# ESCUELA DE INGENIERÍA DE TELECOMUNICACIÓN Y ELECTRÓNICA

# TRABAJO FIN DE GRADO

## Estudio de la vascularización cerebral mediante el uso de imágenes hiperespectrales

**Titulación: Grado en Ingeniería en Tecnologías de la Telecomunicación**

**Autor:** Kenya Espino Gutiérrez

**Tutor:** Dr. D. Gustavo Iván Marrero Callicó

**Cotutor:** Dr. D. Himar A. Fabelo Gómez

**Cotutora**: Dña. Laura Quintana Quintana

**Fecha:** Septiembre 2022

# Índice

# Índice de figuras

# Índice de tablas

# Agradecimientos

Quiero agradecer a todos los profesores que he tenido, tanto en la universidad como fuera de ella, por haber fomentado el desarrollo de mi curiosidad y haberme formado en su gran medida para ser la persona que soy ahora. También quiero agradecer a aquellos que han ayudado a definirme como profesional y como persona, realizando un trabajo que nunca se podrá valorar lo suficiente.

A mis tutores de este trabajo, agradecerles por su infinita paciencia y comprensión ante todas las situaciones en las que me he ido encontrando a lo largo de este periodo, por hacer fácil y llevadero cada momento, aunque no fueran los mejores, nunca podré agradecéroslo lo suficiente.

También a mi familia, que me ha dedicado todo su tiempo, todo su esfuerzo, cariño y todos sus recursos para educarme y formarme lo mejor posible para afrontar la vida.

A mi pareja, por ayudarme a salir de todos y cada uno de los momentos con una sonrisa y hacerme ver el lado bueno de las cosas.

A mis compañeros de trabajo que siempre intentaron hacer las cosas fáciles para que pudiera llegar aquí.

A mis compañeros y amigos, por siempre arrimar el hombro, por saber animarme, darme la mano y motivarme cuando lo necesitaba. Gracias por terminar conmigo trabajos, exámenes y noches eternas. En especial agradecer a Andrea, Sara, Roberto, Bryan, Fernando y Kilian, por ser la familia universitaria que llegó para quedarse, no podría haber encontrado nada mejor.

A las personas que me llenan que no me caben en este papel y a mi querida mujer de verde, gracias.

# Resumen

Los tumores existen y progresan generalmente en ambientes con falta de oxígeno, siendo esta un factor determinante en la agresividad e invasión que desarrollará el tumor. La hipoxia se correlaciona de la misma manera con un aumento de la angiogénesis mediante la activación del factor inducible por hipoxia. En condiciones de falta de oxigenación en el tumor se produce la activación de algunos genes transcriptores que aumentan la agresividad del tumor, activando en el proceso factores de crecimiento de la angiogénesis. La hipoxia y la angiogénesis juegan un papel fundamental en el crecimiento de los tumores y su proliferación por el organismo dando lugar a la metástasis. Varios estudios relacionan la malignidad del tumor, su crecimiento, su desarrollo, su resistencia a fármacos y su mortalidad con ambientes de hipoxia y vascularización.

Conseguir diagnosticar a tiempo estos ambientes con falta de oxígeno y exceso de vasos sanguíneos capaces de fomentar el crecimiento de los tumores, se hace primordial. Los principales avances en el campo de la imagen diagnóstica se centraban en la mejora de la resolución de la imagen y en la introducción de nuevas técnicas de contraste. No obstante, la tecnología de imágenes evoluciona y aparecen en el diagnóstico del cáncer y sus patologías asociadas las imágenes hiperespectrales (HS), caracterizadas por recopilar grandes conjuntos de datos y generar cubos HS que aportan información relevante sobre los elementos que componen la imagen a través de su firma espectral.

Este trabajo de final de grado tiene como objetivo el estudio y análisis de imágenes HS de origen médico con el fin de discriminar entre tejido relacionado al tumor y tejido asociado a vaso sanguíneo a través de un modelo definido en el trabajo basado en el algoritmo de clasificación supervisado *Support Vector Machine* (SVM); y la distinción dentro de este último tejido entre arterias y venas, a través de un modelo definido mediante el algoritmo de segmentación no supervisado *K-means* para determinar si un tumor puede o no estar recibiendo oxígeno.

Tras evaluar los resultados se obtienen una serie de valores que representan las longitudes de onda características para los tejidos tumor y vaso sanguíneo y se estipula que existen nuevas líneas de avances en esta investigación. Además, se concluye a través del modelo no supervisado el algoritmo con mejores resultados y se muestran imágenes adquiridas en el estudio de este modelo de discriminación de vasos y arterias. Finalmente, se encuentra que, aunque el método no supervisado propuesto aporta información relevante, no es posible garantizar la clasificación entre arterias y venas sin supervisión médica y un estudio más exhaustivo caso a caso.

# Abstract

Tumors generally exist and progress in oxygen-deprived environments, which is a determining factor in the aggressiveness and invasiveness of the tumor. Hypoxia correlates in the same way with an increase in angiogenesis through the activation of hypoxia-inducible factor. Under conditions of lack of oxygenation in the tumor there is activation of some transcriptional genes that increase the aggressiveness of the tumor, in the process activating angiogenesis growth factors. Hypoxia and angiogenesis play a fundamental role in the growth of tumors and their proliferation throughout the body, leading to metastasis. Several studies link tumor malignancy, growth, development, drug resistance and mortality to hypoxic and vascularized environments.

Early diagnosis of these environments of oxygen and excess blood vessels capable of promoting tumor growth is of fundamental importance. Major advances in diagnostic imaging have focused on improving image resolution and introducing new contrast techniques. However, imaging technology is evolving, and hyperspectral (HS) images are appearing in the diagnosis of cancer and its associated pathologies. Characterized by the collection of large data sets and the generation of HS cubes that provide relevant information about the elements that make up the image through their spectral signature.

The aim of this final degree work is to study and analyze HS images of medical origin in order to discriminate between tumor-related tissue and tissue associated with blood vessels using a model defined in the work based on the supervised support vector machine (SVM) algorithm; and to distinguish within the latter tissue between arteries and veins, using a model defined using the unsupervised K-means algorithm to determine whether or not a tumor may or may not be receiving oxygen.

After evaluating the results, a series of values representing the characteristics wavelengths of tumor tissues and blood vessels are obtained, and it is stipulated that her are new lines of progress proposed. Furthermore, it is concluded which algorithm works better through the unsupervised model and the images acquired in the study of this vein and artery discrimination model are shown. Finally, it is found that although the proposed unsupervised method provides relevant information, it is not possible to guarantee the classification between arteries and veins without medical supervision and a more exhaustive study of each case.

# Capítulo I: Introducción

# Motivación

La principal motivación para llevar a cabo este trabajo de fin de grado (TFG) es el marco de tecnología aplicada a la medicina en el que se encuentra. Las telecomunicaciones tienen muchas verticales tecnológicas por explorar y el campo médico es uno que aporta valor social y pretende realizar mejoras tanto en la vida de los pacientes como en la de los médicos. El punto de partida para este TFG es la motivación surgida a partir del conocimiento de la existencia del proyecto HELICoiD, llevado a cabo por el grupo de investigadores procedentes del IUMA, cuyo objetivo fue el de detectar y definir perfiles de tumores cerebrales en tiempo real mediante el uso de imágenes HS, proyecto del que se extrae la base de datos de este trabajo.

Este trabajo plantea los problemas que surgen en la detección de tumores al querer pronosticar de forma no invasiva a través de imágenes HS las características que tienen estas lesiones y como proceder en su tratamiento. Concretamente los problemas a los que se enfrenta este trabajo son las patologías asociadas al cáncer, como son los escenarios de hipoxia y angiogénesis en tumores. Es por ello por lo que se plantea como posible solución la definición de un modelo que permita la clasificación y la determinación de los tejidos asociados a tumor y vaso sanguíneo, permitiendo conocer el grado de un tumor a través de la información recopilada y el estado de angiogénesis en el que se encuentra. Por otro lado, se trata de definir un modelo que permita conocer la cantidad de arterias y venas que conforman las lesiones, puesto que esta información podrá representar el grado de oxigenación al que se enfrenta el tumor. Ambas informaciones son primordiales a la hora de realizar un diagnóstico, puesto que constituyen factores que aumentan la peligrosidad del cáncer. A partir de estos problemas y queriendo aportar algunas soluciones, surge este trabajo que parte inicialmente de la base de datos de HELICoiD y del desarrollo por parte de la alumna de modelos de clasificación ajustados para esta aplicación en concreto.

# Objetivo

El objetivo general del presente Trabajo de Fin de Grado (TFG) consiste en el estudio e identificación de la vascularización tumoral mediante el uso de Imagen Hiperespectral (HSI), especialmente en condiciones intraoperatorias en cirugías neuroquirúrgicas. Se realizará una investigación donde se determinen las características más relevantes que permitan realizar una identificación entre los tejidos asociados a tumor y vascularización. Dentro de este último tejido

se buscará distinguir entre venas y arterias, ya que son las encargadas de transportar el oxígeno en el cuerpo y, por ende, las que permitirán saber a través de su discriminación si es posible conocer el estado de oxigenación del tumor. Los datos empleados para la realización de este trabajo son las imágenes tomadas con las cámaras HS del proyecto HELICoiD. Este objetivo general se divide en los siguientes objetivos específicos:

- Estudio del estado del arte en tecnologías de imagen médica aplicada a la detección de vascularización cerebral.
- Estudio de la base de datos de los proyectos asociados a este trabajo.
- Estudio de los algoritmos que permitan realizar la identificación de la vascularización tumoral basada en imágenes HS.
- Aplicación de algoritmos de inteligencia artificial supervisada para la discriminación entre tumor y vaso sanguíneo.
- Aplicación de algoritmos no supervisados para la discriminación del tejido vascular entre venas y arterias.
- Validación de los modelos de discriminación propuestos.

La investigación y desarrollo de sistemas de clasificación mediante imágenes HS es una de las principales líneas de trabajo de la división de *Diseño de Sistemas Integrados* (*DSI*) perteneciente al *Instituto Universitario de Microelectrónica Aplicada* (*IUMA*) de la *Universidad de Las Palmas de Gran Canaria* (*ULPGC*). Con el desarrollo del presente proyecto se explora este emergente campo de investigación, centrado en el estudio de la identificación de los distintos tejidos que podemos encontrar en casos de vascularización de tumores mediante el uso de imágenes HS.

## Organización del documento

Este trabajo está estructurado de la siguiente manera:

**Capítulo 1**: El primer capítulo de este documento consiste en una breve introducción al trabajo de investigación y de los objetivos trazados.

**Capítulo 2**: En él se presenta una contextualización del trabajo. En el estado del arte encontraremos primeramente información médica sobre los tumores, la vascularización. Y contextualización referida a la parte técnica del trabajo basada en el contexto de imágenes HS y su relación con el cáncer.

**Capítulo 3**: Se presenta los diferentes tipos de *algoritmos* utilizados por los clasificadores empleados en el trabajo y en la literatura.

**Capítulo 4:** Se describen los materiales de los que parte la investigación del trabajo y las herramientas que se han necesitado para la elaboración del mismo. En este capítulo se presenta la metodología llevada a cabo para el análisis e identificación de los casos de estudio que se proponía en este proyecto para identificar la distinción de tejidos y la perteneciente a la vascularización.

**Capítulo 5:** Se muestra los resultados obtenidos en el trabajo.

**Capítulo 6:** El último capítulo redactado resume las conclusiones obtenidas durante el desarrollo de este trabajo de investigación y las líneas futuras que pueden seguirse a partir de él.

**Capítulo 7:** Presenta las referencias bibliográficas usadas durante todo el desarrollo del presente trabajo.

# Capítulo 2: Contextualización

## 2.1. Introducción

A lo largo de este presente capítulo se realizará una descripción de los conceptos teóricos médicos y tecnológicos en los que se basará la investigación de este TFG con el fin de poder situar y contextualizar el trabajo realizado. Por ello, en este apartado se realizará una definición, breve introducción y resumen del estudio inicial que se tomará como punto de partida acerca del tratado de imágenes utilizadas para diagnosticar cáncer y su correspondiente vascularización y oxigenación en tumores. Además, es necesario introducir conceptos teóricos que se deberán tener en cuenta para poder realizar el trabajo planteado para este documento, en el cual, se usan casos médicos reales de tumores, los cuales se examinarán y analizarán a lo largo del proceso que seguirá este proyecto.

Para empezar, se presenta una breve descripción de los tumores, tipos existentes y como se ha realizado con anterioridad distintos procesos relacionados con el tratado de imágenes para su diagnosis. Con el fin de conocer la información necesaria y lo más detallada posible, se procederá a definir la explicación de vascularización y su relación en tumores con los conceptos previamente definidos a lo largo del capítulo.

Finalmente, el capítulo acabará con la definición y contextualización de la tecnología de imágenes HS aplicada al diagnóstico del cáncer y un último apartado que explicará los métodos de clasificación de imagen teóricos. Gracias a este capítulo tendremos el contexto teórico y el punto de partida para el desarrollo de todo el proyecto.

## 2.2. Tumor y cáncer

### 2.2.1. Definición

El término tumor suele referirse a cualquier masa de células agrupadas que presentan tejido inapropiado en un grado anormal de multiplicación. Es decir, los tumores son una masa de tejido creado por células que crecen y se replican incontrolablemente sin seguiré el proceso de muerte celular programada. Por lo tanto, cada capa de células producidas se acumulará en la capa anterior, de modo que este organismo anormal crecerá gradualmente [1].

Los tumores pueden ser malignos (cancerosos) o benignos (no cancerosos). Según la forma y suministro de sangre podemos diferenciarlos entre ellos. Los tumores malignos se infiltran

localmente y poseen una forma irregular con bordes diferentes y cubierto por vasos sanguíneos por los que se multiplican y establecen crecimientos secundarios en otras partes del cuerpo. Por otro lado, el tumor considerado benigno se agranda localmente y a menudo está encerrado en alguna zona del cuerpo, pero no invaden tejidos ni se propagan a distancia, este tipo suele tener una forma esférica sin irregularidades llamativas y que se enrojece por vasos sanguíneos de menor tamaño que no son preocupantes en diámetro [2], [3].

Los tumores son, por tanto, el resultado que se produce por una acumulación de células excesiva producida por una patología en la división celular que no consigue alcanzar el equilibrio entre las células muertas y las que deberían componen el organismo. El sistema inmunológico resulta importante en este tipo de situaciones puesto que los pacientes con problemas inmunosuprimidos contarán con peores condiciones al tener un menor control y respuesta en su organismo [2].

El desglose de tumores en malignos nos lleva a incorporar el término cáncer. La Organización Mundial de la Salud (OMS) define la palabra cáncer como la descripción genérica utilizada para designar un amplio grupo de enfermedades que pueden afectar a cualquier parte del cuerpo. Además, la OMS apunta que una de las principales características del cáncer es la multiplicación de células anormales de forma rápida y más allá de los límites establecidos, invadiendo las partes adyacentes del organismo o consiguiendo propagarse a otros órganos, llegando al estado que se conoce como metástasis. El cáncer no tiene un origen determinado, cada patrón es diferente y cuenta con factores de estimulación diversos.

Son diversas las causas que pueden provocar el origen y desarrollo del cáncer. Puede influir en el organismo todo tipo de factores desde la probabilidad genética y las mutaciones asociadas hasta el ritmo de vida y la exposición a agentes externos dañinos para la salud. Uno de los factores importantes es la mutación en genes encargados de funciones vitales que pueden llevar al descontrol en la división celular, reproducción y evasión de la apoptosis[1]. Estas mutaciones se producen debido al crecimiento y reproducción de células que desembocan al reemplazo de células encargadas de funciones importantes, en lugar de seguir con el ciclo natural de muerte celular [3], [4].

[1] Apoptosis: muerte celular programada por el mismo organismo con el fin de mantener el equilibrio entre las células, su crecimiento y desarrollo. Cumple con un propósito fundamental en los organismos al encargarse de eliminar o destruir las células dañadas que pueden producir varios tipos de enfermedades.

Es por ello por lo que algunos genes son más importantes de controlar a la hora de analizar la mutación genética. Entre ellos destacan los oncogenes y genes supresores de tumores. En os oncogenes una anormalidad simple en alguna célula puede desencadenar una mutación dominante. Sin embargo, en los genes supresores de tumores las anomalías deben presentarse en todas las células encargadas de esta función, este tipo se denomina mutación recesiva. En cualquiera de los dos casos, la mutación se activa en los genes y conlleva a una multiplicación incontrolable de células, desembocando en casos de agrupación de células, es decir, masas de tejido anormal, un tumor. A este proceso de descontrol celular y creación de cáncer se le conoce como carcinogénesis, proceso en el que se producen variaciones a nivel celular y molecular que tienen como consecuencia el crecimiento y propagación de células dañadas [4][3].

Existen tres fases identificadas en el desarrollo de la carcinogénesis, se presentan de forma detallada en la Figura 1. Etapas de la carcinogénesis. [4]. La primera fase se conoce como iniciación, en esta fase es cuando se producen los cambios a nivel genético. La siguiente fase se denomina promoción, es la etapa más duradera puesto que posee dos elementos de desarrollo: la inflamación crónica y el estrés oxidativo. Estos componentes son los encargados de expandir las células tumorales y promover la angiogénesis, proceso vital en el desarrollo de tumores. La fase final es conocida como progresión, en ella se lleva a cabo la transformación de las células tumorales en invasivas, llevando células dañadas a otros órganos del organismo y aumentando así el nivel de metástasis.

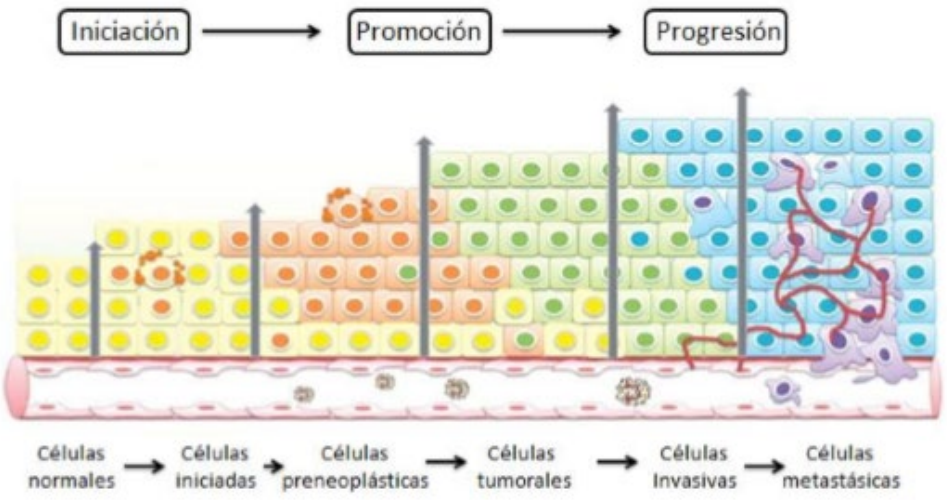



*Figura 1. Etapas de la carcinogénesis.* [4]

Dentro de la lista de lesiones tumorales, el tumor cerebral es el más común dentro del sistema nervioso. Los tumores cerebrales pueden estar causados por diferentes factores celulares, por tanto, son considerados como grupo heterogéneo. Existen dos grupos dentro de este tipo de lesiones: primarias, aquellas que se originan directamente en células que pertenecen al sistema nervioso, y secundarias, son las que tienen su origen en células de otra parte del organismo y que pueden reproducirse en el cerebro como consecuencia de un estado metastásico en el cerebro [5], [6]. Estadísticamente hablando, los tumores primarios más frecuentes son el meningioma y el glioblastoma, en el caso de los tumores secundarios, su origen suele producirse en pulmones, mamas o piel, lo que lleva a que estos tipos de cáncer sean los más agresivos.

Para el caso de los tumores cerebrales primarios existen una clasificación desarrollada por la OMS que clasifica estos en cuatro grados principales La clasificación de estos tumores se realiza para poder conocer un estado de gravedad del tipo de cáncer en el pronóstico del paciente. Separado por grados, se conoce que los grados más bajos de la clasificación I y II, son los que tienen una mayor tasa de supervivencia. Sin embargo, para los grados más altos, III y IV, la peligrosidad aumenta debido a que estadísticamente los casos de éxito son menores. Los tumores cerebrales son una de las principales causas de mortalidad por cáncer, siendo estos de los más aberrantes junto con la leucemia para la edad infantil y el gliobastoma para la edad adulta [4], [7].

En la cirugía, un tumor cerebral será correctamente tratado cuando se elimine la masa de tejido canceroso y un margen de seguridad adicional que evite la reaparición de este mismo tumor en el tiempo. Sin embargo, en el caso de los tumores cerebrales, un exceso margen puede acarrear daños importantes en el paciente, debido a que en el sistema nervioso se encuentran todos los nervios que conectan con funciones vitales, por ello, pueden producirse todo tipo de lesiones. Es por esto por lo que la búsqueda de una herramienta que permita definir con mayor precisión el campo a eliminar y fijar límites de seguridad en operaciones quirúrgicas de este tipo se vuelve primordial en el avance biomédico [8].

### 2.2.2 Tipos de tumores

Tanto a nivel nacional como a nivel mundial, el cáncer es una de las principales causas de mortalidad, en España fue responsable del 22,8% de defunciones en el 2020 [9], [10]. El elevado número de casos de cáncer, así como la cantidad de variantes que estos presentan, han creado la necesidad de buscar parámetros comunes y poder clasificarlos para dar respuesta a los casos

de manera más efectiva. Para dar respuesta a esta necesidad surgen varias clasificaciones según el tejido, el origen primario del tumor o el tipo histológico [11].

Los tipos de tumores pueden clasificarse por tejido y estadio, el tejido determinará el tipo de tumor al que nos referimos, dando por ello nombre particular al cáncer. Sin embargo, el estadio implica el grado de peligrosidad y malignidad en el que se encuentra dicho tumor. Por ello, los tumores pueden estar clasificados inicialmente por el tejido que lo compone y dentro de este mismo tipo tener grados en el estadio del tumor, siguiendo la misma clasificación realizada por la OMS de cuatro grados principales donde los más bajos son los menos peligrosos y los grados más altos lo más mortales. Para categorizar tumores del sistema nervioso según su localización encontramos la distinción de tumores primarios y secundarios (Figura 2). En este TFG las imágenes que sirven como base de datos en el trabajo son pacientes con tumores cerebrales, por ello la clasificación que compete al documento se caracteriza en este tipo de lesión.

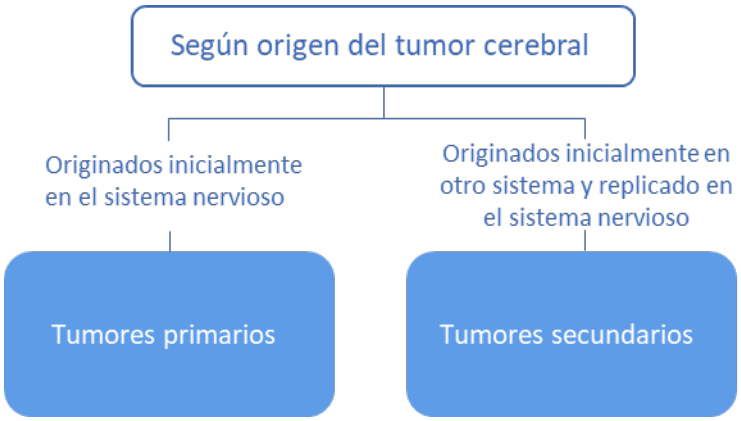



*Figura 2. Clasificación de tumores cerebrales según origen del tumor en el organismo.*

## Tumores Primarios

En 1979, se desarrolló por la OMS un sistema estándar de clasificación para los tumores primarios del sistema nervioso central (TPSNC). Este sistema toma como base de la clasificación el origen del tumor y las características asociadas a su forma o estructura en el pronóstico. Los TPSNC se clasifican entre tumores [12]:

- OMS Grado I: Tumores circunscritos, crecen de forma lenta y cuentan con una baja probabilidad de aumento o transformación a mayor peligrosidad.

- OMS Grado II: Tumores de formas anómalas y difusas, crecen, al igual que los anteriores, de forma lenta. En algunos casos los tumores aumentad su peligrosidad y se transforman en tumores de mayor grado.
- OMS Grado III: Tumores capaces de infiltrase en el organismo, cuentan con células anormales que son capaces de realizar una rápida división celular, extendiendo las células atípicas por el paciente.
- OMS Grado IV: Tumores rápidamente crecientes, poseen una gran velocidad de multiplicación y crecimiento de células cancerosas, en casos avanzados de este grado puede presentarse necrosis y vasos de neoformación

La clasificación llevada a cabo por la OMS presenta que los tumores de grado I y II se definen como "de bajo grado" o "benignos" y tienen un bajo potencial proliferativo, pueden curarse mediante resección quirúrgica. Los tumores de grado II son tumores infiltrantes al igual que los de grado III, pero estos cuentan con baja actividad en la división y propagación celular, aunque en algunos casos, como los gliomas, tienden a progresar a grados superiores (III y IV). Las lesiones OMS II, III y IV, al presentar una morfología difusa suele llevar a que su operación de extracción quede incompleta, necesitando por ello, un seguimiento médico a través de imágenes y, así mismo, en lesiones OMS III y IV ir acompañado de un tratamiento oncológico conocido como radioterapia o quimioterapia. Estos tumores de grado III y IV se denominan tumores de "alto grado" o "malignos" [12].

Afortunadamente en el diagnóstico se conoce que mayormente los tumores suelen ser benignos, aunque estos no pueden ser confirmados histológicamente debido a la dificultad para ver sus características en el cerebro, lo que implica que podría existir un mayor número de casos de incidencia de tumores malignos sin determinar. En los casos conocidos, los tumores malignos inciden mayormente en varones (55%), al contrario que los tumores benignos que suelen encontrarse principalmente en mujeres (64%) [13]

Se encuentra en el sistema nervioso una serie de tumores malignos más frecuentes frente a los distintos tipos existentes. Encabeza la lista de tumores comunes el glioblastoma (3,2 por 100.000 habitantes), seguido por el astrocitoma de grado 3 (0,51 por 100.000 habitantes) y el linfoma (0,43 por 100.000 habitantes). En el caso de tumores benignos, los más encontrados son el meningioma (7,93 por 100.000 habitantes), el adenoma hipofisario (3,65 por 100.000 habitantes) y el schwannoma (1,81 por 100.000 habitantes) [14] Con el fin de entender la

clasificación de tumores primarios que seguirá el desarrollo de este apartado sobre el sistema nervioso se presenta la siguiente Figura 3.

