
| Costes de software | 533,30 |
| Suma (PEM) | 10.955,38 |

Tabla 1.6: Presupuesto de Ejecución de Material del proyecto.

Posteriormente, se calcula el importe total que supondría el desarrollo de este Trabajo Fin de Máster incorporando a los costes de Ejecución de Material la parte proporcional a los costes generales y al beneficio industrial. Además, se incluyen los impuestos y otros gastos. Concretamente, se añade un 13% correspondiente a los gastos generales y un 6% en concepto del beneficio industrial. Por último, a este resultado se le añade un 21% adicional correspondiente al Impuesto sobre el Valor Añadido (IVA). El importe total se detalla en la Tabla 1.7.

| Descripción | Coste (€) |
|--------------------------------------|--------------------|
| Presupuesto de Ejecución de Material | 10.955,38 |
| Gastos generales | 1.424,20 |
| Beneficio industrial | 657,32 |
| Suma | 13.036,90 |
| IVA (21 %) | 2.737,75 |
| PRESUPUESTO TOTAL | 15.774,65 € |

Tabla 1.7: Presupuesto total del presente proyecto.

Por tanto, el coste total de la realización de este proyecto asciende a **QUINCE MIL SETECIENTOS SETENTA Y CUATRO EUROS CON SESENTA Y CINCO CÉNTIMOS.**