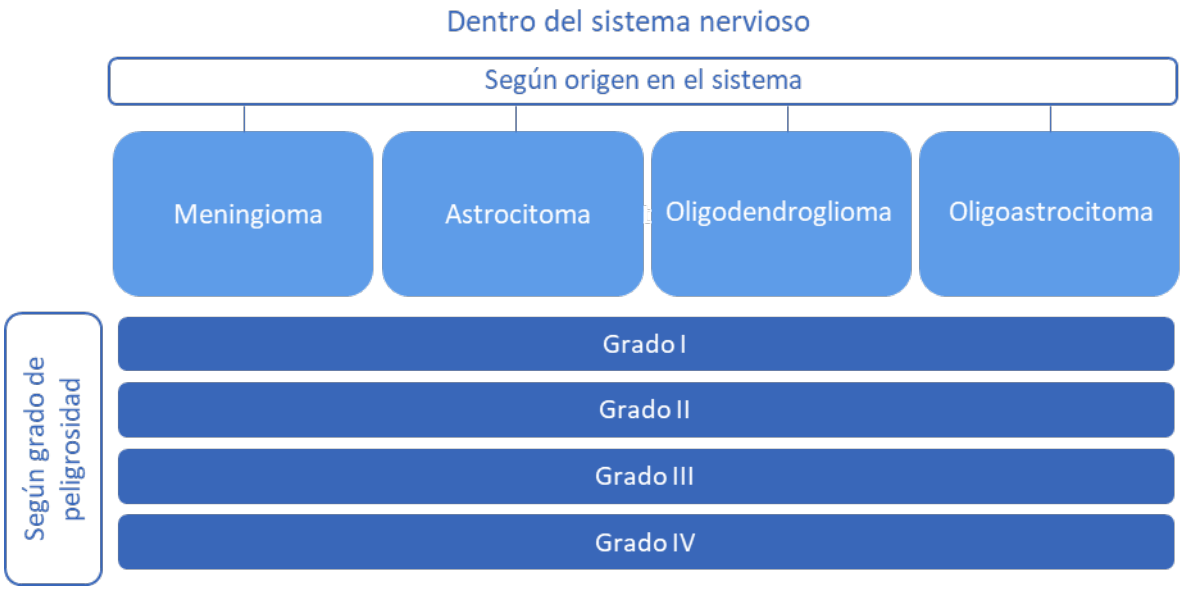



*Figura 3. Clasificación de principales tumores primarios del sistema nervioso.*

### Meningiomas

El nombre meningioma viene dado gracias a que el lugar de origen de este tumor se encuentra en la meninge, células encargadas de cubrir la médula espinal en un organismo como vainas protectoras y recubren a su vez, la fosa craneal del cerebro. Este tipo de tumores suelen ser benignos con crecimiento lento, rodeados por una cápsula, empujan lentamente los tejidos circundantes a medida que crecen sin llegar a formar metástasis en ellos. Rara vez existe una variante del meningioma que se transforma en metástasis, aunque algunos de ellos derivan en otras enfermedades con el paso del tiempo.

Se pueden distinguir diferentes tipos de meningiomas bajo el microscopio. Utilizaremos el método de clasificación de la OMS en el pronóstico para definir la peligrosidad de los diferentes tipos. Un gran número de meningiomas entran en la clasificación de grado I gracias a su lento crecimiento y ser aparentemente asintomáticos, estos casos son casi el 85% dentro de todos los tipos. Alrededor del 10% son clasificados como grado II, aún considerados benignos, aunque denominados como atípicos, ya que el porcentaje de casos es bastante bajo. Para el caso de meningiomas que evolucionan y se transforman en grado II, consiguen el nombre de meningioma anaplásico, estos suponen tan solo un 5% dentro de la clasificación, siendo bastante raros de encontrar [15].

### Astrocitoma

Los astrocitomas consiguen su nombre gracias a su lugar de nacimiento, las células denominadas astrocitos. Los astrocitos son células gliales localizadas en el encéfalo y médula espinal, dan soporte a las neuronas y cumplen funciones vitales para el sistema nervioso central. Este tipo de tumor perdura en el tiempo y puede crecer de forma descontrolada, siendo necesario utilizar la clasificación OMS para determinar su grado de malignidad en función del comportamiento de estos tumores [16], [17].

- Astrocitoma grado I: Incluyen el llamado astrocitoma pilocítico, que suele ser un tumor bien definido, de aspecto sólido o quístico, que suele curarse con cirugía.
- Astrocitoma grado II: conocido como astrocitoma difuso crece lentamente y la cirugía es el primer tratamiento para valorar, dependiendo de la ubicación exacta del tumor y los riesgos de la intervención. Es el astrocitoma más frecuente, suele afectar a niños y adultos por igual (ejemplo en Figura 4. Ejemplo de astrocitoma de bajo grado. Paciente femenina con tumor frontal izquierdo, RMN antes de la cirugía: Tumor frontal izquierdo, aspecto sólido mixto y escaso realce quístico con el contraste [16]

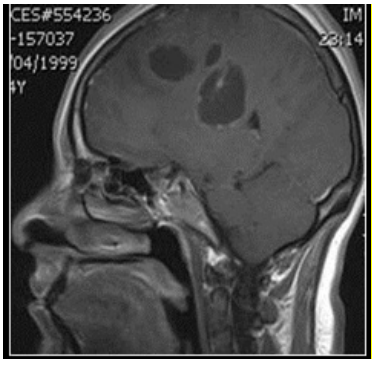



*Figura 4. Ejemplo de astrocitoma de bajo grado. Paciente femenina con tumor frontal izquierdo, RMN antes de la cirugía: Tumor frontal izquierdo, aspecto sólido mixto y escaso realce quístico con el contraste* [16].

- Astrocitoma grado III: El astrocitoma anaplásico, al ser de grado III, tiene un comportamiento más agresivo y se da más en los hombres. Es común considerar en el tratamiento para el paciente como primera instancia la cirugía, aunque dependerá de la ubicación del tumor ya que en algunos casos solo es posible realizar una biopsia para

conocer más información y eliminar, al menos, una parte de la lesión que permita mejorar al paciente. Son tumores de carácter infiltrante lo que dificulta su resección completa.

- Astrocitoma grado IV: Conocido por el nombre de glioblastoma, representa una parte considerable dentro de todos los tumores primarios y es por ello, el más encontrado dentro de los astrocitomas. Son tumores de rápido crecimiento, por lo que producen sintomatología a medida que se propagan por el organismo, como la presión intracraneal. Es un tumor muy agresivo que presenta numerosos vasos sanguíneos a su alrededor, permitiendo que se prolifere y crezca con mayor velocidad, dañando con facilidad el tejido sano. Esta característica dificulta la posibilidad de recibir tratamiento asociado a la cirugía puesto que imposibilita la extracción completa del tejido canceroso.

Normalmente los gliobastomas aparecen directamente en el cerebro y crecen, aunque algunos son el resultado de astrocitomas de bajo grado que evolucionan y terminan convirtiéndose en un tumor maligno (Figura 55) [17].Figura 4. Ejemplo de astrocitoma de bajo grado. Paciente femenina con tumor frontal izquierdo, RMN antes de la cirugía: Tumor frontal izquierdo, aspecto sólido mixto y escaso realce quístico con el contraste [16]

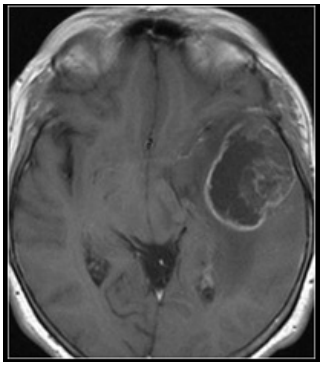


*Figura 5.Ejemplo: glioblastoma multiforme. Paciente femenina con convulsiones y cefalea, de 4 meses de evolución. Resonancia magnética con tumor en el lóbulo temporal izquierdo, con realce en la periferia del tumor a la administración del medio de contraste* [16].

### Oligodendrogliomas

Los oligodendrogliomas reciben su nombre al igual que el resto de los tumores por las células en las que se produce la lesión, en este caso los oligodendrocitos, encargados de producir proteínas y de interactuar con las neuronas para favorecer la activación de estas, además de cumplir con diversos factores neurotróficos. Este tipo de tumor es considerado uno de los

difícilmente transformables en metástasis debido a su lento crecimiento dentro del organismo. Sin embargo, en algunos casos crecen incontrolablemente y producen lesiones de mayor grado [18].

- Oligodendrogliomas grado II: poseen una morfología difusa y suelen ser encontrados dentro de los hemisferios cerebrales o lóbulo frontal. Es un tipo de tumor que puede infiltrarse dentro del tejido sano, siendo capaz de evolucionar a un grado mayor.
- Oligodendrogliomas grado III: El oligodendroglioma anaplásico es un tumor que no suele encontrarse, aunque una vez dado se localizan en el lóbulo frontal y lóbulo temporal. Al ser un tumor de grado III se le considera maligno y su implicación en la enfermedad lleva una serie de patologías como necrosis, problemas con el calcio, hemorragias, etc.

### Oligoastrocitomas

Los oligoastrocitomas son tumores primarios que se caracterizan por originarse a partir de dos células distintas, los oligodendrocitos y los astrocitos, definidas anteriormente. Ambas células pertenecen al mismo grupo de células glia, aquellas encargadas de funciones principales dentro del sistema nervioso. Es por ello por lo que son particularmente extraños dentro los tumores primarios [19] Los casos más frecuentes dentro de este tipo particular son:

- Oligoastrocitomas grado II: Tiene un crecimiento celular moderado con poca o ninguna mitosis. Es posible diferencia en esta lesión las células que lo conforman que pueden ser encontradas mezcladas entre sí o totalmente separadas. Este tipo de tumor benigno es difícilmente diferenciable de los oligodendrogliomas a través de una imagen, requiere de otras técnicas para poder determinar el tratamiento a seguir con el paciente.
- Oligoastrocitomas grado III: tumor de alto grado que se caracteriza principalmente por tener tanto la forma de un astrocitoma como de un oligodendroglioma, siendo difícil identificar cual de sus componentes principales en las células es el que posee malignidad.

## Tumores secundarios.

Hablamos de tumores secundarios cuando las lesiones tienen su origen en tejidos externos al cerebro y que debido a su proliferación son capaces de infiltrase en el tejido cerebral. Según la estadística, afecta entre a un 2,8 y un 11,1 por cada 100.000 habitantes [14]. Los tumores secundarios son los más frecuentes en personas mayores, una vez superados los 50 años se

encuentra una mayor incidencia en la población. Se pueden implantar en varias zonas, principalmente en el hueso del cráneo, las meninges o el parénquima cerebral [14]

Los tipos de cáncer secundario más comunes que acaban en metástasis son los originados en pulmones, mamas, piel y médula espinal. Dentro de estos, el más propenso a propagarse y constituir un organismo metastásico es el melanoma, nombre con el que se conoce al cáncer de piel. En España los tumores que han tenido una mayor mortalidad son los originarios en el pulmón y bronquios, suponiendo en el año 2020 más de 20.000 casos mortales [9].

El diagnóstico de un paciente varía dependiendo de varios factores como el origen del tumor, la edad y estado de salud del paciente, la capacidad de su sistema para hacer frente a enfermedades, variantes genéticas, etc. La supervivencia y pronóstico dependerán del estado en el que se encuentre la metástasis en el organismo, hasta donde ha llegado, órganos afectados, etc. [20]

### 2.2.3. Oxigenación en tumores

A lo largo del capítulo se han presentado que los factores en los que podemos encontrar lesión tumoral son diversos, dentro de estos casos, algunos tumores avanzados se encuentran en situaciones desfavorables de oxigenación que terminan derivando en un estado conocido como hipoxia. Se denomina hipoxia tumoral a la situación que ocurre cuando un tumor cerebral no está siendo provisto del suficiente oxígeno necesario.

Los tumores avanzados continúan creciendo mientras se produce una concentración de oxígeno en el interior de la lesión que no consigue alcanzar la cantidad de oxigenación suficiente. En este caso, se produce una falta de equilibrio entre las células que componen el tumor, aquellas que demandan de oxígeno para poder seguir creciendo y las que se quedan sin poder transportar el oxígeno, lo que deriva en un tumor hipóxico. Esto ocurre debido a que el crecimiento del tumor debilita a los vasos sanguíneos encargados del transporte de oxígeno a todo el organismo, no siendo capaz de abastecer a todas las células que ahora necesitan nutrientes y oxígeno para poder mantenerse sanas y no convertirse en parte del creciente tumor [21] Aquellos tumores en situación de hipoxia presentan zonas diferenciadas, un ejemplo podemos verlo en Figura 6 [4].

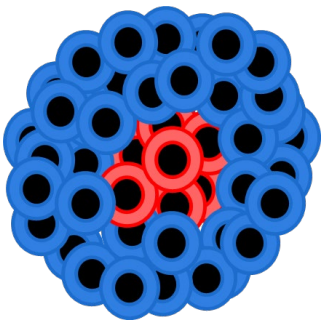


*Figura 6. Tumor primario creciendo sin oxígeno suficiente*

La principal causa de los entornos de tumores hipóxicos es que la acción de proliferación de las células cancerígenas consume la mayor parte del oxígeno que se transporta, dejando a las al tejido sin oxígeno disponible para abastecerse [4], visualmente puede observarse las dos zonas en la Figura 77. Esta falta de oxígeno en el tejido tumoral es la causante de que los tumores aumenten la metástasis o se produzcan cambios en la matriz celular, empeorando el estadio del paciente [22].

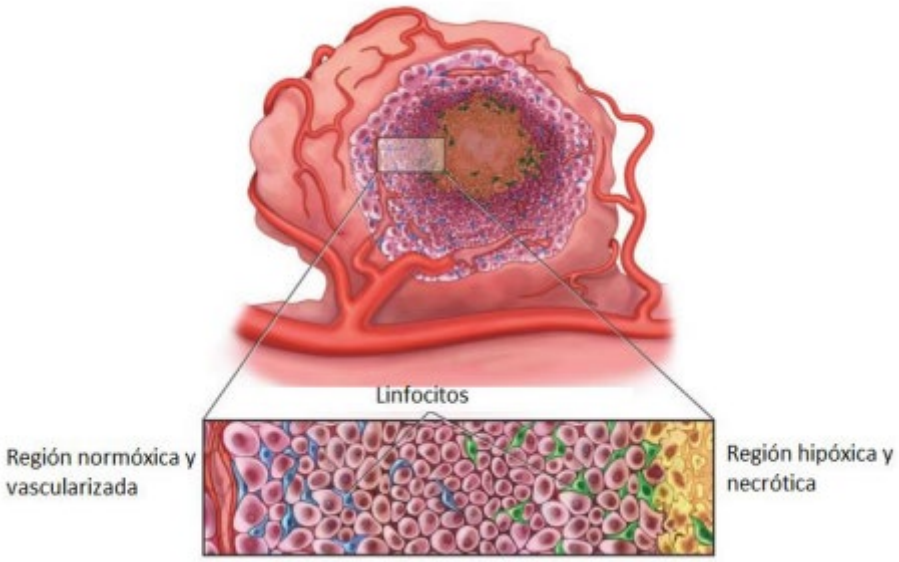



*Figura 7. Tumor hipóxico con distinción entre zona regular de oxígeno e hipoxia [3].*

Las arterias son las encargadas de transporta el oxígeno desde el corazón hasta todos los capilares del cuerpo, que, a su vez, abastecen de oxígeno a los tejidos. El organismo que continua con su transporte habitual de distribución de oxígeno, se encuentra con la problemática de que cuantos más capilares débiles se formen, ya que menor concentración de oxígeno llegará a los

tejidos. Esto produce que las células dentro del tumor más lejanas a estos capilares nuevos no estarán dotadas de todo el oxígeno necesario, dando como consecuencia situaciones nuevas de hipoxia y posible necrosis en estas mismas células que no son del todo alcanzables por los capilares. Caso contrario ocurre en las células a las que los vasos sanguíneos abastecen por completo por su cercanía, permitiendo una mayor proliferación en estas zonas [23].

Anteriormente, en la sección de *Definición*, se había estudiado los daños que producen algunas mutaciones genéticas en determinados genes. En posible garantizar que la hipoxia desestabiliza la entidad genética, haciendo posible que aparezcan las mutaciones genéticas de las que se habló anteriormente y que son perjudiciales para el crecimiento de un tumor maligno canceroso [[24]].

La hipoxia y sus implicaciones en zonas tumorales no solo afectan al crecimiento y empeoramiento a nivel celular, además, supone un gran problema a la hora de recibir tratamiento. Esto se debe a que una de las consecuencias de la hipoxia en el metabolismo es el deficiente suministro a los tejidos en los que también se transportan los fármacos necesarios para mejorar la lesión. Se ha observado en pacientes que los tumores que se enfrentan a hipoxia tienen un mayor rechazo a la quimioterapia y radioterapia. El conocimiento de este obstáculo en el futuro del paciente abre un campo al estudio de nuevas técnicas como tratamiento para pacientes que ofrezcan resistencia a los fármacos convencionales por su fallo vascular.

### 2.2.4. Angiogénesis

La definición del diccionario de la palabra “angiogénesis” cita textualmente: formación de los vasos sanguíneos. Entrando en profundidad en este término el proceso está compuesto por migración, crecimiento y diferenciación de células endoteliales localizadas en el interior de los vasos [25]. El factor vascular de crecimiento endotelial (VEGF) es uno de los responsables de producir señales al organismo para que los receptores del mensaje promuevan el crecimiento y supervivencia de los nuevos vasos sanguíneos. Además del VEFG existen otras señales que producen el mismo efecto [25]. Existen a su vez, señales químicas del organismo que son bloqueantes para la angiogénesis, mandando el mensaje de parar la formación nueva de vasos sanguíneos. En el estudio de la angiogénesis y sus determinantes el VEGF y su receptor son en la actualidad, los que se toman como objeto de estudio para su comprensión y análisis [26].

El crecimiento de nuevos vasos sanguíneos de forma descontrolada y desproporcionada puede derivar en estados de enfermedades o desequilibrio del organismo. En la angiogénesis son varios

los factores encargados de que exista un balance entre todos los agentes que la conforman como proteínas, células implicadas en el desarrollo y receptores [25] La formación de una red de nuevos capilares es un componente esencial en múltiples procesos fisiológicos, como hemos mencionado anteriormente, y también lo es en procesos patológicos tumorales y no tumorales (artritis, degeneración macular, etc.) [13]. En el desarrollo de un tumor maligno, la angiogénesis es un elemento clave en la multiplicación metastásica. El concepto de angiogénesis tumoral se definiría como la capacidad que tiene un tumor para generar una señal química que estimule la creación de nuevos vasos sanguíneos que abastezcan de nutrientes y oxígeno a los tumores para que este sea capaz de crecer en tamaño y forma [27].

Cuando un tumor crece más allá de 1-2 $mm^3$ requiere aporte vascular y el estímulo primario para la formación de nuevos vasos parece ser la hipoxia. Dependiendo del tipo de tumor se usa un mecanismo genético distinto para desencadenar el proceso angiogénico [26] Varios estudios confirman que los tumores dañados por hipoxia son los que comúnmente desarrollan angiogénesis. Por lo tanto, los mecanismos encargados de responder a la falta de oxígeno también parecen ser desencadenantes de la creación de nuevos vasos sanguíneos, considerándose un factor como el VEGF. La angiogénesis es uno de los factores más importantes en el crecimiento y desarrollo de un tumor puesto que se encarga de las necesidades de abastecimiento de este. A la hora de crear nuevos vasos sanguíneos más débiles en el organismo y limitarse el oxígeno que abastece a las células, se consigue no abastecer el tejido tumoral, pero si conseguir la proliferación y crecimiento de células cancerosas [26].

Es inevitable estudiar en estos casos que impacto tiene la hipoxia y la angiogénesis en el cáncer y su desarrollo, específicamente en situaciones de metástasis. El estudio de la oxigenación de tumores que promueve la angiogénesis a través de emisores químicos como el VEGF o la hipoxia en si misma [28]. El impacto de la oxigenación en clínica radica en el establecimiento de nuevas dianas terapéuticas para combatir diversas enfermedades cardiovasculares además de las asociadas a todos los procesos relacionados con el cáncer.

## 2.3. Imágenes hiperespectrales

### 2.3.1. Definición

La imagen espectral, también conocida como espectroscopia de imagen, se refiere a una técnica que integra imágenes y espectroscopia existentes para obtener información espacial y espectral

de un objeto. Las imágenes espectrales pueden dividirse en imágenes multiespectrales, imágenes HS e imágenes ultraespectrales según su resolución espectral, número de bandas, anchura y contigüidad de estas. Los sistemas de imágenes multiespectrales suelen recoger datos en pocas bandas espectrales anchas y relativamente no contiguas, que suelen medirse en micrómetros o decenas de micrómetros. La principal diferencia de una imagen HS y el resto de las imágenes como las multiespectrales consiste en el número de canales espectrales que la conforman (Figura 88).

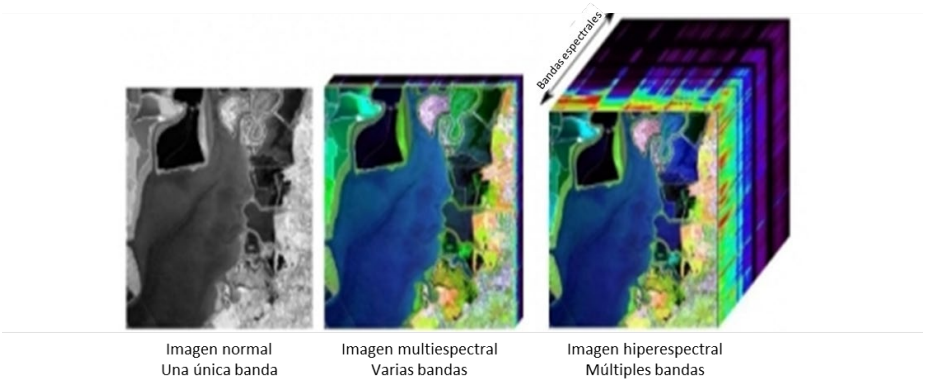



*Figura 8. Comparativa del número de bandas representadas en una imagen normal, multiespectral e HS.* [29]

Se denomina imagen HS a un conjunto de imágenes conformadas por varias y diversas bandas espectrales que permiten la reconstrucción del espectro de reflectancia píxel a píxel de la misma imagen, definiendo el valor del píxel por las medidas de todas las bandas espectrales que conformen la imagen. El conjunto de imágenes que se ha obtenido en la reconstrucción se denomina hipercubo. Este hipercubo tiene tres dimensiones: la componente espacial en los ejes X e Y, y, la componente espectral en el Z (Figura 9) [29].

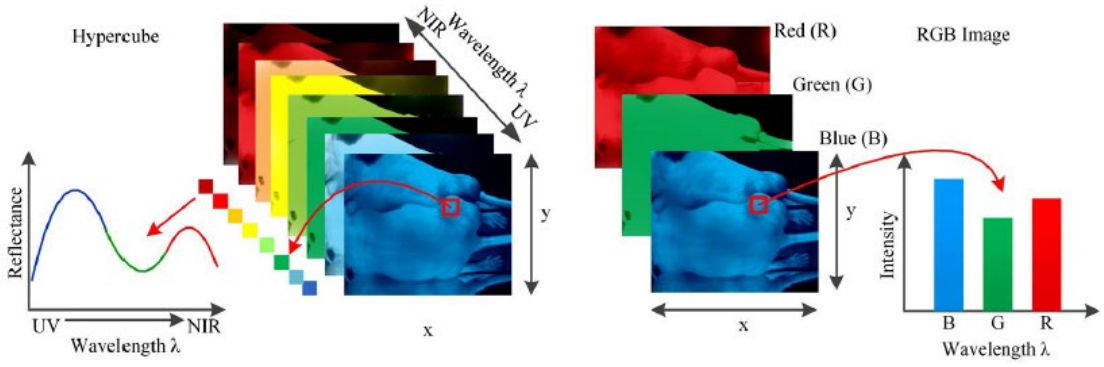



*Figura 9. Información que contiene un píxel en una imagen normal comparado con una imagen HS* [29].

La imagen espectral puede visualizarse como un cubo de carácter tridimensional, o un conjunto de varias imágenes bidimensionales conjuntas, debido a su estructura intrínseca, anteriormente definida en la que la cara del dicho cubo representa a los ejes X e Y, y, la profundidad de este, la función de la longitud de onda[29]. A partir de un solo cubo espectral se pueden crear espectros individuales, mapas y cubos espectrales completos. Los cubos espectrales muestran diferentes grosores estratificados y gráficos en una perspectiva tridimensional lo que permite un estudio general de la imagen. Dado que los cubos espectrales se crean generalmente en áreas más grandes que las que el conjunto de planos focales puede recoger en un solo fotograma, la segunda dimensión de la imagen espacial se crea con el tiempo. Este método de visualización es una forma de observar características espectrales específicas dentro de la escena [30].

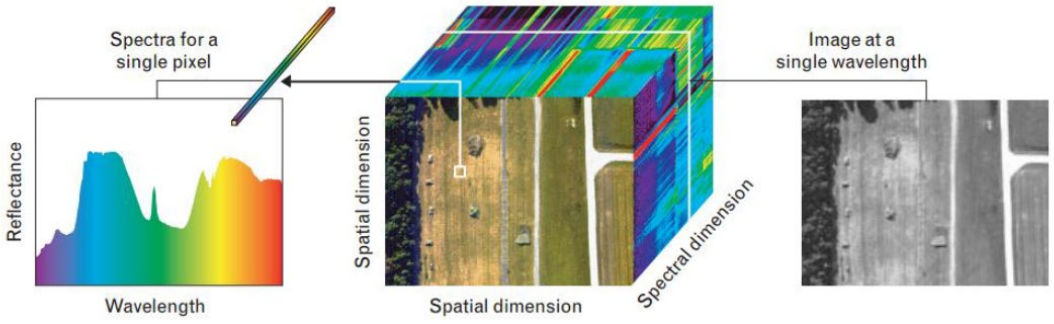



*Figura 10. Desglose de la información que se obtiene de un solo píxel a través de la imagen HS, por cada longitud de onda y la formación del hipercubo [42].*

Las imágenes HS son producidas por instrumentos llamados sistemas de imágenes HS o espectrómetros de imágenes. Estos instrumentos están formados por tres componentes: una fuente de luz, un separador de espectros y un detector. La luz se dirige a la superficie de un objeto. Esta luz después de entrar y atravesar el objeto se refleja, se transmite y/o se absorbe y es entonces cuando pasa por un separador espectral que discrimina espectralmente la luz reflejada en muchas bandas espectrales contiguas en el espectro electromagnético. Finalmente, la luz llega al detector, donde se recoge como una imagen bidimensional que contiene los datos espectrales inherentes a cada píxel y se transfiere a un ordenador para su posterior procesamiento y análisis[29].

Es posible determinar los elementos que componen una imagen debido a las características espectrales específicas que presenta cada material localizado en las longitudes de onda que cubren el espectro visible y el infrarrojo [30]. La forma de clasificar y visualizar los datos, la

información y el mundo que nos rodea es un recurso importante en muchas áreas de análisis, investigación y estudios teóricos. Un analista de imágenes determina el enfoque de clasificación y decide utilizar clases espectrales o clases de información. Así podemos decir que un grupo de píxeles con características espectrales casi idénticas se considera parte de una clase espectral. Un analista utiliza una clase de información cuando intenta identificar elementos o grupos específicos dentro de una imagen. El objetivo principal de un analista de imágenes es tratar de hacer coincidir la clase espectral con una clase de información [29]–[31].

Una vez se ha decidido utilizar clases espectrales o de información, el proceso de clasificación puede ser supervisado o no supervisado. Una clasificación supervisada se basa en algoritmos de detección que utilizan píxeles de muestras de referencia conocidas, normalmente situadas dentro de una misma escena, como base de comparación con otros píxeles de objetos de la misma escena. Por ejemplo, si el analista de la imagen sabe que un área específica es una carretera, todas las demás áreas con el mismo algoritmo de detección serán también una carretera y preparará los datos para que el algoritmo de detección determine que también se trata de una carretera. Por lo tanto, en la clasificación supervisada, el analista suele empezar con clases de información ya conocidas que se utilizan para definir clases espectrales representativas que se ajustan a las muestras de referencia [30], [31]

La clasificación no supervisada es básicamente lo contrario de la clasificación supervisada. Los píxeles de una imagen se agrupan en clases espectrales basándose únicamente en la información de los datos en comparación con las bibliotecas de firmas u otras clases de información conocidas. La clasificación de imágenes puede utilizarse para asignar la información espacial y espectral en varios campos, comúnmente denominados "mapas temáticos".

### 2.3.2. Detección del cáncer

La tecnología HS supone una herramienta potente para el diagnóstico de enfermedades sin tener que realizar técnicas invasivas para el paciente. Gracias al desarrollo de esta tecnología y a las características que poseen las imágenes HS es posible conocer niveles que antes no estaban al alcance la fluorescencia, niveles de absorción o dispersión del tejido en la evolución que lleva consigo una enfermedad y los cambios que se producen mientras. Las imágenes HS han demostrado ser efectivas en el marco de detección y diagnóstico de enfermedades a través de imágenes, posibilitando visualizar los cambios bioquímicos provocados por las distintas enfermedades como el cáncer [30].

Actualmente, se ha convertido en esencial el estudio de imágenes a la hora de realizar un diagnóstico de cualquier tipo de lesión, así como en medicina preventiva y seguimiento de los resultados de un tratamiento. Los resultados de las imágenes HS y su análisis sirven de apoyo al médico para realizar diagnósticos más certeros en cambios de tejido o aumento de este. En casos como tumores cerebrales este análisis se dificulta debido a la aglomeración de células que componen el tejido y a los vasos sanguíneos que se encuentran alrededor que vuelve complejo el visualizar el tumor correctamente. Es por ello por lo que para estos casos es necesario que la información se complete con otro tipo de técnicas como la biopsia y toda la información asociada al caso de paciente y que conforma su historial clínico, etc. No obstante, se ha mejorado a lo largo de los años en esta tecnología aplicada a imágenes para la detección y clasificación de tumores y en la cual se crea el contexto en el que aparecen las imágenes HS en el estudio de detección del cáncer.

La imagen HS es una modalidad de imagen emergente para aplicaciones médicas, especialmente en el para definir un pronóstico de patologías y para la cirugía guiada por imagen [29]. Esta tecnología presenta algunas ventajas frente a las técnicas empleadas actualmente para el diagnóstico de cáncer, como la resonancia magnética (RM), la tomografía computarizada (TC), el ultrasonido (US) y la tomografía por emisión de positrones (PET). La RM se caracteriza por la exposición a radiaciones nocivas, se requiere de un operador formado y se considera una tecnología costosa. La ecografía tiene la desventaja de tener un bajo contraste de imagen. Y, por otro lado, el diagnóstico PET y la TC son económicamente costosas y utilizan altas dosis de radiación. Dado que la mayoría de estos métodos no son económicos y requieren operadores capacitados, el acceso de los pacientes a estas importantes medidas para salvar vidas es limitado. El objetivo a largo plazo de las imágenes HS en la detección del cáncer es desarrollar una herramienta sencilla, no invasiva y sin riesgos, que proporcione una detección temprana de tumores malignos potencialmente mortales y sea asequible para todos. Esta tecnología puede utilizarse en diversos contextos, tanto para mejorar el cribado como para el análisis cuantitativo de los tejidos [30], [32].

Una de las características beneficiosas de esta técnica de procesado HS es que puede adquirir por cada píxel información del espectro de reflectancia, absorción y fluorescencia de la imagen. La luz que llega al tejido biológico se dispersa a causa de la forma heterogénea que tiene la masa de estudio biológica y una absorción en la hemoglobina, la melanina y el agua a medida que se propaga por el tejido. A lo largo de la enfermedad, la forma de estudio cambia y con ella se presupone que también varían las materias de estudio (absorción, fluorescencia y dispersión).

Por lo tanto, la luz reflejada, fluorescente y transmitida del tejido captada por la imagen HS da información de diagnóstico cuantitativa sobre la patología del tejido. Algunos estudios de investigación que emplean imágenes HS como herramienta de diagnóstico se pueden encontrar en [29]. El tipo de enfermedad que se quiere analizar y la técnica de adquisición empleada marcan las principales diferencias entre los distintos trabajos de investigación.

Es conocido, que cada material tiene su propia firma espectral, pero los cambios de las características biológicas y patológicas en los tejidos y órganos también tienen una estrecha relación con los espectros. Es posible visibilizar los cambios que se producen patológicamente gracias a que la firma espectral que se crea en una imagen posee información espectral de las diversas regiones de longitud de onda.

El análisis de imágenes permite extraer información útil para el tratado de lesiones a partir de un gran conjunto de datos HS médicos a nivel tisular, celular y molecular. Dado a su alto detalle para obtener información de los datos, se hace una técnica fundamental para la detección, el diagnóstico y el seguimiento de enfermedades. Los hipercubos que posean una alta resolución espacial y espectral pueden contener potencialmente más información de diagnóstico, pero, dificultan los análisis automáticos de datos HS. Con abundante información de las componentes principales del hipercubo disponibles, se requieren métodos para la clasificación de imágenes y el conjunto de datos HS que consigan extraer, desmezclar y clasificar la información relevante. El objetivo no es sólo discriminar entre diferentes tejidos (sano y maligno) y proporcionar mapas de diagnóstico, se busca el poder descomponer las mezclas en los espectros de los componentes moleculares puros y correlacionar estas huellas moleculares (biomarcadores) con los estados de enfermedad.

Aunque existen diversos estudios del tratado de imágenes HS en ámbitos muy variados como la teledetección, en el ámbito médico su desarrollo y aplicación están evolucionando de manera muy lenta. Las relaciones entre las características espectrales y los mecanismos biomédicos subyacentes no se conocen bien. Los pasos básicos para el análisis de imágenes HS generalmente implican el preprocesamiento, la extracción de características, la selección de estas, y la desmezcla y/o clasificación de los datos [30].

El estudio del tejido in vivo, visualmente en la imagen a continuación (Figura 11), permite proporcionar un diagnóstico no invasivo de la enfermedad y proporciona una herramienta de orientación automática para la cirugía. Tras una cirugía de eliminación del cáncer, el informe de patología señala que el margen quirúrgico es uno de los tres tipos siguientes: claro, positivo o estrecho. En el caso de un margen claro, el tejido normal rodea a las células cancerosas. En caso de margen positivo, las células cancerosas aparecen en los márgenes de los límites de la extracción del tumor, lo que llevaría a una cirugía adicional para eliminar el tejido canceroso restante. Es el caso, por ejemplo, del cáncer de mama donde entre el 20 y el 50% de las cirugías requieren una cirugía adicional para eliminar los tejidos cancerosos restantes.

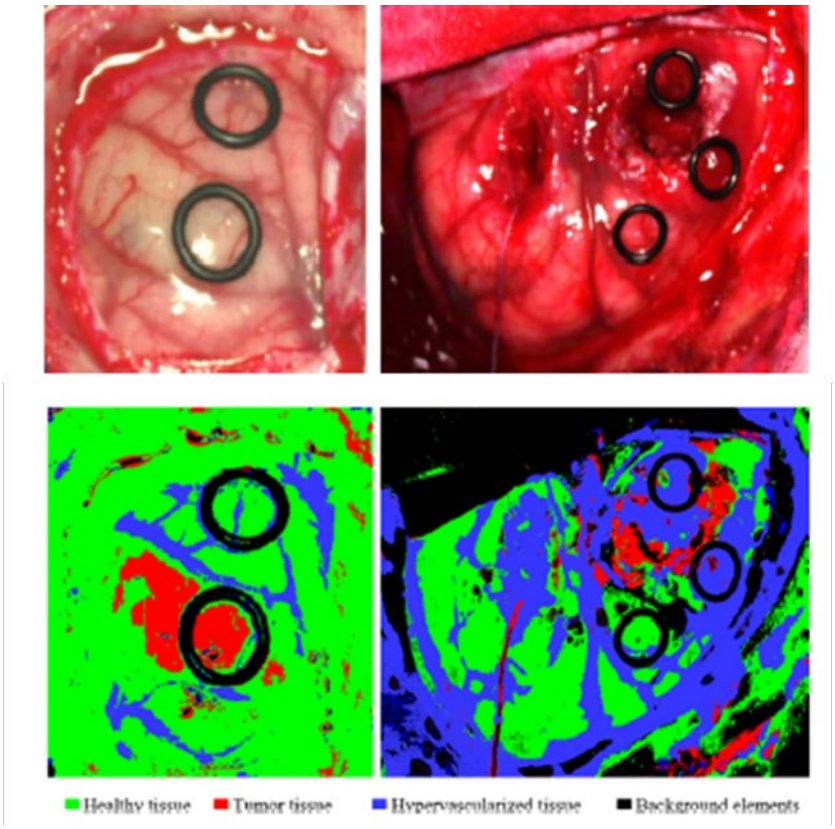



*Figura 11. Casos de estudio para el análisis de hipercubo en la parte superior. Mapa obtenido de los casos de estudio, etiquetas y clasificación de las imágenes en la parte inferior.*

Pocos estudios explican los principios básicos y los sistemas instrumentales para el sistema HSI in-vivo en el campo biomédico [33], y sólo algunos trabajos se ocupan de la detección in-vivo del cáncer gastrointestinal (GI). En uno de ellos [34], empleando la endoscopia rígida, los autores proponen una nueva etapa de preprocesamiento para detectar células cancerosas en la laringe, superando varios problemas relacionados con las interferencias de la imagen como son el error de registro de las imágenes individuales debido a los latidos del paciente y los reflejos especulares. Otro estudio en el campo de la endoscopia gastrointestinal superior se encuentra

en [35]. En él, los autores presentan la calibración y los resultados de las pruebas obtenidas por medio de una reflectancia HIS y una configuración de video endoscopio flexible para la detección de cáncer gastrointestinal in vivo.

El uso de la HSI puede dividirse principalmente en la diagnóstico y detección de patologías . En lo que respecta al cáncer, esta técnica se ha utilizado principalmente en la detección del cáncer de cuello uterino (tanto in vivo [36], [37]como in vitro [38]), para la detección del cáncer de mama (tanto in vivo [39] como in vitro [40]), para la detección del cáncer de piel (tanto in vivo [41] como in vitro [42]) y para la detección del cáncer de cabeza y cuello (tanto in vivo [43] como in vitro [44]). En cuanto a otras patologías, la HSI también se ha utilizado para el estudio de la patología cardíaca y circulatoria (tanto in-vivo [45]como in-vitro [46] y para el estudio de las patologías asociadas a la retina [47]. Recientemente, esta técnica ha demostrado ser de especial relevancia para guiar a los cirujanos en diferentes operaciones como podrían ser la mastectomía [48] , la cirugía de la vesícula biliar [49], la cirugía renal [50] y la cirugía abdominal [51].

Son múltiples los usos que se le pueden dar a las imágenes HS dentro del campo diagnóstico médico. Actualmente los estudios recientes no solo se centrar en encontrar nuevos métodos computacionales de ser más sencillos y ágiles a la hora del procesado de estas imágenes, también se centran en encontrar métodos de clasificación mixta que permita usar varios tipos en clasificadores convencionales. En el siguiente capítulo encontraremos los diferentes algoritmos de clasificación, su estudio y aplicación en la literatura y las posibilidades que plantean.

## 2.4. Algoritmos

En esta sección se presenta como base teórica los algoritmos estudiados en el proyecto y que posibilitan la visualización de datos e información necesaria para investigaciones basadas en el análisis de imágenes HS. Un gran número de píxeles procesados por sensores HS suelen estar compuestos por más de un material, es decir, un mismo píxel se forma con diferentes materiales. Por tanto, los píxeles de una imagen HS se pueden clasificar en:

- Píxeles puros: constituidos por un mismo tipo de material.
- Píxeles mixtos: compuestos por la combinación de varios materiales de distinto tipo.

Por tanto, es por ello por lo que las imágenes HS estarán conformadas en su mayoría por píxeles mixtos, combinando como consecuencia espectros asignados a los distintos materiales. Dentro

de estas imágenes, cada unidad de píxel tiene una firma asociada y característica. Los datos aportados por estas firmas espectrales facilitan la información necesaria para proceder a realizar una identificación y cuantificación de los materiales que componen la imagen.

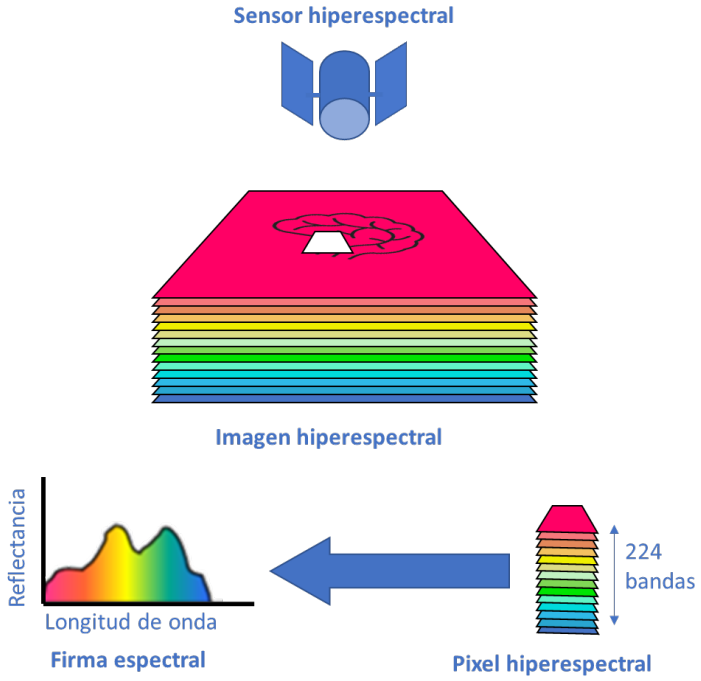



*Figura 12. Firma HS por cada píxel de una imagen HS.*

La imagen procesada está compuesta por las diferentes firmas espectrales pertenecientes a los distintos tipos de objetos que la componen. Por ello, para poder diferenciar estos materiales, es necesario encontrar las diferencias en las características espectrales asociadas a cada elemento.

La detección de estas diferencias espectrales es la encargada de extraer información de los píxeles de la imagen y que tienen una respuesta significativa o inusual. La manera de detectarlos puede ser:

- Supervisada: el método se basa en conocimientos ya conocidos de la imagen.
- No supervisada: no se dispone de ninguna información previa.
- Semisupervisada: se conoce información parcial. Existen datos de entrenamiento con etiquetas y datos sin ellas. Es posible considerar a este tipo de algoritmos como supervisados sin necesidad de poseer etiquetas de todos los datos.

En los subapartados que prosiguen se analizarán los algoritmos que conocemos en la tecnología de imágenes, sus fundamentos teóricos y su posible aplicación al campo médico en este trabajo para la detección de tumor y vascularización asociada que agravan la sintomatología del cáncer.

### 2.4.1. Algoritmos de clasificación no supervisados

En el análisis de datos de imágenes HS, los algoritmos no supervisados tienen un interés especial en ciertos ámbitos, entre ellos destaca el desmezclado espectral, el cual consiste en la obtención de las características principales que conforman una firma espectral recogida en un píxel. Algunos fines son la reducción de la dimensión espectral con el fin de lograr un procesado más rápido y eficiente o reducir la selección de las bandas espectrales con aquellas más discriminantes, o, un preprocesado de los datos para la extracción de características que permita alcanzar unos mejores resultados por clasificadores espectrales [30], [52].

Aunque son varias las técnicas de clasificación de píxeles a través de algoritmos no supervisados, en este proyecto definiremos los métodos destacados en la literatura con mejores resultados en el ámbito biomédico. Se presentan por tanto a continuación los métodos: *K-Means* [53], *Gaussian Mixture*, *Spectral Clustering*, *Agglomerative Clustering* y el método *Birch*.

#### K-Means

El algoritmo *K-Means* se engloba dentro de la rama de clasificación no supervisada. Este método se basa en la agrupación de objetos en *k* grupos con similitudes en sus características, siendo *k* un parámetro definido anteriormente. La agrupación se realiza utilizando los datos recogidos en los vectores que componen cada píxel y que contienen la información espectral. Esta metodología se lleva a cabo a través de un método estadístico basado en espectros promedios de dichos grupos.

El algoritmo actúa siguiendo tres pasos principales:

1. Se determina el número de grupos, denominado como *k,* y se proceden a *establecer k* centroides en el espacio de los datos seleccionándolos aleatoriamente.
2. Se realiza una muestra compuesta por determinados píxeles y para cada una de estas muestras se calcula la similitud del píxel a todos los centroides establecidos, a través

de la técnica de mínima distancia, se incluye en la clase, aquellos que presentan una menor distancia al centroide de estudio.

3. Finalmente se recalcula el centro de gravedad de la clase que ahora conforman los píxeles clasificados en el paso anterior y se vuelven a clasificar todos los píxeles siguiendo el mismo modelo. Es posible realizar una clasificación completa de los píxeles si se limita la desviación estándar o la distancia máxima de búsqueda.

Este algoritmo cuenta con la ventaja de utilizar una metodología de clasificación sencilla, de poca complejidad computacional y rápida que la hace ideal en clasificaciones de bases de datos de tamaños considerables. Por otro lado, una de las desventajas es que surge la necesidad de realizar una búsqueda de técnicas que consiga determinar el número de clases óptimo y que no se produzca una segmentación excesiva o una incompleta. [53], [54].

## Gaussian Mixture

Una mezcla gaussiana (*Gaussian mixture*) es una función compuesta por varias ecuaciones gaussianas, donde cada una es identificada por $k \in \{1, \ldots, K\}$, donde K es el número de grupos o clases de un determinado conglomerado de datos. Los parámetros que se encuentran en esta función compuesta son [55]:

- Una media $\mu$ que define su centro.
- Una covarianza $\Sigma$ que define su anchura. Equivalente a las dimensiones de un elipsoide en un escenario multivariante.
- Una probabilidad de mezcla $\pi$ que define el tamaño de la función gaussiana.

Los coeficientes del algoritmo son probabilidades y deben cumplir la siguiente condición (Ecuación 1):

*Ecuación 1. Condición de probabilidad de coeficientes*

$$\sum_{k=1}^{K} \pi_k = 1$$

Para obtener los valores óptimos para estos parámetros se debe asegurar el garantizar que cada ecuación gaussiana se ajusta a los puntos de datos pertenecientes a cada clúster. Esto es

exactamente lo que hace la máxima verosimilitud. La función de densidad gaussiana la podemos ver en la siguiente Ecuación 2:

*Ecuación 2. Densidad Gausiana*

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Donde x representa el punto de datos, $D$ el número de dimensiones de cada punto de datos, $\mu\ y\ \Sigma$ la media y la covarianza respectivamente [55], [56].

Si tenemos un conjunto de datos compuesto por N = 1000 puntos tridimensionales (D = 3), entonces x será una matriz de 1000 × 3, $\mu$ será un vector de 1 × 3, y $\Sigma$ será una matriz de 3 × 3. Sin embargo, dado que no se trata solo de una única ecuación gaussiana, sino de muchas, este método se complicará cuando llegue el momento de encontrar los parámetros para toda la mezcla.

## Spectral clustering

El agrupamiento espectral *(Spectral clustering*) utiliza la información de los valores propios (espectro) de matrices especiales derivadas del gráfico o del conjunto de datos. Los métodos de agrupamiento espectral son atractivos, fáciles de implementar y considerablemente rápidos, especialmente para conjuntos de datos que se encuentran dispersos. Este algoritmo trata la clasificación de datos como un problema de partición de gráficos sin hacer ninguna suposición sobre la forma de los grupos de datos [56].

El *spectral clustering* permite agrupar también datos no gráficos y no realiza suposiciones sobre la forma de los grupos. Las técnicas de agrupación por clases, como *K-means*, asume que los puntos asignados a un grupo son esféricos alrededor del centroide de este, suposición que puede no ser siempre relevante. En estos casos, el agrupamiento espectral ayuda a crear clases más precisas. Puede agrupar correctamente observaciones que pertenecen al mismo grupo pero que se encuentran más alejadas que las observaciones de otras clasificaciones, esto es gracias a la reducción de la dimensión [56].

Los centros en este algoritmo deben estar conectados, pero no necesariamente tienen límites convexos, a diferencia de las técnicas de clasificaciones convencionales en las cuales la creación de cada clase o grupo se conforma a raíz de puntos de datos. A pesar de ser un algoritmo que presenta grandes ventajas, computacionalmente no tiene buenos resultados en bases de datos de gran tamaño, debido a la necesidad de calcular los valores y vectores propios para la clasificación, por lo que el método pierde precisión.

## Agglomerative Clustering

El agrupamiento aglomerativo es un método de clasificación ascendente en el que las clases tienen subclases, que, a su vez tienen, otras subclases . Puede comenzar colocando cada objeto en una clase y luego mezclar estas clases en clases cada vez más altas hasta que todos los objetos estén en una clase o grupo individual o hasta que necesite una condición de terminación definitiva. Algunos métodos de agrupación jerárquica utilizan este tipo de línea de clasificación, se diferencian unos de otros únicamente en su descripción de la similitud entre clases.

En diferencia con el método *K-means* de clasificación que comienza con un número constante de clases definidas y asigna todos los datos exactamente a ese número, el método aglomerativo en contraposición, consiste en la cohesión de diversos grupos similares entre sí en clases mayores. Este enfoque comienza con cada punto de datos formando su propio grupo y los combinan gradualmente con clases cada vez más altas hasta que todos los puntos se han reunido en un gran grupo [57].

El primer proceso consiste en producir una matriz de similitud. La matriz de similitud es una tabla de algunas distancias por pares o grados de similitud entre clases. Originalmente, la matriz de similitud incluye la distancia por pares entre pares individuales de registros.

Puede parecer que con *N* clases originales para *N* puntos de datos, se necesitan $N^2$ cálculos de medidas para hacer la tabla de distancias. Si la medida de similitud es una métrica de distancia verdadera, sólo se necesita la mitad porque algunas métricas de distancia verdadera siguen el método de que $Distancia(X, Y) = Distancia(Y, X)$.

Seguidamente el proceso se basa en descubrir el valor más pequeño de la misma matriz. Así se reconocen los dos grupos más parecidos entre sí. A través del método aglomerativo se puede

combinar estos dos grupos en uno nuevo y refrescar la matriz de similitud, restaurando las dos filas que describían la clase inicial con una nueva fila que define la distancia entre el grupo fusionado y los grupos restantes.

Ahora hay $N - 1$ grupos y $N - 1$ filas en la misma matriz. Se puede iterar el paso de fusión $N - 1$ veces, por lo que algunos datos pertenecerán al mismo grupo grande. En cada iteración se reconocen los grupos combinados y la distancia entre ellos [57].

## Birch

*Birch* es uno de los algoritmos de clasificación avanzados más útiles y precisos para realizar clasificación en bases de datos con gran tamaño. Este algoritmo de clasificación consiste en la agrupación de conglomerado de datos inicialmente en pequeños grupos, y después estos los resume en clases. No agrupa directamente la agrupación creada en clases. Por eso, este método se utiliza a menudo con otros algoritmos de clasificación, ya que después de hacer el resumen, este también puede ser agrupado por otros algoritmos de clasificación [58].

Birch es un método de clasificación escalable basado en el agrupamiento jerárquico y sólo requiere un escaneo único del conjunto de datos. Este algoritmo se basa en el árbol CF (*clustering features*). Además, este algoritmo utiliza un resumen con estructura de árbol para crear clases.

En el contexto del árbol CF, el algoritmo comprime los datos en los agrupamientos de nodos CF. Los nodos que tienen varias subclases pueden llamarse subclases CF. Estos subconglomerados CF se sitúan en nodos no terminales [58].

El algoritmo Birch por tanto sigue principalmente cuatro fases:

- Escanear los datos en la memoria.
- Condensar los datos (redimensionar los datos).
- Agrupación global.
- Refinar los grupos.

En estas cuatro fases, dos de ellas (redimensionar datos y refinar grupos) son opcionales. Aparecen en el proceso cuando se requiere más claridad. Pero el escaneo de datos es igual que la carga de datos en un modelo. Después de cargar los datos, el algoritmo los escanea en su

totalidad y los encaja en los árboles de jerarquización. En la condensación de datos, reajusta y redimensiona estos para que encajen mejor en el árbol de clasificación. En la agrupación global, envía los árboles para su agrupación utilizando los algoritmos de agrupación existentes. Por último, el refinamiento soluciona el problema de datos valorados mal asignados a diferentes nodos.

### 2.4.2. Algoritmos de clasificación supervisados

En la clasificación de información de imágenes los algoritmos más frecuentes como recurso son los supervisados puesto que poseen un buen resultado en rendimiento para muestras etiquetadas. En el análisis de imágenes HS se requiere de mucha precisión y capacidad de procesado de datos, por lo que elegir este tipo de algoritmos cumple los requisitos para la clasificación de estas imágenes [59].

Dos etapas conforman la clasificación mediante algoritmos supervisados. Por un lado, la etapa de entrenamiento consiste en introducir en el sistema encargado de la clasificación una selección de muestras ya etiquetadas para que se produzca una primera clasificación de estas (Figura 13). Por otro lado, se encuentra la etapa de verificación, en esta etapa se introduce otra selección diferente de muestras etiquetadas y se volverá a realizar la clasificación, de esta manera, se podrá identificar si existe algún error en el sistema clasificador (Figura 14). Para poder realizar las pruebas de entrenamiento y verificación es necesario conocer anteriormente el etiquetado de clases y muestras correspondientes. Para este TFG se han analizado los algoritmos de clasificación principales en el análisis de imágenes HS.

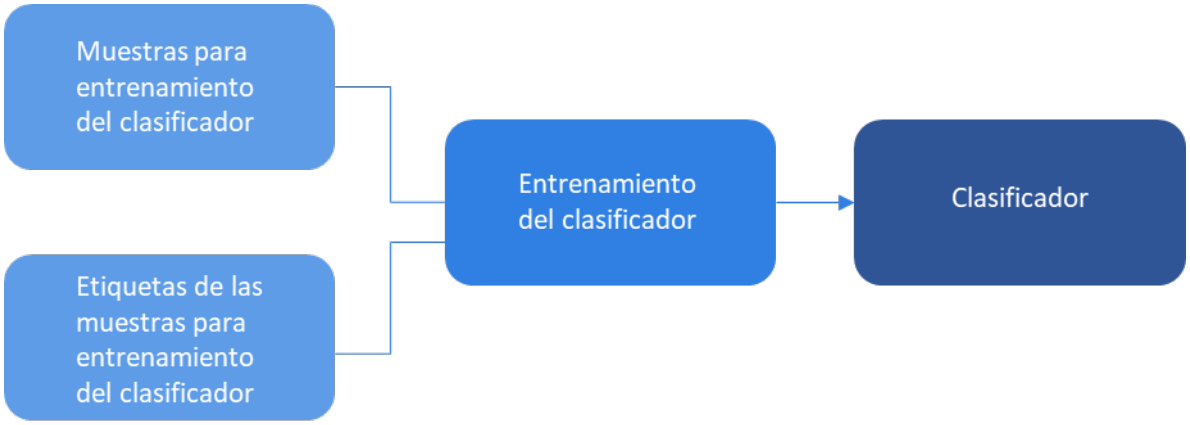



*Figura 13. Etapa de entrenamiento.*

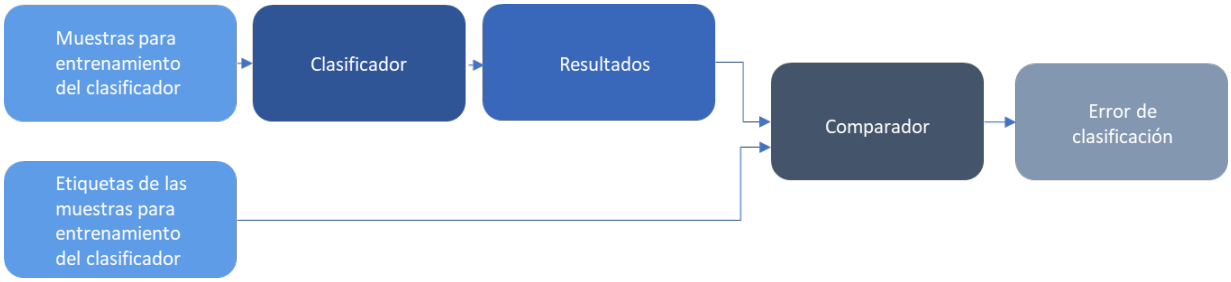



*Figura 14.Etapa de verificación del sistema.*

## Support Vectors Machines

El algoritmo basado en máquinas de soporte vectorial (SVM) fue introducido en la década de 1990 por Vladimir Vapnik [60], basándose en el aprendizaje estadístico en la resolución de cuestiones asociadas a la regresión y clasificación binaria lineal o hiperplano, mediante el uso de las funciones *kernel* [61]. Los *kernels* son funciones matemáticas usadas para mostrar los datos originales en una dimensión del espacio en el que es posible separar linealmente los datos y posibilitando la clasificación. Es decir, los *kernels* son los encargados de delimitar los espacios de gran dimensionalidad mediante el uso de parámetros en la entrada que obtiene como resultado una definición que puede no se ser obligatoriamente lineal. Dependerá, por tanto, de las consideraciones de proximidad de referencia la obtención de unos u otros resultados [62].

Entre las funciones *kernels* de entrada existentes podemos encontrar:

- Polinomio homogéneo:

*Ecuación 3. Ecuación homogénea*

$$k(x, x') = (x \cdot x')^d$$

- Polinomio heterogéneo:

*Ecuación 4. Ecuación heterogénea*

$$k(x, x') = (x \cdot x' + 1)^d$$

- Función de base radial:

*Ecuación 5. Función de base radial*

$$k(x, x') = e^{-\gamma\|x-x'\|^2}, \text{ para } \gamma > 0$$

- Función de base radial gaussiana:

*Ecuación 6. Función de base radial gaussiana*

$$k(x, x') = e^{\frac{\|x-x'\|^2}{2\sigma^2}}$$

- Sigmoide:

*Ecuación 7. Sigmoide*

$$k(x, x') = tan(x \cdot x' + c), \text{ para algunos } k > 0 \text{ y } c < 0$$

En el grupo de algoritmos que utilizan *kernels* para la clasificación, el algoritmo SVM ha sido ampliamente estudiado y utilizado para el caso de imágenes HS debido a que el modelo no necesita grandes muestras de datos para generar un sistema clasificador completo. El modelo de clasificación por tanto tiene un objetivo, definir el hiperplano en dos zonas que permitan una clasificación binaria, esto lo consigue ampliando los márgenes de los bordes que existen entren las muestras que se quieren clasificar y el hiperplano que separa ambas zonas, de tal manera que permita clasificar las muestras que se encuentren cerca de la zona de decisión.[59].

El clasificador SVM permite utilizar distintos tipos de *kernels* en el cálculo del hiperplano óptimo. Así pues, en función del *kernel* de entrada que se utilice, se podrá delimitar la zona en forma de recta, curva, etc.[63], [64]. En el modelo de clasificación por SVM se representan puntos en el espacio que hacen referencia a los datos muestreados, estos se encuentran en un espacio de $N$ dimensiones en el que $N$ representa el número de clases que se encuentran en ese espacio separado por el hiperplano que es el encargado de realizar la clasificación dividendo el espacio en tantos subespacios como clases existan. La clasificación, por ende, se basa principalmente en el cálculo de hiperplanos (Figura 15.A).

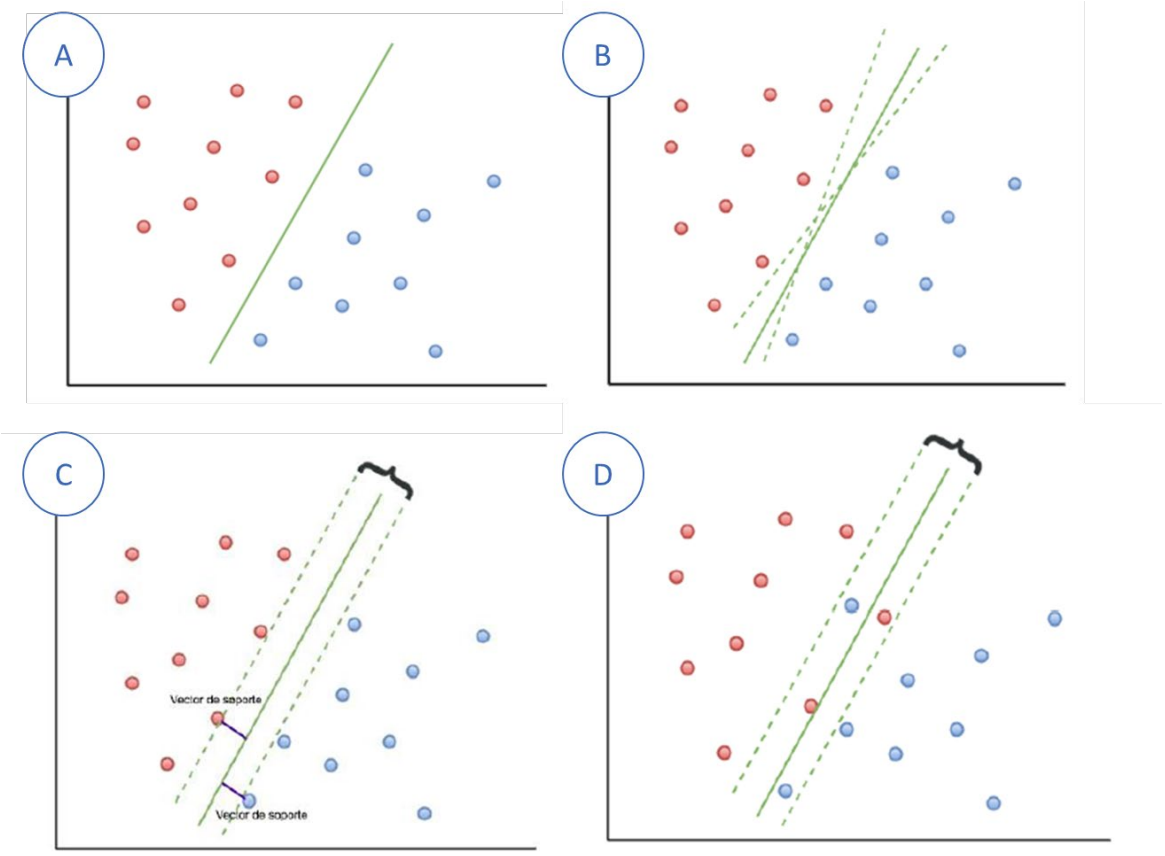



*Figura 15.Fases en la clasificación por SVM a través de un hiperplano.*

En la clasificación por hiperplano se necesita elegir de entre todas las posibilidades de hiperplano que existan (ejemplo en el caso B de la Figura 15) aquel que sea óptimo para realizar la separación de clases. En caso de que más de un hiperplano obtenga buenos resultados en la separación de clases, se escogerá el hiperplano que tenga mayor margen de seguridad. Este término define la distancia que existe entre los diferentes puntos distribuidos en el plano y los más cercanos al hiperplano. Los puntos que se encuentran en los límites del hiperplano y que son los llamados ventores de soporte. Es decir, si observamos la Figura 15.C, vemos como los puntos más cercanos al hiperplano que separa las clases son los puntos que sirven de vector soporte y por ello delimitan el margen de seguridad que existe.

En bases de datos de gran tamaño con gran cantidad de datos y clases, es posible no encontrar un hiperplano que cumpla con la separación de clases de manera óptima y, por ende, agregue error. En este tipo de casos habría que emplear un sistema clasificador que cuente con un margen de seguridad e introduzca un nuevo valor C, que representa el coste asociado a error por el que los puntos que se encuentren dentro del margen de seguridad pueden clasificarse erróneamente. Cuanto mayor sea el valor que se le dé a C, mayor probabilidad de éxito tendrá

el clasificador [64]. En la Figura 15.D, el clasificador cuenta con un margen de seguridad con valor C que reduce el error en la clasificación puesto que elimina las variantes con posibilidad de fallo.

La elección del *kernel* a usar en la clasificación y la definición de sus parámetros de entrada es una de las elecciones con mayor importancia para desarrollar el modelo clasificador. El problema de la elección de *kernel* óptimo para cada aún es un tema de estudio que no ha podido resolverse y por ello, una de las mayores desventajas del SVM, puesto que se puede llegar a perder tiempo en encontrar el clasificador idóneo [65]. Otra problemática es el tiempo y espacio, dado que el modelo tarda en procesar los datos, siendo lento y el almacenamiento que requiere para poder realizar las fases de entrenamiento y validación.

# Capítulo 3: Estado del arte

En este capítulo se sitúa el estado actual del tema de estudio, en este caso, el tratado de imágenes HS y los métodos usados para la clasificación de tumores en el diagnóstico a través de imágenes. También se mencionan brevemente los estudios concretos que se han analizado para poder llevar a cabo un posterior desarrollo del trabajo y las conclusiones de partida.

Una imagen HS por sí sola puede ser considerada como un conjunto de datos, la cual, bajo la hipótesis de que las variaciones existentes en la reflectividad de los tejidos a través de las longitudes de onda pueden aportar información para diferenciar distintos tejidos de interés. No obstante, la existencia de características o variables dependientes no correlacionadas con la señal de salida pueden introducir ruido o distorsión en los modelos, repercutiendo negativamente en el desempeño de este y dando origen a problemas relacionados con *overfitting*. Es por esto por lo que el pre-procesado y limpieza de datos o canales espectrales, es considerado una etapa fundamental, previa al modelado o entrenamiento de algoritmos predictivos. Aunque existen técnicas como el análisis de componente principal (PCA) y los auto codificadores para reducción de ruido, que permiten reducir la dimensionalidad de la señal de entrada y corregir errores en la misma, ninguno de estos algoritmos permite establecer la dependencia o contribución que posee cada variable sobre las variables dependientes.

Se establece, por ello, que la utilización de algoritmos de regresión lineal para el estudio de dependencia entre las variables dependientes e independientes del sistema aporta valor a la clasificación. A priori, una regresión lineal por sí sola no puede ser utilizada como herramienta para determinar la dependencia existente entre dos o más señales. Sin embargo, al tener definido un criterio α de significancia estadística como referencia y el estudio del $P$ Valor asociado a cada regresor, es posible determinar a través de la validación de hipótesis nula, si dicho predictor está relacionado a los cambios presentes en la variable de respuesta.

En la literatura, es posible evidenciar casos de estudios en los que se emplea un valor referencial de $\alpha = 0.05$ (intervalo de confianza del 95%), asumiendo una probabilidad de error cercana al 5%. De esta forma, es posible reducir una incógnita del sistema de ecuaciones que se usará en este trabajo, introduciendo al estudio posibles errores asociados a la metodología subjetiva empleada para la elección de dicho coeficiente. Además de establecer el valor de $\alpha$, es necesario también establecer la señal de referencia o información que se desea modelar con los datos presentes en las imágenes HS. El cual en este estudio está asociado a la caracterización de los tejidos asociados a la vascularización cerebral y cáncer.

Basado en la teoría de la información presentada por Claude E. Shannon y Warren Weaver a finales de la década de los años 1940. Es posible establecer que la información está constituida única y exclusivamente como los mensajes o estados que cambian el estado del conocimiento del sujeto o sistema receptor. De esta forma, la caracterización de dichos tejidos puede ser modelada como una señal binaria, que modela los estados de presencia o no de cada tipo de tejido en estudio.

En la literatura también se han realizado estudios sobre distintos algoritmos, concretamente en el estudio [66] se plantea como objetivo comparar el rendimiento de diferentes algoritmos de aprendizaje automático supervisado. Para ello se seleccionaron 336 artículos que utilizaban más de un algoritmo de aprendizaje automático supervisado para la predicción de enfermedades. De 281 artículos, sólo 155 utilizaron uno de los algoritmos de aprendizaje automático supervisados considerados en el estudio. De este estudio se obtuvo que el algoritmo SVM se aplica con mayor frecuencia (29 estudios). En el estudio [66] se observa que el algoritmo SVM muestra una precisión superior en la mayoría de las ocasiones para enfermedades cardíacas, diabetes y enfermedad de párkinson.

Analizando estudios para el diagnóstico del cáncer y sus respectivas patologías, se encuentran en [67] que muchas de las investigaciones se centran en el diagnóstico y el pronóstico automatizados y los métodos de aprendizaje automático tienen el potencial de producir herramientas clínicas que proporcionen a los oncólogos mayores conocimientos sobre los resultados probables de las distintas vías de gestión y tratamiento de la enfermedad. Dentro de [68] en el estudio de la clasificación de las células cancerosas, como benignas o malignas, el trabajo, que compara 4 algoritmos supervisados obtiene como resultado que el algoritmo SVM da mejores resultados en la clasificación. Finalmente, en [69] se han estudiado algoritmos de aprendizaje supervisado y semisupervisado en el diagnóstico del cáncer.

Los resultados de todos los modelos obtienen como conclusión que las precisiones de los algoritmos no supervisados están muy próximas a las de los algoritmos supervisados, definiendo que, aunque los algoritmos supervisados tienen una exactitud del 98%, los no supervisados son totalmente capaces de realizar diagnóstico de tipo de tumor con precisiones mayores a 90%. En resumen, en diagnóstico de tumores y patologías adyacentes al cáncer, los algoritmos que obtienen mejores resultados en la aplicación según la literatura es el SVM.

# Capítulo 4: Materiales y métodos

## 4.1. Materiales

Para el desarrollo de este TFG se han utilizado distintos materiales para poder lograr los objetivos planteados para este proyecto. En este capítulo se presentan las herramientas necesarias para poder desarrollar todos los pasos que se presentarán posteriormente en la metodología seguida en este proyecto.

Partimos inicialmente de elementos básicos como son un ordenador con capacidad para realizar grandes computaciones de datos, puesto que las imágenes HS son de gran de tamaño y requieren de bastante capacidad del ordenador. Otro elemento básico es contar con un software que permita la herramienta de programación, en este caso nuestra herramienta es Anaconda 3-2021 para Windows con lenguaje Python 3.9.

El elemento primordial del que parte este TFG es de la base de datos desarrollada por el proyecto HELICoiD, la cual creó una base de datos de imágenes HS del cerebro humano in vivo. Se han obtenido de 22 pacientes diferentes en el Hospital Universitario Doctor Negrín hipercubos con información sobre tejidos. Los tipos de tumores captados en este estudio incluye tanto tumores cerebrales primarios como secundarios. Este tipo de imágenes se obtienen a través de cámaras HS con medición píxel a píxel, la intensidad de la luz para una longitud de onda en concreto. Los sensores de esta cámara son los encargados de recopilar toda la información como un conjunto de imágenes [70]. Las imágenes fueron obtenidas tomando medidas secuenciales de los espectros línea a línea de la región del cerebro estudiado. Se realizaron con dos cámaras HS como las que se pueden observar en la figura usando rango VNIR 400 nm- 1000nm en una de ellas y NIR 900nm-1700nm en otra [70].

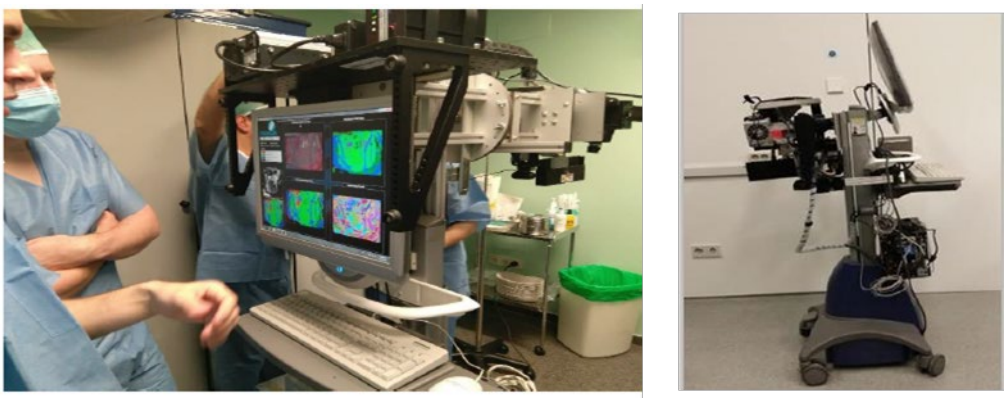


*Figura 16. Cámaras para la extracción de imágenes HS.*

Utilizando la información proporcionada por los médicos se desarrolla la base de datos de imágenes, en la cual el equipo de investigación formado para el proyecto HELICoiD etiqueta algunos píxeles de cada imagen, utilizando una herramienta de etiquetado, *Labelling Tool,* desarrollada en MATLAB, con el fin de generar el etiquetado que constituirá el punto de partida para el entrenamiento clasificador. En este punto del proyecto para poder realizar los objetivos es necesario partir de un análisis de la base de datos exhaustiva, ya que conforman la base de este proyecto. La Tabla 1 detalla el número total de píxeles etiquetados para cada imagen. El conjunto de datos se ha reducido a 4 clases diferentes (tejido normal, tejido tumoral, vasos sanguíneos y *background*).

*Tabla 1. Resultados del análisis de la base de datos*

| Clase | | | #Etiquetado pixel base de datos |
|---|---|---|---|
| **Normal** | | | 117.242 |
| **Tumor** | **Primario (G-IV)** | **GBM** | 16.449 |
| **Vaso sanguíneo** | | | 73.874 |
| ***Background*** | **Meningiomas** | | 434 |
| | **Genérico *Background*** | | 185.684 |
| **Total:** | | | **393.683** |

Previo al análisis de resultados obtenidos durante la fase de optimización y entrenamiento, es necesario conocer la distribución y estructura de los tejidos contenidos en las imágenes HS. Este análisis mostrado en la Tabla 2 es una extensión de la Tabla 1.

*Tabla 2. Distribución de identificadores en imágenes HS*

| Identificador de Tejido | Clasificación Principal | Subclasificación | Detalle del Identificador | Cantidad de Muestras (Pixeles) | Representatividad | Grupo |
|---|---|---|---|---|---|---|
| 100 | Normal | Sin definir | Sin definir | 117242 | 31,08% | Normal |
| 200 | Tumor | Primario-GIV | Gliobastoma | 12641 | 3,35% | Tumor |
| 220 | Tumor | Primario -GIII | Oligodendroglioma | 1844 | 0,49% | Tumor |
| 250 | Tumor | Secundario | Estómago | 1964 | 0,52% | Tumor |
| 300 | Otros | Sangre | Generico | 14073 | 3,73% | Vaso |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | Sanguíneo |
| 301 | Otros | Sangre | Venas-Vasos sanguíneos | 0 | 0,00% | Vaso Sanguíneo |
| 302 | Otros | Sangre | Arterias-Vasos sanguíneos | 0 | 0,00% | Vaso Sanguíneo |
| 303 | Otros | Sangre | Sin definir-Vasos sanguíneos | 43356 | 11,49% | Vaso Sanguíneo |
| 320 | Otros | Meningioma | Dura-Mater | 434 | 0,12% | Otros |
| 331 | Otros | Externo | Hueso | 0 | 0,00% | Otros |
| 400 | *Background* | BG-Genérico | BG- Genérico | 68059 | 18,04% | Otros |
| 410 | *Background* | BG-Genérico | Genérico | 74586 | 19,77% | Otros |
| 411 | *Background* | BG-Genérico | Genérico | 19488 | 5,17% | Otros |
| 412 | *Background* | BG-Genérico | Genérico | 0 | 0,00% | Otros |
| 422 | *Background* | Elemento quirúrgico | Marcador | 23551 | 6,24% | Otros |

En la Tabla 2 se presentan la totalidad de muestras o píxeles obtenidos de las imágenes HS establecidas para el estudio, junto a los identificadores de tejidos previamente definidos por el equipo médico. Para cada muestra se presentan 3 niveles de clasificación, identificando una clase primaria de tejido y 2 niveles de subclasificación para especificar los detalles del tejido asociado a ese marcador. Adicionalmente, se ha establecido en dicha tabla una columna "Grupo" que muestra el identificador al que ha sido recategorizado cada tipo de tejido con la finalidad de ser asociados a los tejidos a caracterizar en este estudio.

Finalmente, al etiquetar los pixeles presentados en la Tabla 2 de acuerdo con las nuevas clases de estudio se obtuvo la distribución obtenida en la Figura 17, en la cual, es posible evidenciar el desbalance natural que pueden presentar las clases de tejido contenido las imágenes HS. El tejido tumoral y los vasos sanguíneos constituyen menos del 20% de todos los datos de entrenamiento, debido a su baja presencia en la constitución de los tejidos cerebrales. Estos datos validan la necesidad de aplicar una técnica de análisis que presente mejor desempeño ante clases desbalanceadas, donde muestras como el tejido asociado a tumor representa apenas 1 de cada 20 muestras en estudio.

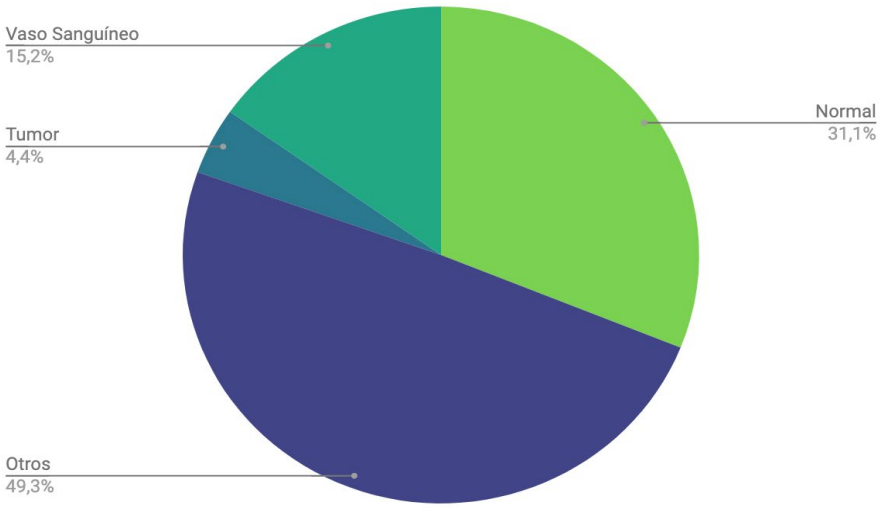



*Figura 17. Distribución de las clases de interés en imágenes HS de la base de datos.*

A partir de la información recopilada y el etiquetado de datos, se comienza la identificación de los tejidos que caracterizan los tumores y vascularización en los pacientes recopilados a través de modelos supervisados. Seguidamente, se complementa el análisis con otro estudio utilizando algoritmos no supervisados para poder definir dentro del tejido asociado a la vascularización cual está asociado a venas y cual, a arterias, siendo de gran utilidad parar reconocer el grado de oxigenación que posee el tumor en cada caso.

## 4.2. Discriminación entre tumor y vaso sanguíneo

### 4.2.1. Metodología

En este aparado se explicará toda la metodología que se ha seguido para poder llevar a cabo el trabajo y cumplir los objetivos que nos planteamos en el anteproyecto de este TFG. Para empezar, la base de datos HELICoiD, además de presentar diversas imágenes de los casos de estudio, adiciona datos reales de etiquetas generadas por especialistas sobre cada tipo de tejido, constituyendo este el conjunto de datos de la fuente de la verdad para esta etapa de la investigación. Estas etiquetas constituyen la base para generar las señales que permitirán la detección de cada tipo de tejido y que ya han sido preprocesadas y estudiadas a la hora de comenzar a desarrollar el trabajo.

Para definir un diagrama de flujo de trabajo que refleje el modelo que se ha diseñado para poder lograr el objetivo de distinguir entre los tejidos tumor y vaso sanguíneo a través del algoritmo supervisado SVM es necesario considerar dos aspectos fundamentales del comportamiento que

caracterizan a los algoritmos supervisados. Haciendo referencia a su susceptibilidad ante el ruido y que los resultados obtenidos en solo un primer ciclo de entrenamiento no reflejan la tendencia de clasificación real del modelo, puesto que estos datos son de carácter probable, es por esto por lo que un cambio en la distribución de los datos dentro de los grupos de entrenamiento puede ocasionar variaciones en los resultados obtenidos para un mismo modelo de entrenamiento.

La base de datos de HELICoiD proporciona imágenes HS que aportan datos con 128 canales de profundidad distribuidos de forma homogénea en el rango VNIR y NIR. No obstante, mayor cantidad de canales no significa un mayor aporte de información, ya que la información depende estrictamente del receptor y la capacidad que tenga para generar un cambio y ser procesada por este, por ello, solo es posible considerar como información los canales que efectivamente aporten información relevante. Basado en este concepto, es posible establecer que la fuente de datos aporta hasta 128 características por píxel, los mismos, no necesariamente aportan información al sistema, repercutiendo de forma negativa en el modelo, ya que, al no contener información relevante, aporta ruido al modelo.

Para lograr este objetivo intermedio y validar que los datos utilizados para el entrenamiento del modelo supervisado aporten información al sistema, es necesario establecer un modelo de discriminación, en el cual, se posea como finalidad validar la hipótesis de que el receptor haya cambiado o no su estado de conocimiento sobre la información contenida en los datos. De esta forma, si al eliminar características del mensaje, el receptor no presenta variaciones en su desempeño, será posible afirmar que los datos eliminados del sistema corresponden a ruido y no dan información importante al receptor.

En este punto del procedimiento es necesario establecer una forma eficiente de análisis que permita discriminar qué características de los datos aportan información al modelo de entrenamiento, reduciendo lo más posible la cantidad de pruebas o conjuntos de datos utilizados para entrenar y probar el modelo. Para esto se decidió utilizar un modelo de regresión logística y considerar como datos de referencia los coeficientes de P valores obtenidos en cada modelo resultante. De esta forma, al utilizar un valor de referencia conocido como significancia estadística como umbral de segmentación, será posible despreciar las longitudes de onda que no aporten información en la clasificación de los tejidos en estudios.

La regresión logística incorpora a la solución una herramienta fundamental para discriminar qué longitudes de onda o características de los datos son de interés para el receptor y cuáles se

consideran ruido, evitando la necesidad de evaluar todas las posibles combinaciones de datos de entrada en el sistema. Sin embargo, este proceso incorpora dos nuevos requisitos, el primero asociado a establecer cuáles serían las señales de entrada y salida del modelo de regresión logística y cuál sería el valor de significación estadística adecuado para segmentar los datos. Dado que el modelo en estudio pretende distinguir los tejidos tumorales y los vasos sanguíneos del resto de los tejidos, las etiquetas presentadas en la fuente de verdad de la base de datos se establecen como las etiquetas utilizadas durante el ajuste de la regresión logística. De este modo, el modelo buscará asociar los marcadores de los tejidos con las 128 características presentes en los datos de entrenamiento.

Para considerar al modelo de regresión logística como modelo de discriminación adecuado para el proceso, es necesario establecer el α (nivel de significación) adecuado para discriminar el ruido y la información en el sistema. La mayoría de las literaturas y autores asumen un valor de α=0,05 (intervalo de confianza del 95%) [71], que se considera en muchas investigaciones como el valor de referencia apropiado para el análisis de datos. Y se establece en muchos casos empíricamente durante la investigación, asumiendo que el impacto de tales datos tiene un error mayor al 5%, su contribución al modelo sería insignificante. Establecer la posibilidad de introducir errores de procedimiento de origen humano en la investigación.

En esta investigación, se decidió no establecer de forma empírica el umbral de segmentación de los datos, utilizando una estrategia combinada para evaluar el impacto que poseen los cambios sobre nuestro receptor en estudio. Se le ha denominado estrategia combinada debido a que plantea la utilización del modelo de entrenamiento supervisado como validador de las características seleccionadas de los datos. De esta forma, al aplicar un ajuste de regresión logística, se seleccionan los canales que posean P valores menores al α de significancia estadística establecida, finalmente los datos seleccionados son puestos a prueba por el modelo durante la etapa de entrenamiento y prueba. Si los canales eliminados de la señal de entrada solo contenían ruido, será posible observar que el desempeño del modelo no se ve afectado o mejora debido a la eliminación del ruido en los datos, en caso contrario, al eliminarse canales de información, se observará un deterioro en el rendimiento del modelo. Dicho planteamiento, establece de forma axiomática que el valor adecuado de α será aquel que permita remover la mayor cantidad de canales de los datos HS, sin repercutir negativamente sobre el desempeño del modelo de clasificación supervisada.

Junto a la lógica de procesamiento de datos, denominada estrategia combinada, es necesario añadir una etapa de preprocesado abocada a transformar las imágenes en datos que puedan ser consumidos e interpretados por el algoritmo de análisis, conformando finalmente este conjunto de subprocesos, el algoritmo de procesamiento necesario para entrenar y optimizar un modelo de entrenamiento supervisado de tipo SVM. De esta forma y con esta premisa, se desarrolla el siguiente diagrama de flujo que permitirá seguir los pasos llevados a cabo en esta parte del trabajo distinguen 3 módulos constituidos por la etapa de preprocesado (caja verde), procesado general (resto del diagrama) y la fase de validación (caja roja).

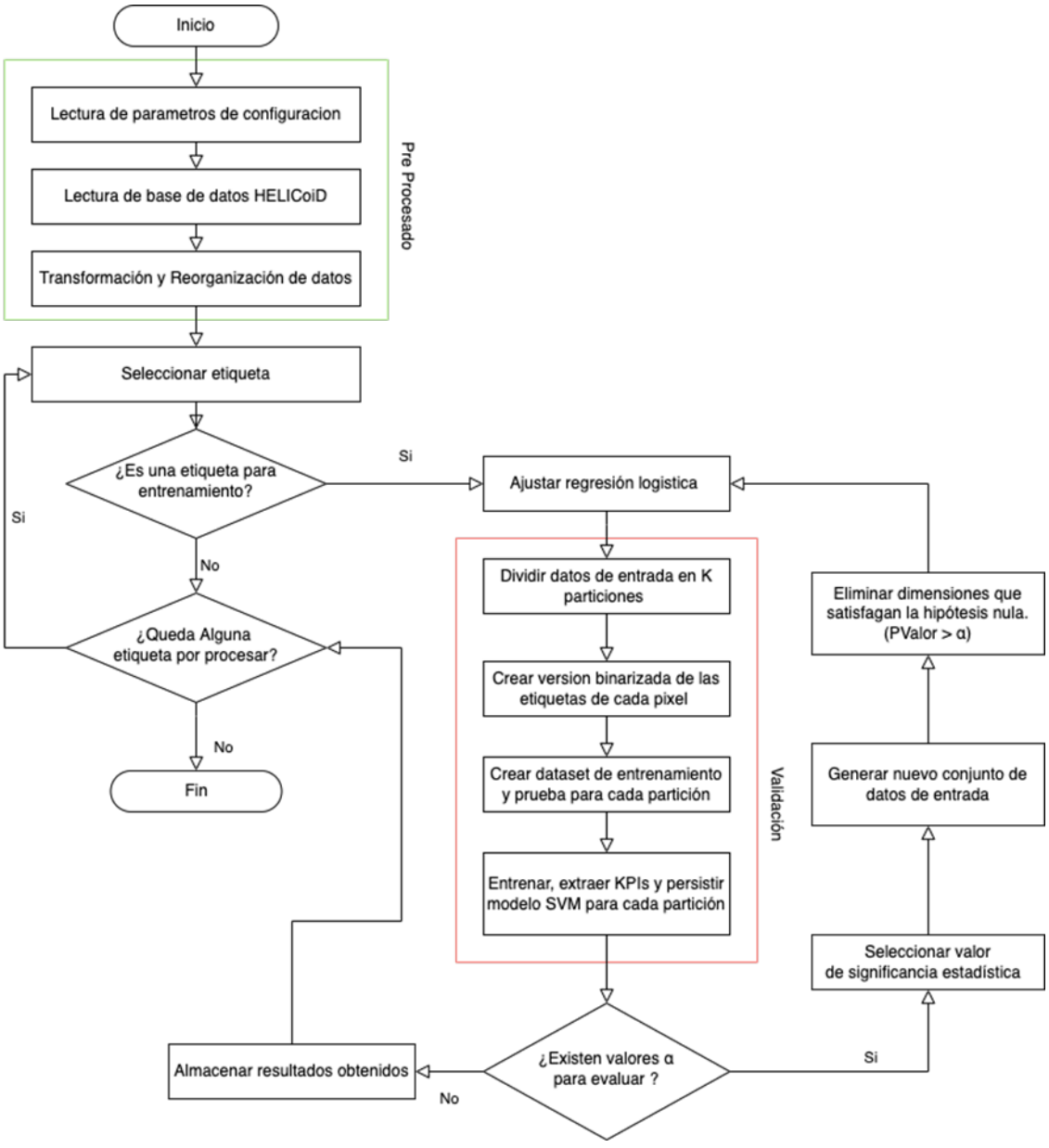



*Figura 18. Diagrama de flujo del sistema de trabajo para la discriminación entre tumor y vaso sanguíneo.*

El proceso general, inicia a través de la etapa de preprocesado, la cual posee como finalidad leer las imágenes HS y transformarlas en registros que puedan ser procesados por el modelo de aprendizaje supervisado y la regresión logística. Para esto el sistema de preprocesado utiliza un archivo de configuración, establecido con la finalidad de poder recategorizar etiquetas provenientes originalmente de la fuente de la verdad de la base de datos. Esto permite eliminar la sobre especificidad de algunas etiquetas y agruparlas de acuerdo con el objetivo de esta investigación. Al observar la Tabla 2, bajo la columna Grupo, se presentan las agrupaciones establecidas con la finalidad de detectar tejido tumoral y vasos sanguíneos en las imágenes, reagrupando 15 marcadores de tejidos en 4 grupos principales, 2 de interés y 2 que caracterizan otros tipos de tejidos predominantes en las imágenes HS.

Al conocer la configuración utilizada es necesario iniciar la etapa de carga y transformación y datos que alimentará la etapa principal del proceso. Para esto, el sistema va consumiendo y procesando cada imagen con la finalidad de almacenar solo información necesaria de cada registro, constituidos en su totalidad por las imágenes HS y sus correspondientes fuentes de la verdad previamente etiquetadas. No obstante, la estructura de datos que componen dichos conjuntos de datos no puede ser consumidos directamente por los modelos de clasificación, por lo cual, es necesario reorganizarlos para ser usados en la etapa de análisis. La Figura 19 establece el modelo de reorganización de datos diseñado para convertir los datos en tensores consumibles por el modelo de clasificación y presenta la fase final de transformación contenida en la etapa de preprocesado. Esta posee como finalidad reorganizar los datos en una estructura vectorial que pueda ser utilizada como datos de entrenamiento en las etapas siguientes, para esto, el sistema transforma las imágenes HS de dimensión: ancho $A$, Alto $H$ y profundidad $P = (A * H\ pixeles)$, en un tensor de longitud $h = (H * A)$ manteniendo la profundidad de cada punto de las $N$-dimensiones. Por lo que es posible reorganizar los datos contenidos en ambos planos, siempre que se mantenga la asociación 1:1 entre píxel y marcador, permitiendo de esta forma convertir un conjunto de imágenes en dos listas de datos, un constituida por los pixeles y otra con los marcadores de cada uno de estos píxeles en el orden correspondiente. De esta forma, la etapa de preprocesador culmina al suministrar a la etapa de análisis los tensores de entrenamiento reorganizados y recategorizados en función de los marcadores establecidos dentro de la configuración.

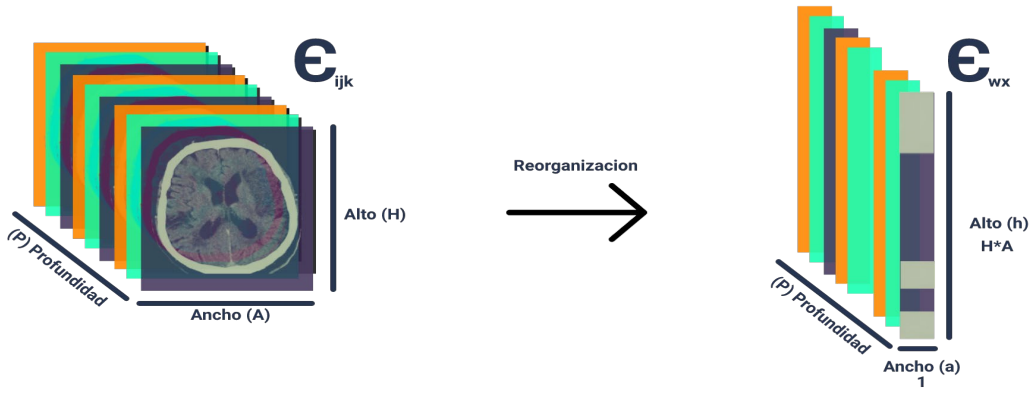



*Figura 19. Transformación de datos de entrada.*

El modelo de entrenamiento supervisado de tipo SVM y el ajuste por regresión logística para el análisis de la significancia son aspectos principales de la etapa de procesado y validación. Subsecuentemente, la Tabla 3 expone los intervalos de confianza utilizados para la extracción de las características de interés contenidos en los datos multiespectrales. Los mismos están constituidos por un umbral de referencia α=1, que constituye el caso de entrenamiento con las 128 características del tensor, los intervalos de 90,95 y 99% de confianza, junto a valores intermedios entre los puntos de interés. Finalmente, el proceso de análisis tomará de referencia cada valor de α para generar un subgrupo de características para el entrenamiento del modelo SVM.

*Tabla 3. Umbrales de filtrado para los P Valores de la regresión logística.*

| **Umbral α** | 1 | 0.2 | 0.1 | 0.075 | 0.05 | 0.03 | 0.02 | 0.01 |
|---|---|---|---|---|---|---|---|---|
| **Intervalo de confianza** | 0% | 80% | 90% | 92.50% | 95% | 97% | 98% | 99% |

De esta forma, se establecen los parámetros del sistema asociados a la regresión logística y la evaluación de los $P$ valores, no obstante, para establecer la estrategia de entrenamiento y validación de los resultados obtenidos para el modelo SVM, es necesario considerar la naturaleza de los datos de entrenamiento. La Figura 17, presenta la distribución de los tejidos en función de los marcadores empleados para este estudio, en el cual, es posible observar el desbalance natural que poseen los tejidos. Este desbalance es generado principalmente por la distribución natural que suelen tener los tejidos, presentando una predominancia del tejido saludable por sobre las otras clases. Entre las cuales, los tejidos tumorales y vasos sanguíneos son los de menor presencia. Este hecho hace necesario establecer una estrategia de

entrenamiento que mitigue el efecto del desbalance de clases sobre rendimiento de modelo entrenado. Para esto, se planteó la implementación de un esquema uno contra el resto (One-vs-Rest o OVR)[72]. Dicha estrategia permite maximizar los resultados del modelo ante conjuntos de datos desbalanceados y mejora el rendimiento de algoritmos como los modelos SVM debido a que estos han sido diseñados de forma nativa para clasificación binaria.

Para la etapa de validación es necesario establecer una metodología que permita mejorar la precisión del rendimiento del modelo entrenado, estableciendo a la validación cruzada de tipo *K Fold* estratificada como la técnica empleada para validar los resultados obtenidos por el modelo SVM. La cual consiste en tomar los conjuntos de datos de entrenamiento y dividirlos en $K$ particiones conservando la misma distribución de clases en cada subgrupo de muestras. De esta forma es posible evaluar $K$ veces el modelo de entrenamiento elegido, esta técnica permite estimar no solo el desempeño del sistema sino también la precisión de dichas medidas.

Finalmente, el proceso general el cual inicia posterior al preprocesado de datos contempla la evaluación de cada etiqueta en estudio (tumor y vaso sanguíneo) por separado con la finalidad de obtener el mejor desempeño del modelo para cada etiqueta. De esta forma, el sistema consume los tensores suministrados por la etapa de preprocesado y aplica un ajuste por regresión logística sobre todos los datos, obteniendo los $P$ valores a evaluar, con estos $P$ valores de referencia, el proceso ejecuta los ciclos de discriminación y entrenamiento del modelo supervisado para cada valor de α. Durante la fase de discriminación el proceso elimina las características del tensor de entrada que no satisfacen la significancia estadística establecida obteniendo un tensor de menor dimensionalidad a la inicial ($N$ < 128). Estos datos son empleados como datos de entrenamiento para la máquina de soporte vectorial. Durante la etapa de entrenamiento el modelo es sometido a 20 iteraciones con particiones de datos igualmente distribuidos, en el cual, al momento de entrenar se reemplazan las etiquetas como 1 en caso de que el marcador corresponda a la etiqueta de estudio o 0 en caso de que no correspondan, y posteriormente se almacenan las métricas de rendimiento asociadas al modelo generado. Este ciclo de ejecución es realizado secuencialmente con cada uno de los valores de α, para posteriormente ser evaluados y seleccionar en conjunto de características que mejor rendimiento presenta sobre los datos.

Para elegir el modelo que mejor rendimiento presente, se evaluarán las clasificaciones en función del umbral α de segmentación con 4 métricas principales: precisión, exhaustividad, exactitud y valor F1 [73].

- Precisión: mide la calidad del modelo SVM. Contesta a la pregunta de qué porcentaje de respuestas de las obtenidas se asocian verdaderamente a los tejidos de estudio [73] .

*Ecuación 8. Cálculo de precisión*

$$precisión = \frac{Verdadero\ Positivo}{Verdadero\ Positivo + Falso\ Positivo}$$

- Exhaustividad: mide la cantidad capaz de identificar. Contesta a la pregunta de qué porcentaje de los píxeles totales es capaz de clasificar el modelo en los tejidos [73].

*Ecuación 9. Cálculo Exhaustividad*

$$exhaustividad = \frac{Verdadero\ Positivo}{Verdadero\ Positivo + Falso\ Negativo}$$

- Valor F1: combina los datos de precisión y exhaustividad en un único valor, calculando la media entre ambos valores [73].

*Ecuación 10. Cálculo valor F1*

$$F1 = 2\frac{precisión \cdot exhaustividad}{precisión + exhaustividad}$$

- Exactitud: mide porcentaje de casos que el modelo ha supervisado. Aunque no funciona correctamente en clasificaciones desbalanceadas como ocurre con estos tejidos, por lo que solo se tomará como un valor de métrica, pero no de referencia [73].

*Ecuación 11. Cálculo exactitud*

$$exactitud = \frac{Verdadero\ Positivo + Verdadero\ Negativo}{Verdadero\ Positivo + Verdadero\ Negativo + Falso\ positivo + Falso\ Negativo}$$

Con los resultados obtenidos de todas estas métricas y una vez seleccionado el umbral de referencia que tomaremos como válido, se realizará el estudio de las longitudes de onda que son características de los tejidos tumor y vaso sanguíneo, logrando un modelo optimizado y supervisado.

## 4.3. Discriminación entre arteria y vena

### 4.3.1. Metodología

En esta segunda parte del TFG, el objetivo principal es encontrar un modelo de clasificación que permita definir dentro de los tejidos asociados a vasos sanguíneos cuales están asociados a venas y cuales a arterias. Este objetivo, como se ha nombrado a lo largo de todo el presente trabajo permitiría conocer el grado de angiogénesis y oxigenación que presentan los tumores, información relevante en etapas diagnósticas y de seguimiento de lesiones tumorales.

La base de datos HELICoiD , como fuente de datos de información posee gran cantidad de descriptores para clasificar los tejidos presentados en las imágenes HS, estableciendo de forma jerárquica 3 niveles de agrupación que permiten variar el nivel de especificidad con el que son asociados los datos de cada imagen. No obstante, como se puede observar en los datos ya recogidos en la Tabla 2 dentro de este mismo capítulo en la sección de *Materiales*, al estudiar las clasificaciones establecidas para los tejidos de vasos sanguíneos, no es posible determinar qué tejidos están asociados a venas y arterias. Por lo cual, a la hora de definir el diagrama de trabajo, se asume que no es posible utilizar una metodología basada en modelos supervisados para generar un modelo capaz de distinguir estos elementos de interés ya que como se define en la sección *Algoritmos* del *Capítulo 2,* para los algoritmos supervisados es necesario partir de una base de datos previamente etiquetada. Este escenario, es el punto de partida de esta segunda parte del TFG en la que se plantea la posibilidad de emplear modelos no supervisados para entender el orden natural en que los tejidos se asocian entre ellos y posteriormente establecer que tan correlacionado puede estar dichas agrupaciones con los tejidos sanguíneos.

Son diversos los métodos que pueden ser propuestos para la agrupación de datos, los cuales, dependiendo de su diseño pueden o no ser configurados para generar una cantidad específica de agrupaciones. Aunque existen diversos métodos de agrupación no supervisada, donde el número de agrupaciones finales varía en función de los umbrales de segmentación, en esta investigación solo se evaluarán métodos en los cuales pueda ser preestablecido la cantidad de agrupaciones a generar. Entre los algoritmos a estudiar se presentan *K-Means, Spectral Clustering, Gaussian Mixture y Agglomerative Clustering* que son métodos de agrupación estudiados en la literatura y presentados teóricamente a lo largo de este trabajo, y que permiten distribuir los datos de entrenamiento entre $N$ grupos predefinidos Como planteamiento inicial,

se establece generar 2 grupos basados en los tejidos clasificados como vasos sanguíneos para evaluar posteriormente si estas agrupaciones permiten efectivamente diferenciar venas y arterias o si responden a otro tipo de características de las muestras asociadas a vasos sanguíneos. En la Figura 20, es posible observar cómo existe 1 solo marcador presentado en la fuente de verdad de cada imagen, asociado a vasos sanguíneos. De esta forma es posible evidenciar 2 aspectos relevantes de los datos de estudio, entre los cuales se establece que no hay pixeles previamente diferenciados entre venas y arterias dentro de las imágenes HS y que a su vez los tejidos etiquetados como vasos sanguíneos no poseen bordes definidos pudiendo contener muestras de tejidos que no corresponden directamente a vasos sanguíneos. Este será el punto de partida para esta segunda parte, ya que, al no tener etiquetas definidas para arterias y venas, la búsqueda de un nuevo modelo de clasificación da paso a los algoritmos no supervisados.

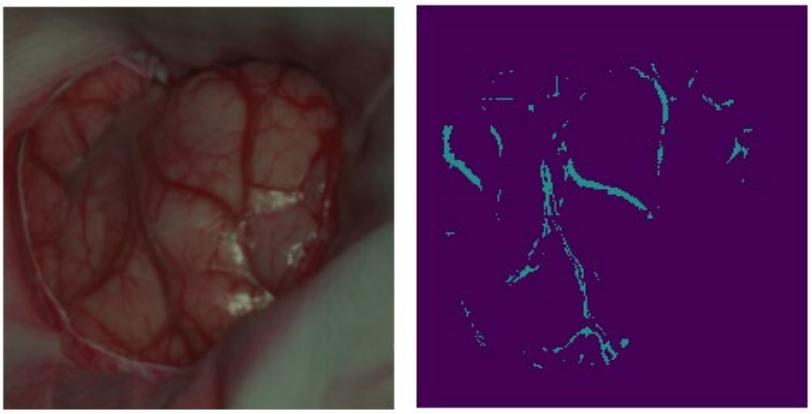


*Figura 20.Ejemplo de identificadores de tejido asociados a vasos sanguíneos.*

La Figura 21 presenta el diagrama de flujo establecido para realizar la evaluación de los modelos de entrenamiento no supervisados y la necesidad de realizar una elección del modelo con mejor rendimiento en el análisis de los datos en estudio. Como se puede ver (Figura 21), se tienen de nuevo dos etapas principales. La primera etapa y remarcada en color verde, consiste en sacar de la base de datos, fuente principal del trabajo, la etiqueta que se pretende subclasificar, vasos sanguíneos. No obstante, debido a la gran dimensionalidad presentada en las imágenes HS se establece aplicar una etapa de reducción dimensional con la finalidad de disminuir la cantidad de variables independientes que consume el modelo de agrupación. Como técnica de reducción dimensional se establece aplicar un PCA seleccionando las $N$ componentes principales que expliquen al menos el 95% de la varianza.

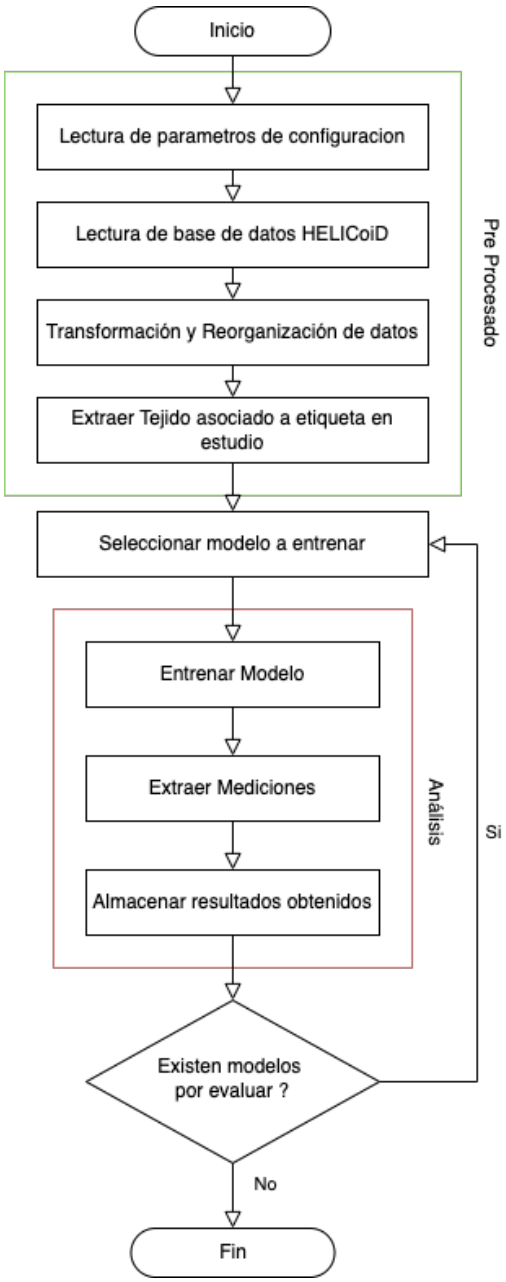



*Figura 21. Diagrama de flujo del sistema de análisis para la elección del modelo de entrenamiento.*

La segunda etapa seleccionada por el rectángulo rojo consiste en el análisis y selección de todos los algoritmos no supervisados que se han elegido en el estudio para conocer el óptimo para realizar el procesado de imagen. A la hora de evaluar el desempeño de un modelo de clasificación no supervisado, es importante no solo medir la eficacia de este a la hora de separar y modelar las clases, sino también, considerar cuán eficiente es el mismo a la hora de procesar los datos de nuestro estudio. Es por esto por lo que los indicadores elegidos para esta fase de la investigación contemplan métricas asociadas a la caracterización de las clases y el consumo de recursos durante el proceso de evaluación.

En esta segunda etapa de selección del modelo utilizaremos la forma de medida denominada *silhoutte* o llamada en español como silueta. La *silhoutte* es un método de interpretación y validación de la coherencia dentro de los clasificadores por grupos. Por lo tanto, corresponde como el indicador de eficacia más apropiado para esta etapa de análisis, ya que la misma permite cuantificar que tan segmentadas o definidas están las clases. En contraposición, como indicadores de eficiencia se estableció la medición del tiempo de entrenamiento y el consumo máximo de memoria RAM (*Random Access Memory)* muestras utilizadas durante el entrenamiento.

Establecido el algoritmo de agrupación en estudio, la siguiente parte del modelo se basa en establecer la correlación existente entre las clases generadas a partir del algoritmo de clasificación seleccionado y los tejidos asociados a venas y arterias. Proceso en el cual, se necesitará contar con validación externa por parte de profesionales médicos puesto que son los únicos que pueden confirmar si la clasificación es o no correcta.

# Capítulo 5: Resultados

## 5.1. Resultados obtenidos en la discriminación entre tumor y vasos sanguíneos

Previo al procesado de los datos fue necesario establecer el *kernel* a utilizar por el modelo de entrenamiento con la finalidad de obtener el mayor desempeño posible durante la etapa de validación y entrenamiento del modelo. Para dicho análisis, se procedió a ejecutar el flujo presentado en la Figura 18. Estableciendo una sola partición con la totalidad de los datos, emulando el entrenamiento del modelo sin aplicar validación cruzada en análisis de los resultados.

Al etiquetar los pixeles presentados en la Tabla 2 de acuerdo con las nuevas clases de estudio se obtuvo la distribución obtenida en la Figura 17, en la cual, es posible evidenciar el desbalance natural que pueden presentar las clases de tejido contenido las imágenes HS. El tejido tumoral y los vasos sanguíneos constituyen menos del 20% de todos los datos de entrenamiento, debido a su baja presencia en la constitución de los tejidos cerebrales. Estos datos validan la necesidad de aplicar una técnica de análisis que presente mejor desempeño ante clases desbalanceadas, donde muestras como el tejido asociado a tumor representa apenas 1 de cada 20 muestras en estudio.

En la Figura 22 se presentan los resultados obtenidos bajo el algoritmo SVM establecido con anterioridad. Aunque dicho proceso no permite establecer una medición exacta del rendimiento del modelo entrenado, si es posible determinar la tendencia que posee cada *kernel* para modelar y distinguir las etiquetas de vaso sanguíneo y tumor del resto de los tejidos contenidos en las imágenes HS. De esta forma, es posible observar en la figura que de los 3 *kernels* estudiados en esta etapa (lineal, polinomial y función de base radial), el *kernel* polinomial presenta mayor rendimiento en todos los ciclos de prueba. Al comparar el rendimiento promedio de cada *kernel* en dichas configuraciones, podemos observar una mejora nominal del 0.32% y 3.03% del *kernel* polinomial sobre el *kernel* de función de base radial y el *kernel* lineal respectivamente. Es a partir de los resultados presentados anteriormente, que se estableció la utilización de un *kernel* polinomial para el análisis y filtrado de los canales asociados a los tejidos de estudio.

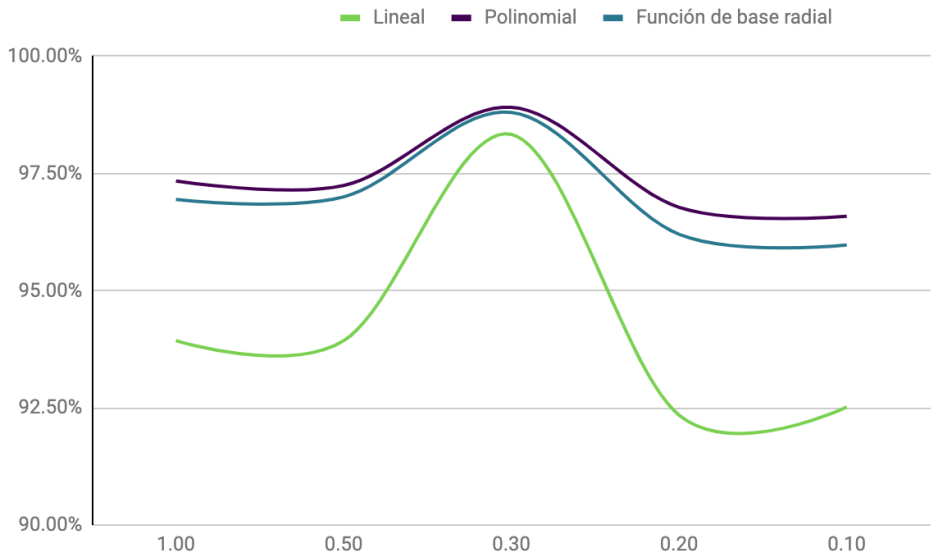



*Figura 22.Exactitud de predicción promedio para las etiquetas de tumor y vaso sanguíneo para los canales de información en función de la significancia estadística.*

Las figuras a continuación (Figura 23 y Figura 24) presentan los resultados obtenidos de la etapa de optimización y filtrados de los canales espectrales de las imágenes de estudio bajo el esquema de análisis presentado en el algoritmo de procesamiento. En dichas gráficas es posible observar como el desempeño del modelo de clasificación varía en función del umbral de segmentación de canales basado en la significancia estadística. Esto puede ser interpretado a su vez, como el desempeño que presentan los modelos de clasificación al eliminar los canales espectrales que no aportan información a la detección de los tejidos.

Para el estudio de desempeño del modelo, se estableció la utilización del Valor F1, la exhaustividad y precisión junto a la exactitud con la finalidad de tener información más precisa del rendimiento del modelo. En primera instancia al evaluar el desempeño del modelo, considerando exclusivamente la exactitud de clasificación, se pudo observar que el sistema logró mejorar su exactitud sobre la etiqueta de vaso sanguíneo a medida que se eliminaban canales espectrales que pudieran generar ruido en la señal a través de la regresión logística. No obstante, para el tejido asociado a tumor es posible observar como la exactitud promedio del modelo se mantiene invariante a los cambios en la señal de entrada.

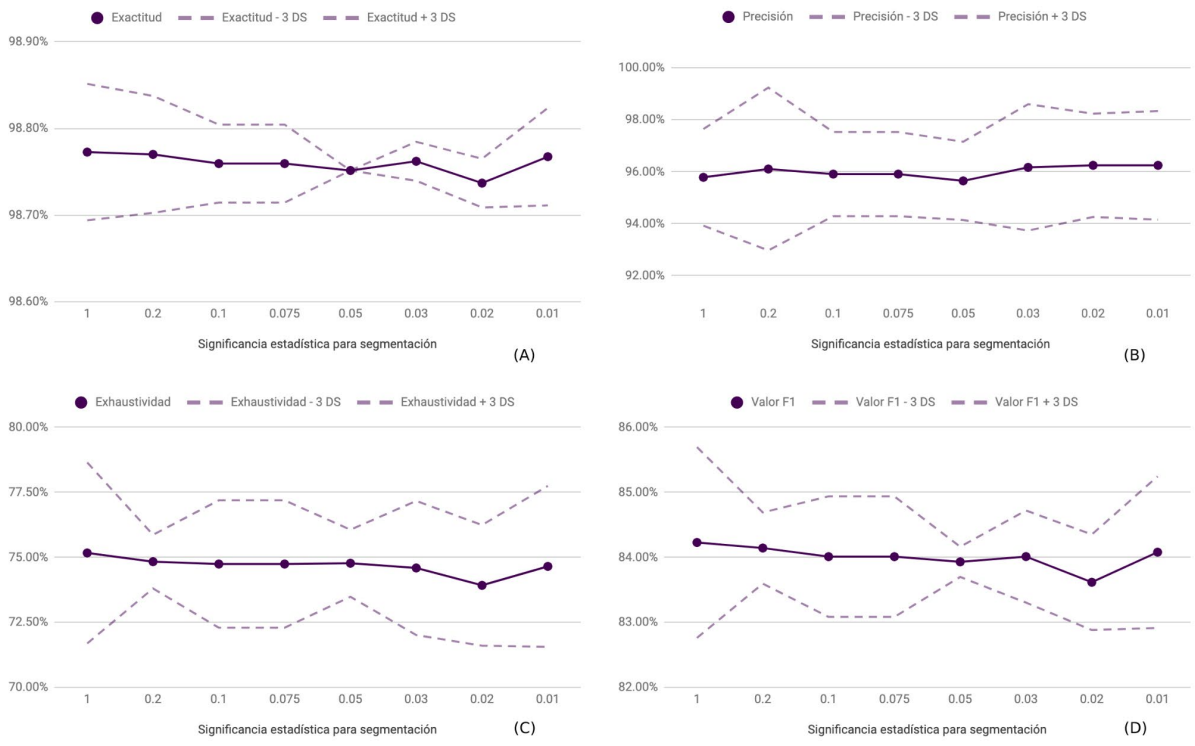



*Figura 23. Exactitud, Precisión, Exhaustividad y valor F1 en función del umbral α de segmentación para la etiqueta de tumor contra tejido normal, vasos sanguíneos y otros*

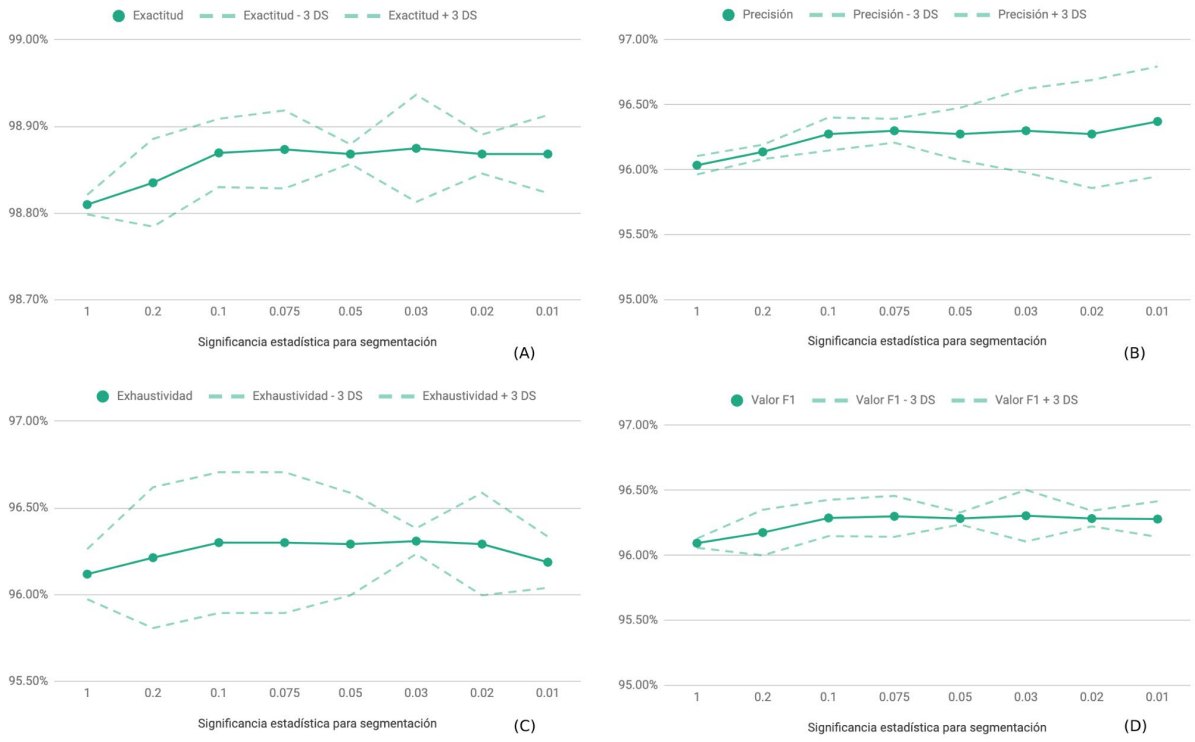



*Figura 24. Exactitud, Precisión, Exhaustividad y valor F1 en función del umbral α de segmentación para la etiqueta de vasos sanguíneos contra tejido normal, tumor y otros.*

La exactitud del modelo de clasificación por sí sola, no corresponde una referencia determinante para el desempeño del modelo. Por esto, junto a la exactitud, la precisión y exhaustividad, son métricas empleadas para el estudio del rendimiento de los modelos de entrenamiento supervisados. Evaluando en primera instancia la precisión y exhaustividad del modelo para detección de tumor (Figura 23 gráficas B y C), es posible observar que el valor medio obtenido a través del entrenamiento bajo la técnica de *K Folds* estratificados, se mantiene sin presentar variaciones y que los cambios del mismo se observan principalmente en la exactitud de la métrica, presentando para valores como $\alpha$=0.05 y $\alpha$=0.2, menor dispersión o entropía.

Al estudiar los resultados obtenidos para la exactitud y exhaustividad del modelo de detección de tejido sanguíneo (Figura 24, gráficas B y C), podemos observar un comportamiento contrario al presentado en modelo de detección de tumores, en el cual, el valor promedio de ambas métricas aumenta hasta en un 0.30%, aumentando a su vez la incertidumbre de las medidas.

Al comparar los resultados obtenidos para la exactitud, precisión y exhaustividad de ambos modelos, no es posible determinar que configuración posee mejor rendimiento para todas las métricas estudiadas. Para esto se decidió emplear el estudio del valor F1 como métrica principal de estudio, ya que esta permite determinar una media armónica entre la precisión y la exhaustividad. De esta forma, el valor F1 estableció el criterio de análisis y elección del parámetro de segmentación $\alpha$, con el mejor desempeño para el sistema. Al observar las gráficas asociadas al mismo en las figuras (Figura 23 (D) y Figura 24 (D)), es posible observar como la tendencia de ambas métricas (precisión y exhaustividad) se combinan, maximizando las variaciones entre los puntos de estudio de la variable independiente.

En primera instancia al estudiar el comportamiento del valor F1 asociado al modelo de detección de tumor (Figura 23 (D) ) es posible observar como el desempeño del modelo decrece en función del umbral de segmentación, presentando una variación de hasta el 0.2% a medida que el parámetro $\alpha$ decrece. Sin embargo, al estudiar la precisión de los resultados obtenidos (Figura 23 (B)) , es posible denotar como a medida que el valor de F1 decrece, existe mayor dispersión o incertidumbre en el resultado obtenido. Por lo cual, al combinar ambas métricas, es posible establecer que el umbral de segmentación de mayor desempeño es el asociado a $\alpha$=0.05, ya que el mismo es el que maximiza los resultados obtenidos para el valor F1 y la precisión, presentando la menor desviación estándar entre las posibles soluciones.

La gráfica de barras apiladas y escalonadas, presentadas en la Figura 25, representan los canales o longitudes de onda elegidos durante la etapa de optimización para caracterizar los tejidos de

tumor y vasos sanguíneos. Ambas señales pueden ser modeladas como filtros ópticos que extraen los canales de interés de la señal HS, recolectando solo los canales de información que permiten la caracterización de los tejidos. Ambas señales han sido establecidas con módulo 1 y aunque algunas referencias presentan valores de módulo 2, estas solo hacen referencia a la presencia de información asociada a ambos tejidos, lo cual denota que dicho canal de información aporta características al tejido de tumor y vasos sanguíneos de forma simultánea.

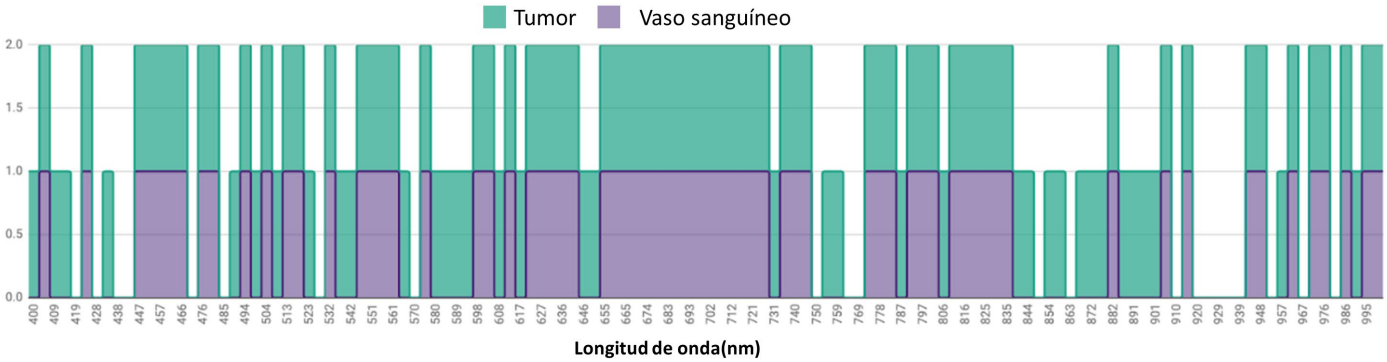



*Figura 25. Longitudes de onda asociadas a cada tejido.*

La Figura 26 presenta la distribución de los canales o longitudes de onda asociadas a las imágenes HS en función de los nuevos grupos de información obtenidos. De esta forma, se observa cómo el 42.2% de los canales aporta información a la detección de ambos tejidos y que un 17.2% de longitudes de onda no presentaron correlación alguna o una correlación débil generando ruido en la caracterización de estos. En concordancia a esto, un 40.6% de los canales HS están exclusivamente asociados a un tipo de tejido, entre los cuales el tumor demostró estar asociado a un mayor número de canales (28.9%).

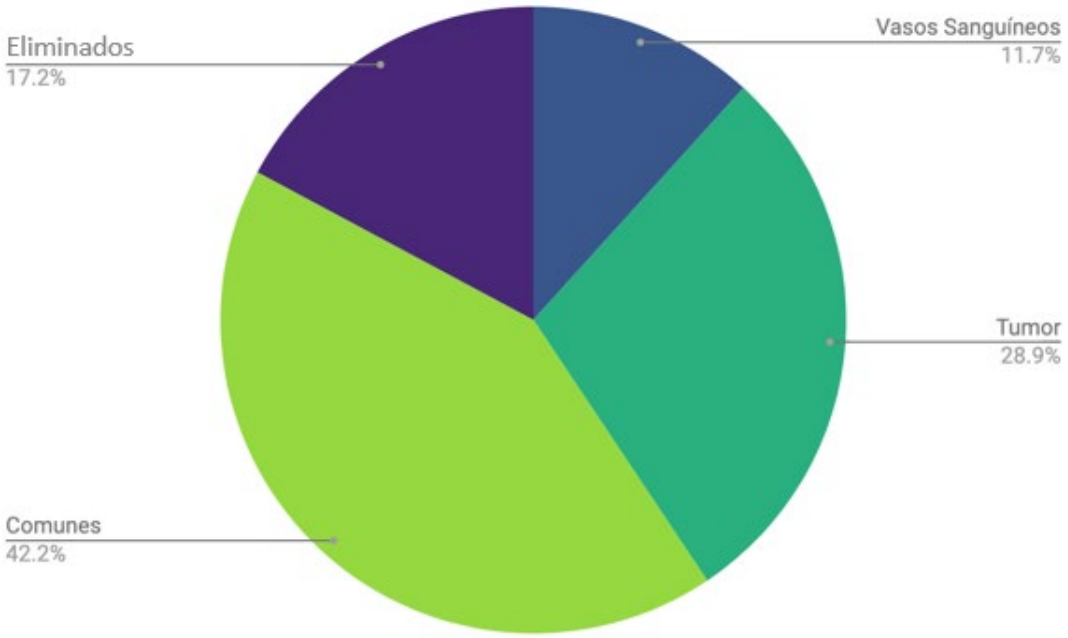



*Figura 26. Distribución de longitudes de onda de las imágenes HS en función de los tejidos.*

## 5.2. Resultados obtenidos en la discriminación entre arterias y venas.

### 5.2.1. Selección del algoritmo de entrenamiento

Para analizar el rendimiento de los algoritmos en estudio se estableció un barrido sobre la cantidad de agrupaciones empleadas con valores de 2, 4, 8, 10 y 20 grupos y se tomó como valor de referencia la media obtenida en cada proceso de agrupación. Los datos presentados a continuación han sido obtenidos exclusivamente con las muestras etiquetadas como vasos sanguíneos en las imágenes HS de nuestra base de datos del proyecto HELICoiD.

La Figura 27 presenta los resultados obtenidos al establecer el factor de *silhoutte* para cada algoritmo de entrenamiento no supervisado a medida que varía la cantidad de grupos de salida para el modelo. En la misma, es posible observar cómo el desempeño de cada algoritmo posee un comportamiento inversamente proporcional al número de clases que se establecen para el modelo de la salida. Dicho comportamiento denota la naturaleza de los datos y los algoritmos de clasificación no supervisados, en la cual, al aumentar el número de clases para modelar los

datos, existe menor separación entre las agrupaciones, lo que puede generar grupos menos diferenciados de los vecinos cercanos.

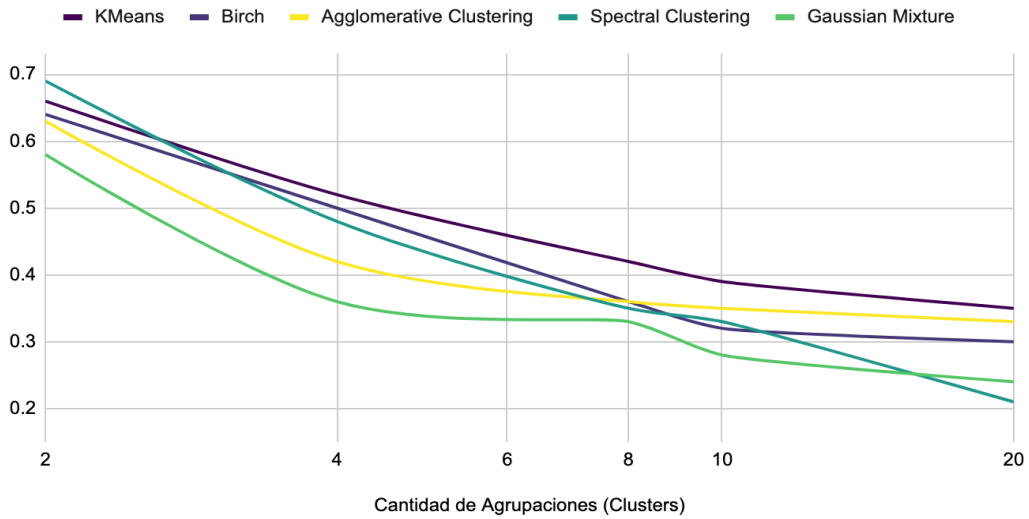



*Figura 27. Silueta en función de la cantidad de agrupaciones.*

La Figura 28 presenta los valores promedio y la tendencia del valor de silueta para cada modelo de entrenamiento no supervisado. Estableciendo como referencia el desempeño promedio de todos los modelos (0.416 Silueta), es posible evidenciar como *K-means* presento un mayor desempeño (12.5%) respecto a todos los modelos, seguido de *Birch* con un 1.92% más que la media. Al combinar esta información con la presentada en la Figura 27, se observa como *K-means* presentó una menor variación de desempeño durante todos los ciclos de ejecución de las pruebas, derivando en una menor desviación estándar y mayor exactitud de la medida del modelo.

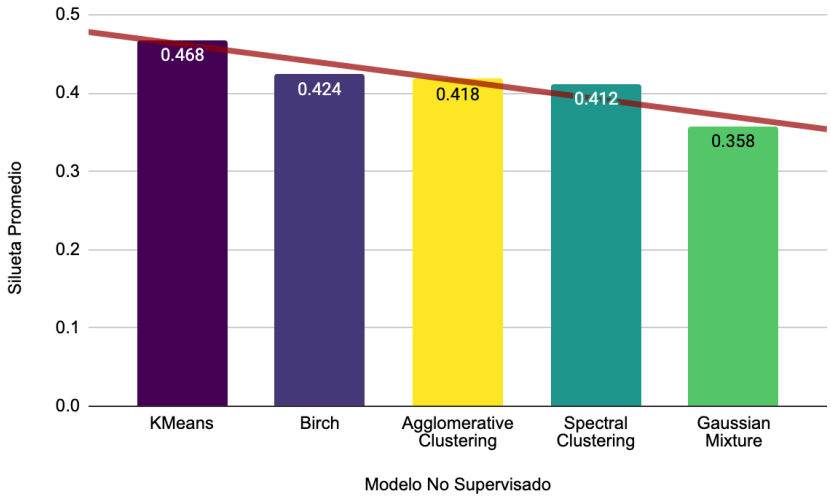



*Figura 28. Silueta promedio para cada modelo de entrenamiento.*

Debido a las características finitas de los sistemas de informática, el consumo de memoria es un aspecto de interés en el análisis del rendimiento de un algoritmo de clasificación, ya que la memoria disponible en el sistema es finita. La gráfica representada en Figura 29 presenta el consumo de memoria RAM en función de la cantidad de muestras utilizadas para el entrenamiento del modelo. En el cual, los algoritmos de *Spectral clustering, Birch, Gaussian mixture* y *Agglomerative clustering* presentan un comportamiento exponencial, requiriendo significativamente más memoria que algoritmos como *K-means* y *Gaussian mixture*.

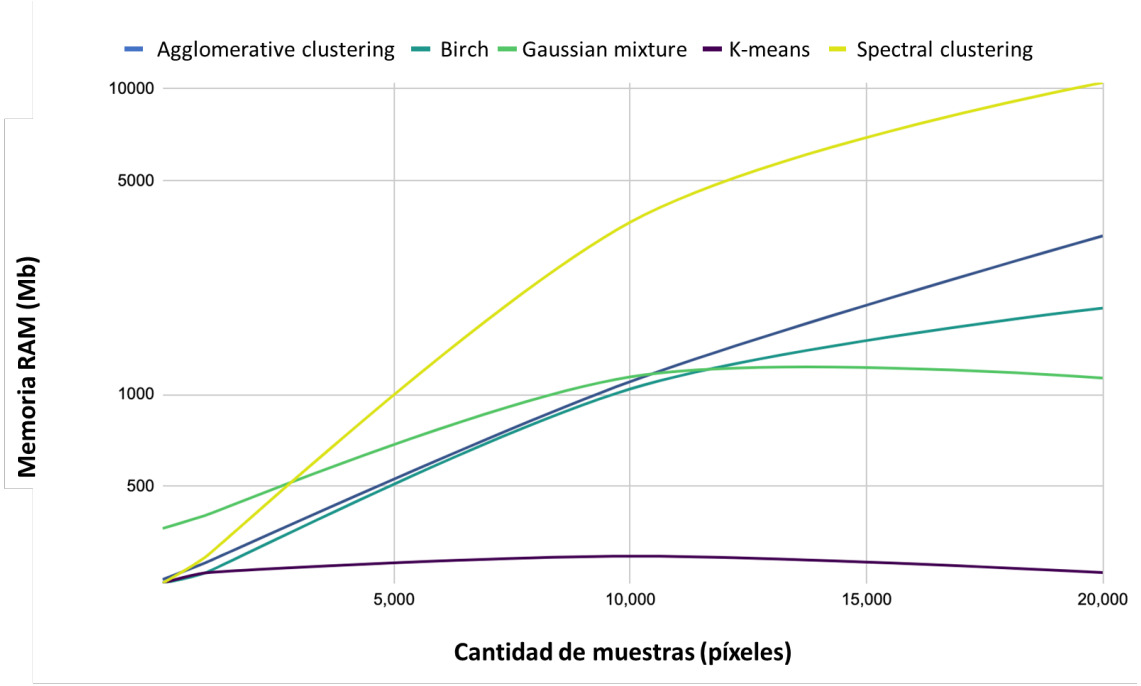



*Figura 29. Consumo de memoria RAM en megabytes en función de la cantidad de muestras seleccionadas para cada algoritmo en estudio*

Otro de los aspectos relevantes en el estudio de los algoritmos de agrupación está asociado al tiempo de ejecución, con la finalidad de identificar qué tan eficientes son los métodos en estudio para grandes volúmenes de datos. Es por esto, que la Figura 30 presenta los datos asociados al tiempo de entrenamiento en función de la cantidad de muestras en estudio. Es posible observar cómo los algoritmos de *Spectral clustering*, *Agglomerative clustering* y *Birch* poseen un comportamiento exponencial, mientras que algoritmos como *K-means* y *Gaussian mixture* presentan una relación lineal entre el tiempo de ejecución y el número de muestras.

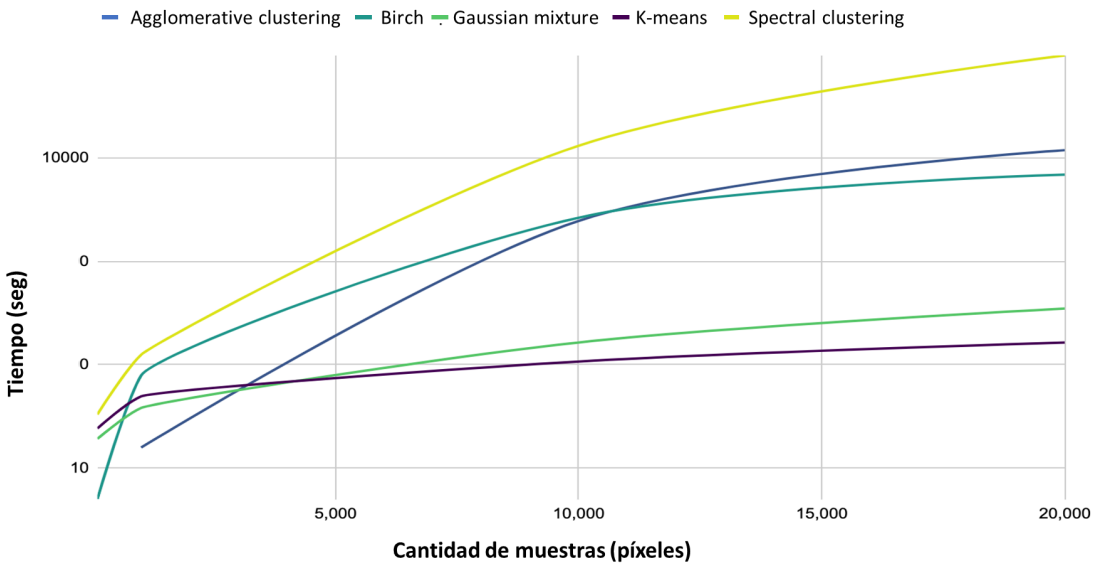



*Figura 30.Tiempo de ejecución en función de la cantidad de muestras seleccionadas para cada algoritmo en estudio.*

Inicialmente, entre los aspectos de estudio adicionales a la selección del algoritmo de agregación, también se estableció determinar la similaridad existente entre los resultados obtenidos bajo cada método de agrupación. La Figura 31 presenta la métrica *Measure Score* (media de puntuación en español) que se caracteriza por ser un valor ponderado entre qué tan similares y contenidos entre sí son los grupos obtenidos en cada método de clasificación, lo cual establece, que no existe similitud significativa entre los grupos obtenidos entre los métodos de agrupación. Entre los cuales *K-means* presenta el mayor puntaje promedio (0.74) y *Spectral clustering* (0.66), siendo respectivamente los algoritmos con mayor y menor similitud a los otros métodos estudiados.

Al estudiar los resultados obtenidos para cada modelo de agrupación, se consideró evaluar de forma complementaria si existía similitud alguna entre las agrupaciones generadas por los modelos estudiados, de esta forma, si varios modelos tendían a agrupar los datos de la misma

forma que el algoritmo seleccionado, podríamos afirmar que las conclusiones de esta etapa de la investigación serían válidas para dichos modelos con alta similitud. Para esto se estableció la utilización de la métrica *measure score* que permite conocer el parecido que existe entre los grupos generados por dos modelos de agrupación.


*Figura 31.Measure Score promedio para cada algoritmo de agrupación en estudio*

La Figura 31 presenta finalmente las puntuaciones medias obtenidas al comparar de forma cruzada cada uno de los algoritmos evaluados en esta etapa de investigación. Aunque algunos algoritmos como *K-means* y *Birch* o *K-means* y *Gaussian mixture,* presentaron similitudes mayores al 70%, no es posible afirmar que las agrupaciones generadas por varios modelos de entrenamiento no supervisado posean una misma tendencia en la agrupación de los datos.

Es por esto, que, al comparar los resultados obtenidos para los cinco algoritmos en estudio, es posible establecer que *K-means* como algoritmo de agrupación no solo posee el mejor desempeño en la segmentación de clases, sino que también es el algoritmo que mejor rendimiento presenta en el consumo de memoria RAM y tiempo de ejecución. Por lo cual, se establece como el algoritmo óptimo para ser utilizado en la caracterización de venas y arterias.

### 5.2.2. Resultados del modelo

Teniendo definido el modelo de entrenamiento no supervisado y los datos en estudios, es posible analizar los datos con la finalidad de generar un modelo capaz de diferenciar en grupos los vasos sanguíneos en función de la información sobre el tejido descritas en las imágenes HS. Se propone una hipótesis inicial, de evaluar el desempeño del modelo distribuyendo los datos en dos grupos de tejidos con la finalidad de determinar si estos grupos presentaban o no correlación con los identificadores de tejido de venas y arterias.

Establecido el algoritmo de clasificación, los datos de entrenamiento del modelo para agrupar los datos asociados de vasos sanguíneos, se procedió a aplicar el procedimiento presentado en la Figura 21, generando finalmente un modelo predictivo de dos clases bajo el algoritmo de *K-means*. Posterior a esto, se procedió a realizar una interpretación visual de los resultados, basado en las imágenes de referencia presentadas en la base de datos. Al procesar las imágenes, para representar en los resultados se han elegido las imágenes obtenidas de la base de datos, OP5C1, OP22C1 y OP12C2. En la imagen (Figura 32) se presentan tres columnas, la primera columna son las imágenes sacadas literalmente de la base de datos, la segunda son las clases generadas cuando tenemos $K$=2, siendo la clase 0 el color verde y la clase 1 el color amarillo. Para el fondo se ha usado el color -1, el cual se usa solo en la tercera columna donde solo se presentan el procesado de las partes de la imagen que han sido previamente clasificadas como vaso sanguíneo por la base de datos.

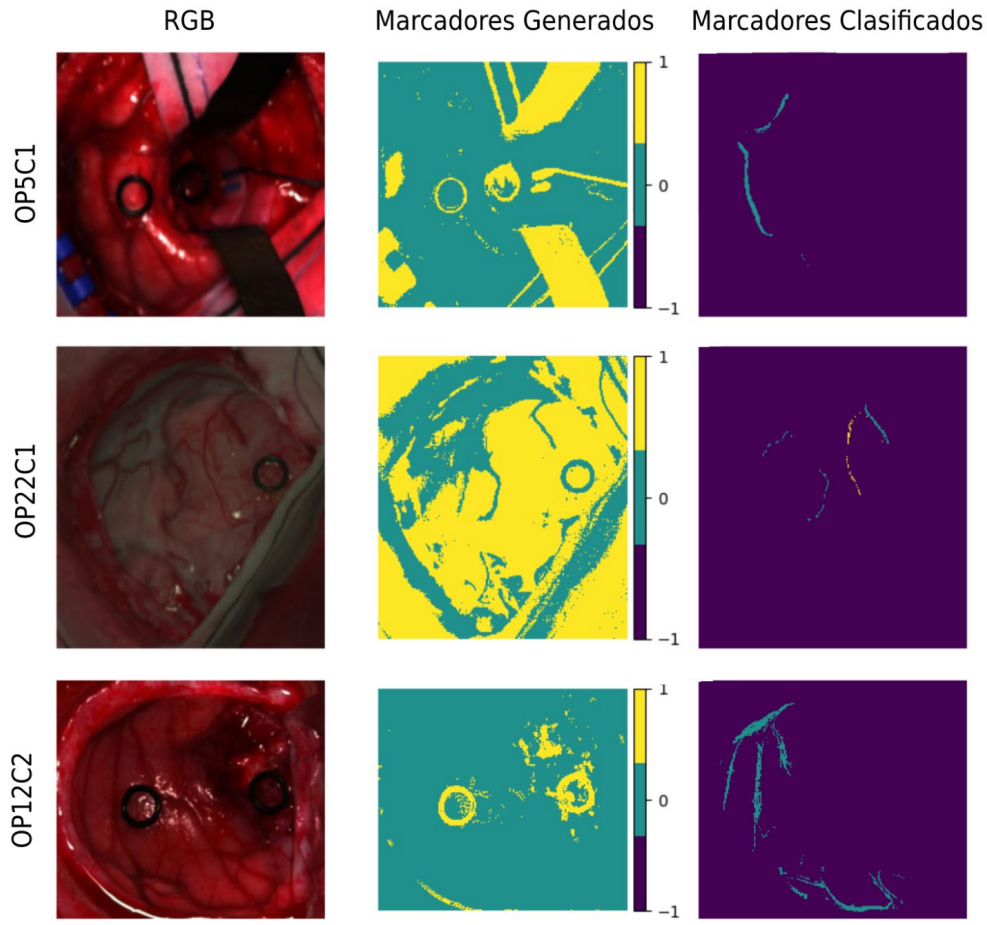



*Figura 32.Procesado de Imágenes HS de la base de datos HELICoiD bajo modelo K-means utilizando K=2.*

La Figura 32 ejemplifica los resultados obtenidos para 3 casos de referencia en pacientes, no obstante, las consideraciones presentadas en esta sección harán referencia a lo observado en las 37 imágenes que componen el estudio y de las cuales se puede hacer uso en el *Anexo* de este mismo documento. Las imágenes seleccionadas como referencia han sido escogidas de la base de datos debido a que ejemplifican distintas condiciones de iluminación, tejidos y elementos quirúrgicos, permitiendo comparar los resultados obtenidos en distintos escenarios. Cada trio de imágenes presentada en la Figura 32 contiene la representación RGB (*Red, Green, Blue*), el mapa generado a partir de la clasificación en grupos realizada por *K-means* de los píxeles de cada imagen HS, junto al enmascaramiento de este mapa generado a partir de los descriptores etiquetados como vasos sanguíneos en la base de datos. Estas 3 representaciones permiten en primera instancia observar qué objetos y tejidos constituyen cada imagen y como el modelo generado los interpreta.

Al considerar los marcadores generados asociados a los casos clínicos OP5C1 y OP12C2, se observa predominancia del identificador 0 (clase 0) sobre la imagen, que al contrastarlo con la representación RGB, presenta una correlación del 84% con los pixeles clasificados como tejido orgánico, mientras el identificador 1 se correlaciona 72% con los marcadores asociados a instrumentación quirúrgica, no obstante, al observar la imagen OP22C1, se evidencia el caso contrario, en el cual, ambos identificadores están altamente correlacionados con el tejido orgánico, denotando que posiblemente no exista correlación entre las clases generadas y los tejidos en estudio.

Al estudiar los marcadores clasificados, presentados en dichas imágenes HS es posible evidenciar como casi la totalidad de los tejidos son asociados a un solo identificador de clase (clase 0), lo cual sugiere que el modelo entrenando no está aportando información de los tejidos predominante, sino que está detectando anomalías o muestras con características divergentes a la de la clase predominante.

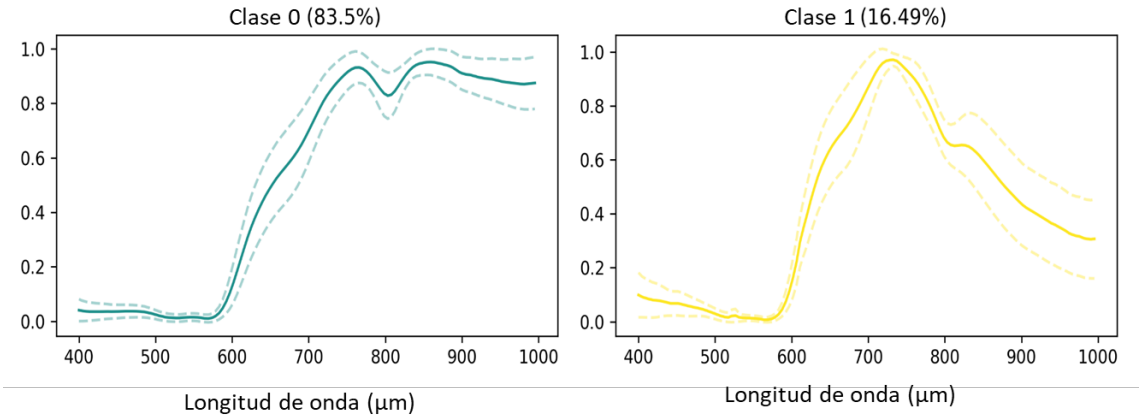



*Figura 33.Reflectancia Normalizada en función de la longitud de onda para cada grupo de muestras.*

La Figura 33 presenta la reflectancia normalizada en función de la longitud de onda, caracterizando la respuesta espectral de los tejidos presentados en los datos de entrenamiento. Al estudiar la naturaleza de la curva de reflectancia modulada para la Clase 0, podemos evidenciar un comportamiento que se asocia directamente a la curva de reflectancia asociada a tejidos orgánicos que pueden ser constituidos por tejido sano, tumor y vasos sanguíneos. En contraparte, la naturaleza de la curva de reflectancia presentada en la Clase 1, se asocia a la curva que estaría presente al caracterizar instrumentación quirúrgica o elementos inorgánicos.

Considerando estos aspectos junto a las proporciones que constituyen cada clase ( 83% y 17%), respectivamente, es posible aseverar que los resultados obtenidos durante este ensayo generaron un modelo para diferenciar entre tejido orgánico e inorgánico.

Este resultado obtenido puede ser atribuido a diferentes aspectos técnicos y procedimentales que pudieran afectar la calidad de los datos y el procedimiento empleado para la construcción del modelo predictivo. Entre las fuentes principales de errores puede identificarse la posibilidad de que los datos puedan contener píxeles etiquetados que no correspondan a la clase en estudio junto a los técnicos mediante el cual opera el algoritmo de agrupación seleccionado.

*K-means* como modelo de entrenamiento no supervisado ha demostrado ser un algoritmo de clasificación eficiente en el consumo de recursos. Sin embargo, al estudiar el funcionamiento del modelo, se observa que el mismo está diseñado para agrupar las muestras utilizadas en el entrenamiento de forma tal, que las clases obtenidas tengan la mayor separación posible entre ellas. Al combinar dicha información junto a la hipótesis de que los tejidos venas y arterias pueden tener una firma espectral similar, es posible establecer que la utilización de un algoritmo de agrupamiento distribuyendo los datos en solo 2 agrupaciones, no garantiza la distensión de venas y arterias.

Con la finalidad de mitigar el efecto que puede introducir muestras clasificadas incorrectamente, se decidió analizar el procedimiento presentado, considerando incluir todos píxeles contenidos en cada imagen HS. Esta nueva metodología requiere a su vez, aumentar el número de clases o grupos con los cuales serán modelados datos, ya que los datos de entrenamiento incluyen información de elementos que no corresponde a los tejidos en estudio.

Las siguientes figuras (Figura 34,Figura 35 y Figura 36) presentan los resultados obtenidos para 3 imágenes que ejemplifican cómo el modelo entrenado clasifica la totalidad de tejidos existentes en las imágenes HS, en agrupaciones que varían de 2 a 20 grupos por modelo

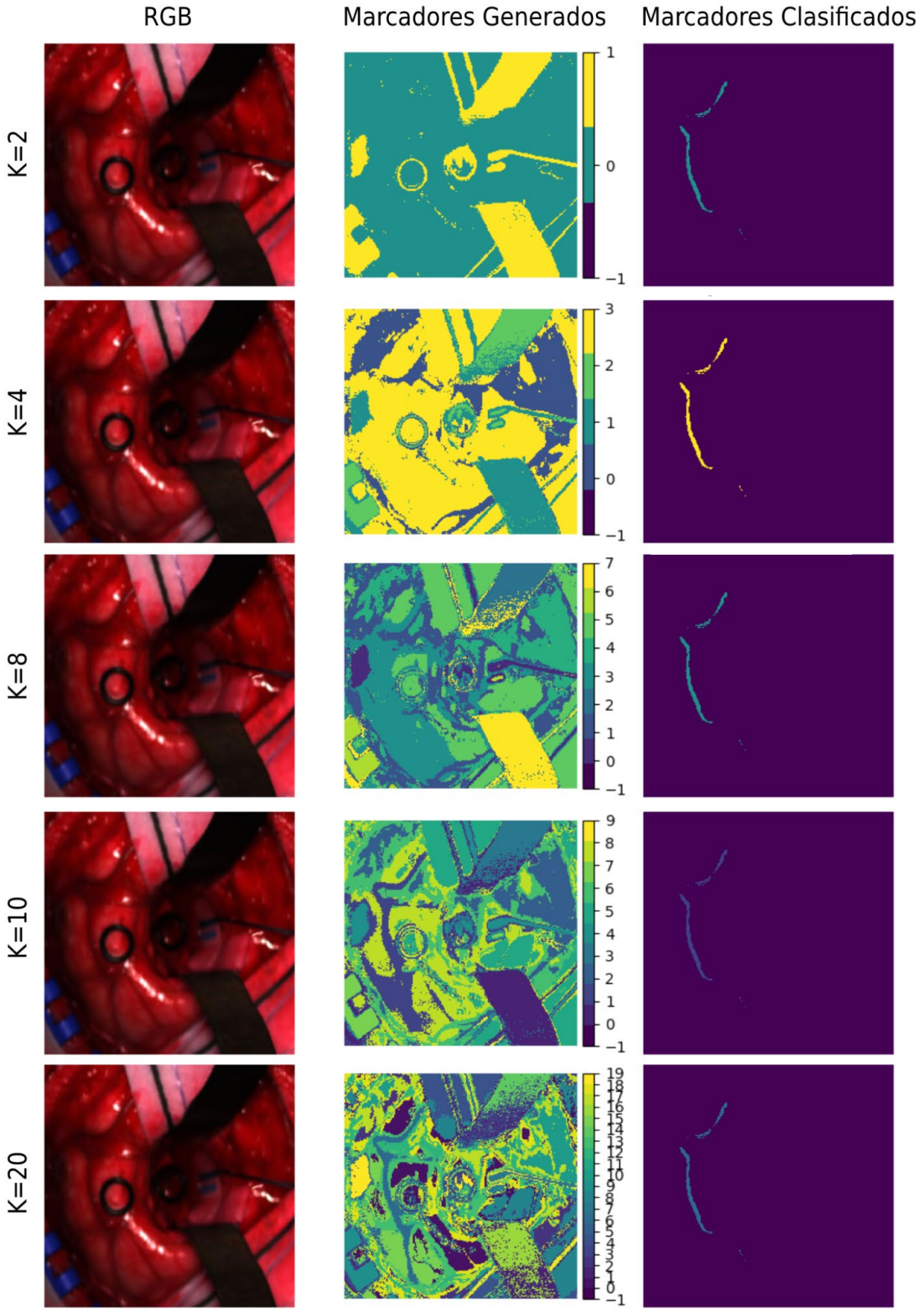



*Figura 34.Comparativa de resultados obtenidos al procesar el registro OP5C1 mediante modelos no supervisados basado en K-means con K grupos de clasificación*

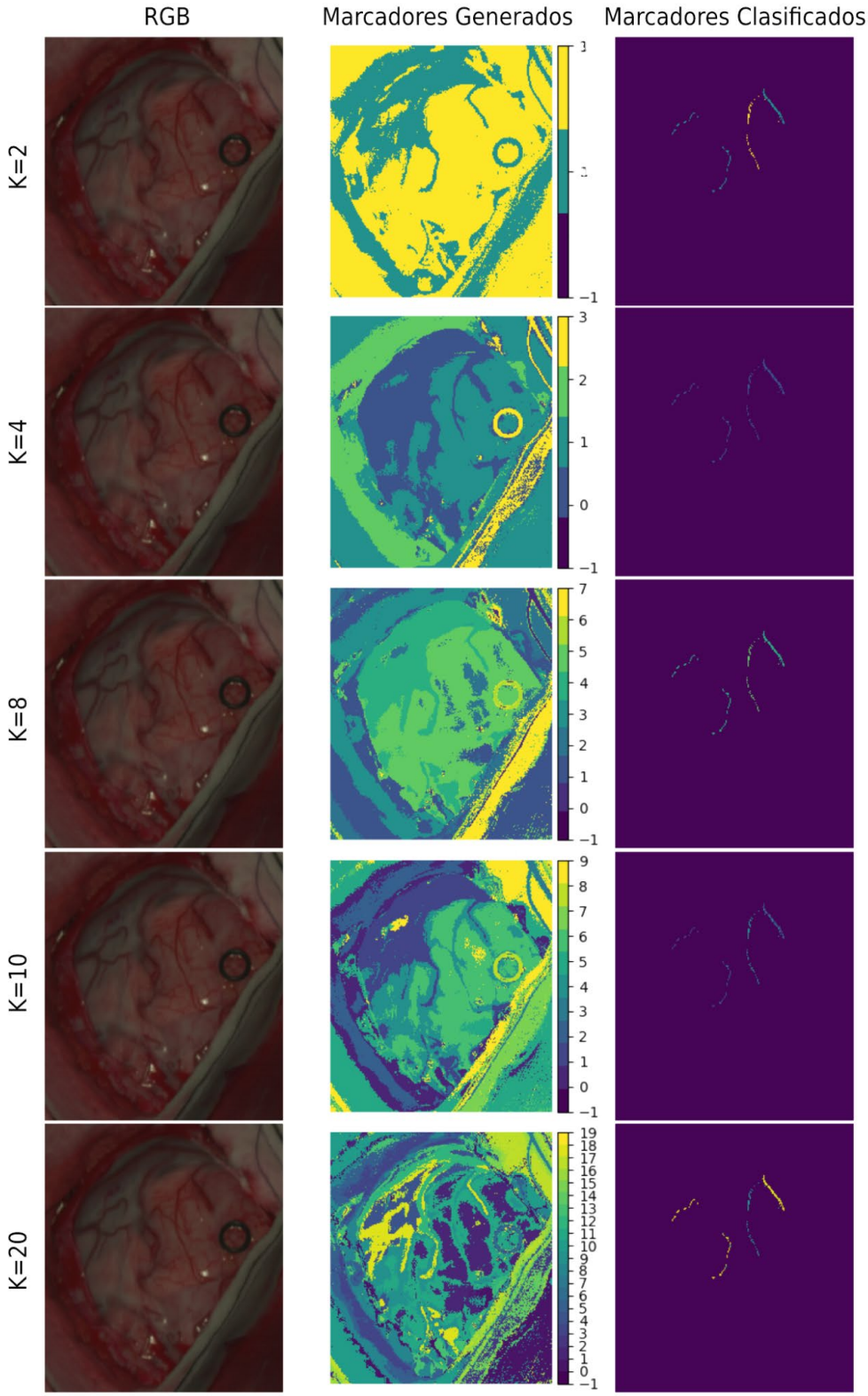



*Figura 35.Comparativa de resultados obtenidos al procesar el registro OP22C1 mediante modelos no supervisados basado en K-means con K grupos de clasificación.*

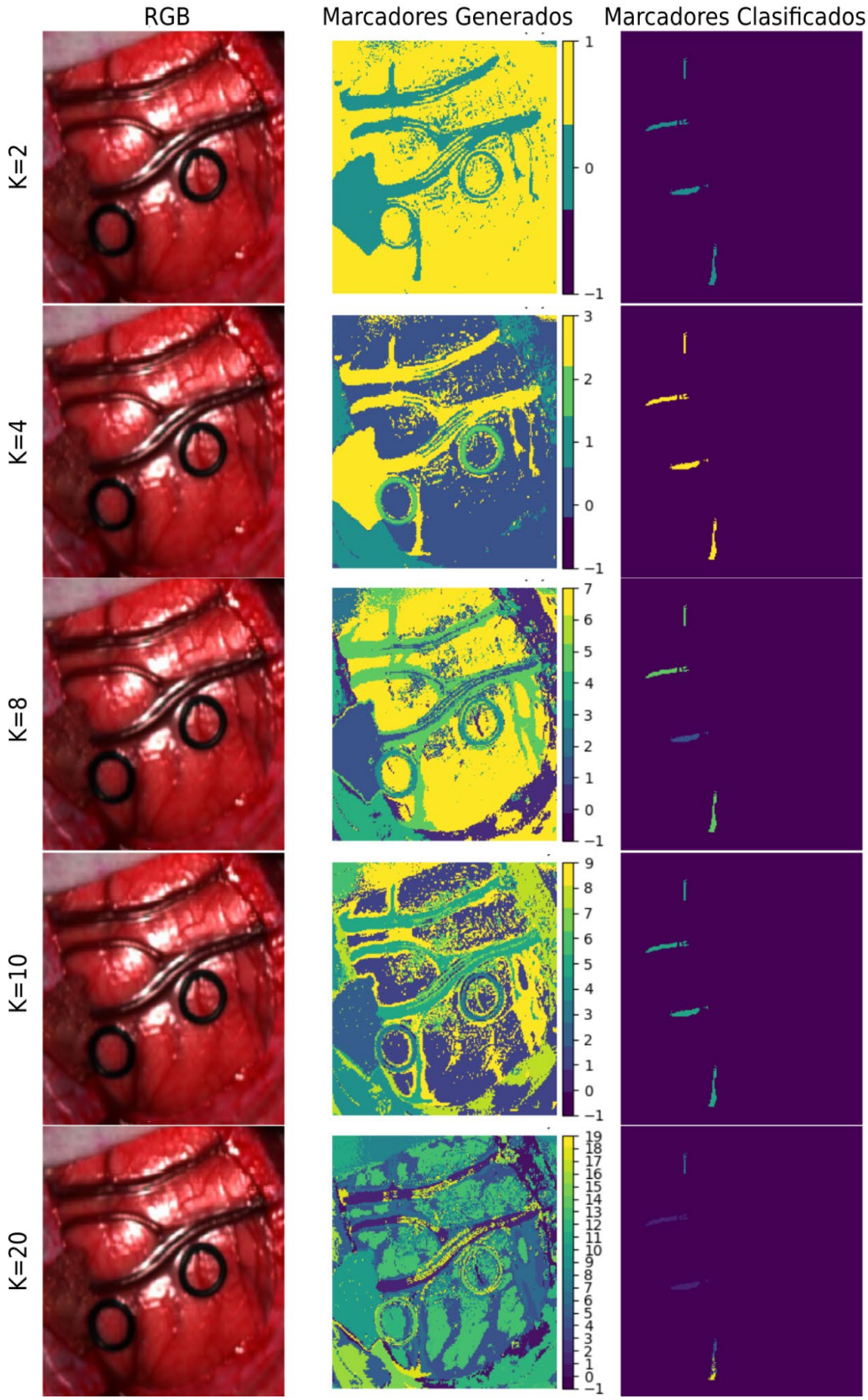



*Figura 36.Comparativa de resultados obtenidos al procesar el registro OP4C2 mediante modelos no supervisados basado en K-means con K grupos de clasificación.*

Al comparar los datos obtenidos en las figuras anteriores (Figura 32, Figura 34, Figura 35 y Figura 36), es posible observar cómo los modelos de agrupación de 2 y 4 clases (K=2 y K=4), no proporcionan información suficiente para diferenciar los tejidos orgánicos entre sí, diferenciando principalmente las regiones que contienen tejido orgánico de elementos como herramientas quirúrgicas o partes del cráneo. En contraposición, a medida que se aumenta la cantidad de clases disponibles para ordenar los datos contenidos en las imágenes HS, es posible destacar como aparecen contornos más definidos para tejidos como venas, arterias, tejido normal, cráneo y elementos inorgánicos utilizados en la intervención quirúrgica.

No obstante, el aumento de clases para el entrenamiento del modelo de clasificación aumenta la complejidad de los grupos generados, disminuyendo a su vez el distanciamiento entre las clases, lo cual se traduce visualmente en la introducción de ruido en la representación de las imágenes. De esta forma, se observa como en la Figura 36 al incrementar el número de clases a 20 agrupaciones, los tejidos que habían sido diferenciados con mayor detalle en las demás clases, pierden especificidad, derivando en imágenes con segmentaciones menos precisas.

La etapa de validación de datos ha sido establecida en dos etapas con la finalidad de determinar si existen indicadores de tejido que puedan estar asociados a las venas y arterias, posteriormente determinando con ayuda del equipo médico si dichos marcadores corresponden a los tejidos en estudio.

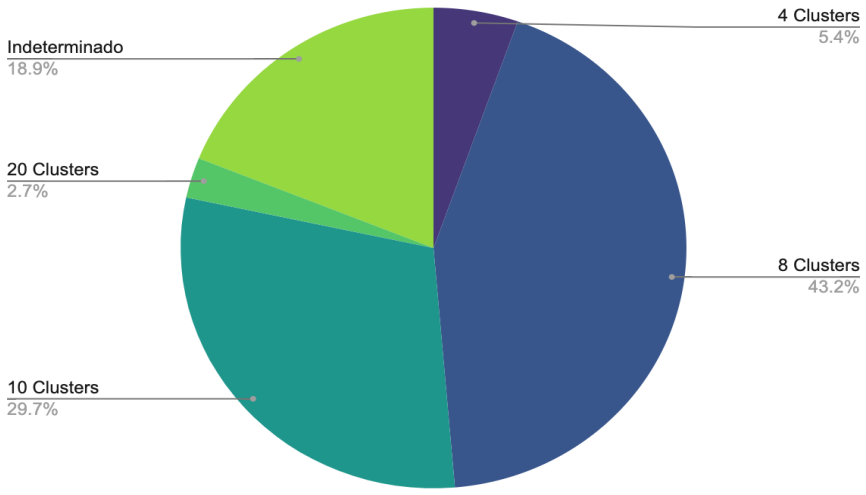



*Figura 37. Distribución de imágenes en función de los identificadores de tejido que presentan información de vasos sanguíneos.*

La Figura 37 presenta los resultados obtenidos tras analizar cada imagen contenida en la base de datos HELICoiD e identificar qué modelo presenta mejor segmentación para los tejidos

asociados a vasos sanguíneos. De la totalidad de 37 imágenes evaluadas en esta investigación, solo el 18.9% de estas no mostró un patrón interpretable en el cual pudiera distinguirse el tejido asociado a vasos sanguíneos. Presentando a su vez mayor correlación con los tejidos sanguíneos para distribuciones de 8 y 10 clases. Combinando dichos resultados, se observa como los resultados obtenidos para modelos de 8 y 10 agrupaciones presentan líneas de contorno para tejidos que pueden ser asociados a vasos sanguíneos y a su vez, caracterizan regiones en los es posible observar acumulación de sangre en el tejido. La presente Tabla 4 contiene las etiquetas asociadas a cada clase generada por el modelo, las cuales han sido determinadas analizando los marcadores generados para las 37 imágenes en estudio, asociando a cada clase la etiqueta con mayor predominancia entre todas las imágenes. Es decir, para las 8 clases que se han generado para el modelo *K-means* con *K=8,* se muestra las etiquetas que corresponderían a cada clase. Estas referencias pueden ser tomadas como base para validar con el equipo clínico, la correlación existente entre cada uno de los marcadores de tejido y las clases de venas y arterias, pudiendo proporcionar información acerca de regiones que contengan mayor o menor irrigación sanguínea y en consecuencia identificar regiones poco oxigenadas.

*Tabla 4. Asociación de tejidos y etiquetas de agrupación establecidos a partir de la etapa de validación manual de imágenes con 8 grupos.*

| Identificador | Tipo de tejido |
|---|---|
| Clase 0 | Normal |
| Clase 1 | Otros |
| Clase 2 | Normal |
| Clase 3 | Vaso Sanguíneo |
| Clase 4 | Vaso Sanguíneo |
| Clase 5 | Vaso Sanguíneo |
| Clase 6 | Otros |
| Clase 7 | Normal |

Las curvas de reflectancia presentadas en la Figura 38, denotan la firma espectral de los tejidos que constituyen cada grupo. Dicha información puede ser empleada para identificar y validar que grupos se correlacionan con los tejidos en estudio (tejido orgánico, inorgánico o vasos sanguíneos). Al combinar dichas gráficas, con los identificadores presentados en la Tabla 4, observamos como las clases 0 y 2, presentan una firma espectral idéntica a la esperada para el tejido sano o normal [74]. Caso contrario el presentado en las clases 1 y 6, las cuales, presentan una reflectancia asociada a goma o elementos quirúrgicos [74].

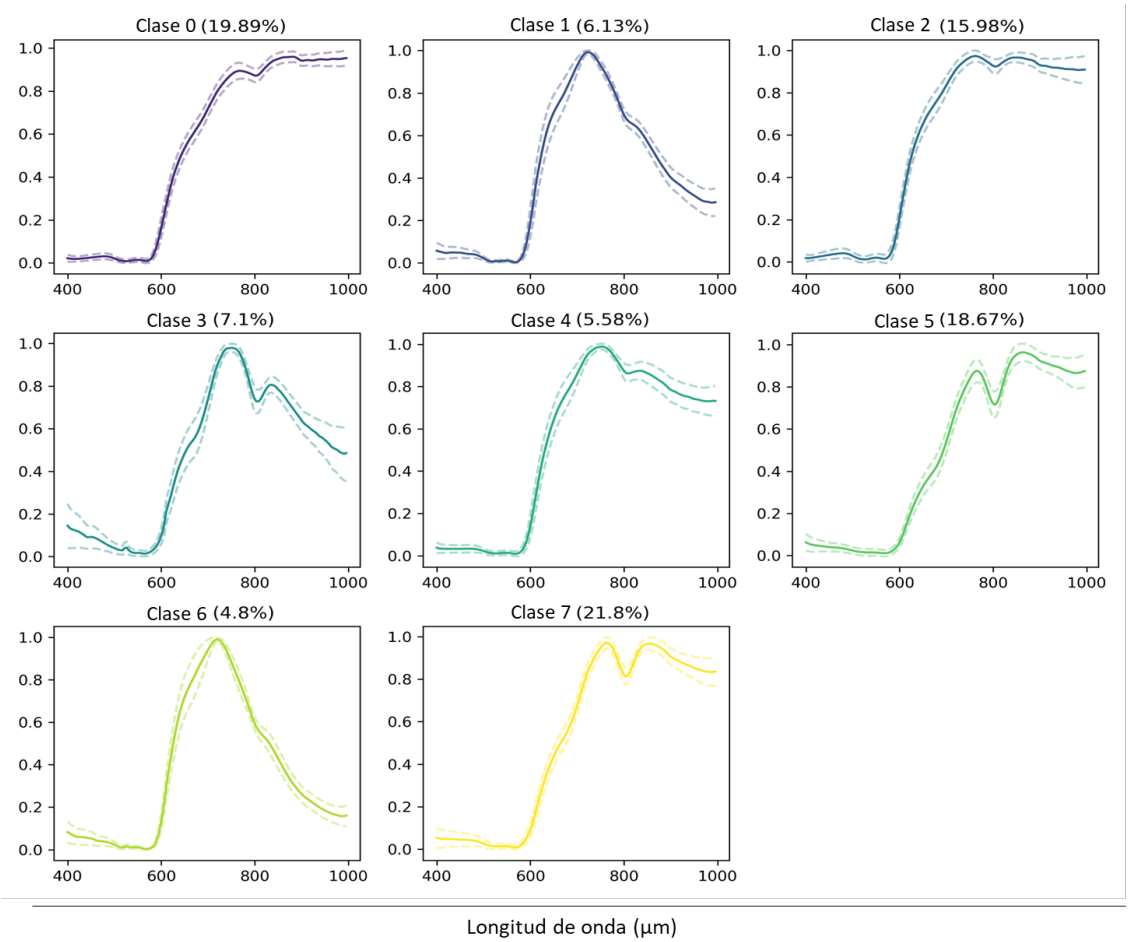



*Figura 38.Reflectancia Normalizada en función de la longitud de onda para cada grupo de muestras.*

Estos planteamientos, proponen que existen 4 sub clases generadas por el modelo de clasificación no supervisado, en cual, la respuesta impulsiva de las muestras no se asocia a tejido normal o quirúrgico, presentando una curvatura similar al que reflejan los tejidos asociados a vasos sanguíneos, con la principal característica de que estos grupos presentan una mayor o menor depresión en la banda de 800 µm [75], aspecto que otras investigaciones ha sido asociado a la mayor o menor presencia de oxígeno en sangre . Para que este modelo sea eficiente y pueda aportar valor a la comunidad científica es necesario validar los resultados con un equipo médico o con imágenes establecidas con etiquetas que reflejen que tejido dentro del vaso sanguíneo es arteria o cual vena. En este estudio, lamentablemente, no se ha podido contar con la validación médica necesaria puesto que el personal sanitario no se encuentra ahora mismo con posibilidad de analizar detenidamente cada resultado.

# Capítulo 6: Conclusiones y líneas futuras

Tras finalizar el presente proyecto, se exponen las conclusiones que se han sacado del desarrollo de los métodos de clasificación para la distinción de tejidos. En este capítulo también se proponen mejoras a los métodos desarrollados que han surgido y que no han sido desarrolladas porque escapan del alcance de este TFG.

En el desarrollo de la primera parte del proyecto destinada a la discriminación de tejido asociado a tumor y tejido asociado a vaso sanguíneo mediante el algoritmo SVM, se concluye que el 42,2% de los canales aporta información a la detección de ambos tejidos y que un 17,2% de longitudes de onda no presentaron correlación alguna o una correlación débil, generando ruido en la caracterización de estos. En concordancia a esto, un 40,6% de los canales HS están exclusivamente asociados a un tipo de tejido, entre los cuales el tumor demostró estar asociado a un mayor número de canales (28,9%). Esta información tiene sentido puesto que la lesión principal que observamos en las imágenes es la del tumor cerebral. Se determina también en esta primera parte que la reducción de la imagen en el proceso aporta valor al sistema de clasificación puesto que los resultados obtenidos definen las longitudes de onda determinantes sin necesidad de recurrir a grandes cantidades de computación en el procesado de imágenes HS.

En la segunda parte, basada en la discriminación de arterias y venas en el tejido asociado a vasos sanguíneos, se extrae que, al estudiar los modelos de entrenamiento no supervisados, no es posible evaluar su desempeño de la misma forma que se estudian los modelos supervisados, ya que estos, carecen de una fuente etiquetada o valores de referencia con los cuales contrastar el desempeño del modelo. Además, utilizar métodos que permitan definir el número de clases de salida, ha permitido medir de forma más eficiente el rendimiento y eficiencia de los métodos estudiados.

En el análisis del modelo no supervisado, se concluye que los algoritmos de agrupación son poco eficientes cuando la dimensionalidad es mayor a 10 componentes y que el algoritmo *K-means* como modelo de clasificación no supervisado ha demostrado ser un algoritmo eficiente en el consumo de recursos. Al combinar dicha información es posible establecer que la utilización de un algoritmo de agrupamiento distribuyendo los datos en solo dos agrupaciones, no garantiza la discriminación entre venas y arterias.

Con la finalidad de mitigar el efecto que pueden introducir las muestras clasificadas incorrectamente, se decidió analizar el procedimiento presentado, considerando incluir todos píxeles contenidos en cada imagen HS. Esta metodología requiere a su vez aumentar el número

de clases o grupos con los cuales serán modelados los datos, ya que los datos de entrenamiento incluyen información de elementos que no corresponde a los tejidos en estudio. De la totalidad de 37 imágenes evaluadas en este trabajo, solo el 18,9% de estas no mostró un patrón interpretable en el cual pudiera distinguirse el tejido asociado a vasos sanguíneos, presentando a su vez mayor correlación con los tejidos sanguíneos para distribuciones de 8 y 10 clases. Combinando dichos resultados se observa como los resultados obtenidos para modelos de 8 y 10 agrupaciones presentan líneas de contorno para tejidos que pueden ser asociados a vasos sanguíneos y a su vez, caracterizan regiones en los que es posible observar acumulación de sangre en el tejido. Aunque no ha sido posible validar los resultados con un equipo médico o con resultados de oxigenación similares para estas imágenes, se podría concluir con que la línea de clasificación aporta información relevante en la línea de desarrollo y mejora para estudios de investigadores de la rama biomédica en el diagnóstico por imagen.

Finalmente, a través del desarrollo del trabajo y de las incógnitas que han podido surgir en la realización del presente TFG se presentan como iniciativa otras líneas de investigación que pueden ser abordadas en el futuro por la comunidad científica:

- Comprobar que otros algoritmos supervisados pueden obtener iguales o mejores resultados en la clasificación de este tipo de imágenes HS.
- Definir otros modelos de clasificación que permitan corroborar los datos obtenidos en ésta y otras agrupaciones realizadas que hayan partido de la misma base de datos.
- Modelos de validación de clasificación que permitan determinar que la validación ha sido correcta.
- Trabajo mano a mano con un equipo médico en la validación de tejido no clasificado todavía como venas y arterias que aportan información en la peligrosidad de tumores.
- Aplicar dichos modelos, supervisado y no supervisado, a otras imágenes con lesión tumoral y ampliar el estudio y las conclusiones extraídas.

# Capítulo 7: Bibliografía

[1] "Definition of tumor - NCI Dictionary of Cancer Terms - NCI." https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor (accessed Aug. 17, 2022).

[2] "¿Qué es Tumor? » Su Definición y Significado [2021]." https://conceptodefinicion.de/tumor/ (accessed May 11, 2021).

[3] E. Los, T. Astrocitarios, and T. Oligodendrociticos, "Tumores cerebrales Introducción," pp. 1–5, 2012, Accessed: Apr. 02, 2018. [Online]. Available: https://www.seom.org/es/info-sobre-el-cancer/tumores-cerebrales?showall=1

[4] "'Estudio de la actividad antitumoral en modelo de tumor sólido de cáncer de mama murino del pro-fármaco selectivo en hipoxia'".

[5] "Evaluation and Diagnosis of Brain Tumors," *Department of Neurosurgery, Unicersity of Colorado School of Medicine*, 2016. http://www.ucdenver.edu/academics/colleges/medicalschool/departments/Neurosurgery/patientcare/multi-disciplinaryprograms/AdultBrainTumorProgram/Pages/EvaluationandDiagnosisofBrainTumors.aspx (accessed Apr. 02, 2018).

[6] "Tumor cerebral." http://www.cirugia-neurologica.org/tumor-cerebral.ws (accessed Apr. 02, 2018).

[7] B. T. Classification, "Clasificación de los tumores cerebrales," pp. 339–342, 2017.

[8] "Tumores cerebrales - diagnóstico, evaluación y tratamiento." https://www.radiologyinfo.org/sp/info.cfm?pg=braintumor (accessed Apr. 02, 2018).

[9] "INEbase / Sociedad /Salud /Estadística de defunciones según la causa de muerte / Últimos datos." https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176780&menu=ultiDatos&idp=1254735573175 (accessed Aug. 17, 2022).

[10] "• Tipos de cáncer más mortales en el mundo en 2020 | Statista." https://es.statista.com/estadisticas/636256/mortalidad-por-cancer-muertes-a-nivel-mundial-por-tipo/ (accessed Jun. 22, 2021).

[11] D. N. Louis *et al.*, "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary," *Acta Neuropathologica*, vol. 131, no. 6, pp. 803–820, 2016, doi: 10.1007/s00401-016-1545-1.

[12] M. Sinning, "CLASIFICACIÓN DE LOS TUMORES CEREBRALES," *Revista Médica Clínica Las Condes*, vol. 28, no. 3, pp. 339–342, May 2017, doi: 10.1016/J.RMCLC.2017.05.002.

[13] R. G. Figueiras, A. R. Padhani, J. C. Vilanova, V. Goh, and C. Villalba Martín, "Imagen funcional tumoral. Parte 1 ARTICLE IN PRESS," vol. 52, no. 2, pp. 115–125, 2010, doi: 10.1016/j.rx.2009.12.008.

[14] L. E. Contreras, "EPIDEMIOLOGÍA DE TUMORES CEREBRALES," *Revista Médica Clínica Las Condes*, vol. 28, no. 3, pp. 332–338, May 2017, doi: 10.1016/J.RMCLC.2017.05.001.

[15] "Meningioma - Diagnóstico y tratamiento - Mayo Clinic." https://www.mayoclinic.org/es-es/diseases-conditions/meningioma/diagnosis-treatment/drc-20355648 (accessed Oct. 27, 2021).

[16] "Astrocitoma de bajo grado – Cirugía Neurológica." http://www.cirugia-neurologica.org/blog/astrocitoma-de-bajo-grado/ (accessed Oct. 27, 2021).

[17] "Tratamiento de los astrocitomas infantiles (PDQ®)–Versión para profesionales de salud - Instituto Nacional del Cáncer." https://www.cancer.gov/espanol/tipos/cerebro/pro/tratamiento-astrocitomas-infantiles-pdq (accessed Oct. 27, 2021).

[18] "Oligodendroglioma - Instituto Nacional del Cáncer." https://www.cancer.gov/rare-brain-spine-tumor/espanol/tumores/oligodendroglioma (accessed Oct. 27, 2021).

[19] "oligoastrocitoma: ¿qué es? - tumores - 2021." https://energymedresearch.com/14746-oligoastrocytoma-what-is-it (accessed Oct. 27, 2021).

[20] "Tipos de carcinoma." https://www.sanitas.es/sanitas/seguros/es/particulares/biblioteca-de-salud/cancer/carcinoma.html (accessed Oct. 27, 2021).

[21] "Hipoxia en la malignidad del cáncer: Revisión." http://ve.scielo.org/scielo.php?pid=S0535-51332009000400012&script=sci_arttext&tlng=pt (accessed May 03, 2021).

[22] U. de Jaén, A. : Almudena, and P. Camacho, “Facultad de Ciencias Experimentales Estudio de las bases moleculares de la respuesta a la hipoxia,” 2014.

[23] S. M. Catalán, J. Mbmcg, F. de Biología, and S. M. Catalán Jiménez, “Efecto de la hipoxia sobre la quimiosensibilidad al 5-fluorouracilo en células Hela”.

[24] C. Boticario Boticario and M. Cascales Angosto, “Hipoxia y cáncer CORE View metadata, citation and similar papers at core.ac.uk provided by Real Academia Nacional de Farmacia: Portal Publicaciones,” *An. R. Acad. Nac. Farm*, vol. 76, no. 3, pp. 379–408, 2010.

[25] “Inhibidores de la angiogénesis - Instituto Nacional del Cáncer.” https://www.cancer.gov/espanol/cancer/tratamiento/tipos/inmunoterapia/hoja-informativa-inhibidores-angiogenesis (accessed May 17, 2021).

[26] R. Letelier, “Angiogénesis y cáncer,” *Medwave*, vol. 7, no. 3, Apr. 2007, doi: 10.5867/medwave.2007.03.3546.

[27] “ANGIOGÉNESIS - SEOM: Sociedad Española de Oncología Médica © 2019.” https://www.seom.org/43-Socios%20-%20Formaci%C3%B3n%20y%20Recursos/Bases%20de%20la%20Oncolog%C3%ADa/157-angiogenesis?start=1 (accessed May 17, 2021).

[28] M. Á. Castilla, S. Justo, and A. J. de Solis, “Response to hypoxia. A systemic mechanism based on the control of gene expression,” 2006. Accessed: May 03, 2021. [Online]. Available: https://www.researchgate.net/publication/7065522

[29] G. Lu and B. Fei, “Medical hyperspectral imaging: a review,” *Journal of Biomedical Optics*, vol. 19, no. 1, p. 010901, 2014, doi: 10.1117/1.JBO.19.1.010901.

[30] E. Sánchez Bernabé, J. Manuel, D. López, J. M. Rueda, and E. Secretario, “ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA Y SISTEMAS DE TELECOMUNICACIÓN PROYECTO FIN DE GRADO TÍTULO: Procesado de imágenes hiperespectrales Miembros del Tribunal Calificador: PRESIDENTE”.

[31] G. Alfonso and O. Lobo, “ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA Y BIOCIENCIAS NEKAZARITZAKO INGENIARITZAKO ETA BIOZIENTZIETAKO GOI MAILAKO ESKOLA TEKNIKOA”.

[32] A. Sahu *et al.*, "Characterization of mammary tumors using noninvasive tactile and hyperspectral sensors," *IEEE Sensors Journal*, vol. 14, no. 10, pp. 3337–3344, 2014, doi: 10.1109/JSEN.2014.2323215.

[33] V.-D. T *et al.*, "A hyperspectral imaging system for in vivo optical diagnostics. Hyperspectral imaging basic principles, instrumental systems, and applications of biomedical interest," *IEEE Eng Med Biol Mag*, vol. 23, no. 5, pp. 40–49, Sep. 2004, doi: 10.1109/MEMB.2004.1360407.

[34] R. B *et al.*, "Development of an image pre-processor for operational hyperspectral laryngeal cancer detection," *J Biophotonics*, vol. 9, no. 3, pp. 235–245, Mar. 2016, doi: 10.1002/JBIO.201500151.

[35] M. Hohmann *et al.*, "Preliminary results for hyperspectral videoendoscopy diagnostics on the phantoms of normal and abnormal tissues: Towards gastrointestinal diagnostics," *Optics InfoBase Conference Papers*, 2011, doi: 10.1117/12.889829.

[36] A. Malpica *et al.*, "Multispectral digital colposcopy for in vivo detection of cervical cancer," *Optics Express, Vol. 11, Issue 10, pp. 1223-1236*, vol. 11, no. 10, pp. 1223–1236, May 2003, doi: 10.1364/OE.11.001223.

[37] B. S. Sorg, B. J. Moeller, O. Donovan, Y. Cao, and M. W. Dewhirst, "Hyperspectral imaging of hemoglobin saturation in tumor microvasculature and tumor hypoxia development," *Journal of Biomedical Optics*, vol. 10, no. 4, p. 044004, 2005, doi: 10.1117/1.2003369.

[38] K. J. Zuzak *et al.*, "Hyperspectral imaging utilizing LCTF and DLP technology for surgical and clinical applications," *https://doi.org/10.1117/12.816279*, vol. 7170, pp. 71–79, Feb. 2009, doi: 10.1117/12.816279.

[39] S. BS, M. BJ, D. O, C. Y, and D. MW, "Hyperspectral imaging of hemoglobin saturation in tumor microvasculature and tumor hypoxia development," *J Biomed Opt*, vol. 10, no. 4, p. 044004, 2005, doi: 10.1117/1.2003369.

[40] L. E. Boucheron, Z. Bi, N. R. Harvey, B. S. Manjunath, and D. L. Rimm, "Utility of multispectral imaging for nuclear classification of routine clinical histopathology imagery," *BMC Cell Biology*, vol. 8, no. SUPPL. 1, Jul. 2007, doi: 10.1186/1471-2121-8-S1-S8.

[41] D. Hattery, M. Hassan, S. Demos, and A. Gandjbakhche, "Hyperspectral imaging of Kaposi's Sarcoma for disease assessment and treatment monitoring," *Proceedings - Applied Imagery Pattern Recognition Workshop*, vol. 2002-January, pp. 124–130, 2002, doi: 10.1109/AIPR.2002.1182265.

[42] D. DT *et al.*, "Differentiation of normal skin and melanoma using high resolution hyperspectral imaging," *Cancer Biol Ther*, vol. 5, no. 8, pp. 1033–1038, 2006, doi: 10.4161/CBT.5.8.3261.

[43] D. M. Roblyer *et al.*, "Multispectral optical imaging device for in vivo detection of oral neoplasia," *https://doi.org/10.1117/1.2904658*, vol. 13, no. 2, p. 024019, Mar. 2008, doi: 10.1117/1.2904658.

[44] H. Akbari, L. v. Halig, H. Zhang, D. Wang, Z. G. Chen, and B. Fei, "Detection of cancer metastasis using a novel macroscopic hyperspectral method," *Medical Imaging 2012: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 8317, p. 831711, Mar. 2012, doi: 10.1117/12.912026.

[45] C. JA, W. EC, and K. MR, "Evaluation of hyperspectral technology for assessing the presence and severity of peripheral artery disease," *J Vasc Surg*, vol. 54, no. 6, pp. 1679–1688, Dec. 2011, doi: 10.1016/J.JVS.2011.06.022.

[46] E. L. P. Larsen, L. L. Randeberg, E. Olstad, O. A. Haugen, A. Aksnes, and L. O. Svaasand, "Hyperspectral imaging of atherosclerotic plaques in vitro," *Journal of Biomedical Optics*, vol. 16, no. 2, p. 026011, 2011, doi: 10.1117/1.3540657.

[47] J. Schweizer, J. Hollmach, G. Steiner, L. Knels, R. H. W. Funk, and E. Koch, "Hyperspectral imaging - A new modality for eye diagnostics," *Biomedizinische Technik*, vol. 57, no. SUPPL. 1 TRACK-P, pp. 293–296, Sep. 2012, doi: 10.1515/BMT-2012-4375.

[48] P. SV *et al.*, "Medical hyperspectral imaging to facilitate residual tumor identification during surgery," *Cancer Biol Ther*, vol. 6, no. 3, pp. 439–446, 2007, doi: 10.4161/CBT.6.3.4018.

[49] Z. KJ, N. SC, A. G, H. D, B. K, and L. E, "Intraoperative bile duct visualization using near-infrared hyperspectral video imaging," *Am J Surg*, vol. 195, no. 4, pp. 491–497, Apr. 2008, doi: 10.1016/J.AMJSURG.2007.05.044.

[50] E. O. Olweny *et al.*, "Renal oxygenation during robot-assisted laparoscopic partial nephrectomy: Characterization using laparoscopic digital light processing hyperspectral imaging," *Journal of Endourology*, vol. 27, no. 3, pp. 265–269, Mar. 2013, doi: 10.1089/END.2012.0207.

[51] H. Akbari, Y. Kosugi, K. Kojima, and N. Tanaka, "Blood vessel detection and artery-vein differentiation using hyperspectral imaging," *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, pp. 1461–1464, 2009, doi: 10.1109/IEMBS.2009.5332920.

[52] M. E. Paolettia, J. M. Haut, J. Plaza, and A. Plaza, "A comparative study of techniques for hyperspectral image classification," *RIAI - Revista Iberoamericana de Automatica e Informatica Industrial*, vol. 16, no. 2, pp. 129–137, 2019, doi: 10.4995/riai.2019.11078.

[53] J. Theiler and G. Gisler, "A contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation".

[54] X. Wen and X. Yang, "An Unsupervised Classification Method for Hyperspectral Remote Sensing Image Based on Spectral Data Mining," *Advances in Data Mining Knowledge Discovery and Applications*, Sep. 2012, doi: 10.5772/50135.

[55] A. Fabisch, "gmr: Gaussian Mixture Regression," *Journal of Open Source Software*, vol. 6, no. 62, p. 3054, Jun. 2021, doi: 10.21105/joss.03054.

[56] Z. Zhang, X. Liu, and L. Wang, "Spectral Clustering Algorithm Based on Improved Gaussian Kernel Function and Beetle Antennae Search with Damping Factor," *Computational Intelligence and Neuroscience*, vol. 2020, 2020, doi: 10.1155/2020/1648573.

[57] M. R. Ackermann, J. Blömer, D. Kuntze, and C. Sohler, "Analysis of agglomerative clustering," in *Algorithmica*, May 2014, vol. 69, no. 1, pp. 184–215. doi: 10.1007/s00453-012-9717-4.

[58] F. Ramadhani, M. Zarlis, and S. Suwilo, "Improve BIRCH algorithm for big data clustering," in *IOP Conference Series: Materials Science and Engineering*, Jan. 2020, vol. 725, no. 1. doi: 10.1088/1757-899X/725/1/012090.

[59] M. E. Paolettia, J. M. Haut, J. Plaza, and A. Plaza, "A comparative study of techniques for hyperspectral image classification," *RIAI - Revista Iberoamericana de Automatica e Informatica Industrial*, vol. 16, no. 2, pp. 129–137, 2019, doi: 10.4995/riai.2019.11078.

[60] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to Statistical Learning Theory", Accessed: Oct. 26, 2021. [Online]. Available: http://www.kyb.mpg.de/~bousquethttp://www.lri.fr/~boucherohttp://www.econ.upf.es/~lugosi

[61] I. Gil Leiva, P. Díaz Ortuño, J. Vicente, and R. Muñoz, "Técnicas y usos en la clasificación automática de imágenes."

[62] G. Mercier and M. Lennon, "Support Vector Machines for Hyperspectral Image Classification with Spectral-based kernels," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2003, vol. 1, pp. 288–290. doi: 10.1109/igarss.2003.1293752.

[63] "TFG-2402-ARTOLA".

[64] E. Secretario, "ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA Y SISTEMAS DE TELECOMUNICACIÓN PROYECTO FIN DE GRADO."

[65] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery 1998 2:2*, vol. 2, no. 2, pp. 121–167, 1998, doi: 10.1023/A:1009715923555.

[66] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making 2019 19:1*, vol. 19, no. 1, pp. 1–16, Dec. 2019, doi: 10.1186/S12911-019-1004-8.

[67] N. Murali, A. Kucukkaya, A. Petukhova, J. Onofrey, and J. Chapiro, "Supervised Machine Learning in Oncology: A Clinician's Guide," *Dig Dis Interv*, vol. 4, no. 1, p. 73, Mar. 2020, doi: 10.1055/S-0040-1705097.

[68] I. J. Jacob, A. S. Nadhan, S. Jain, S. Gunti, and D. Sathya, "Predicting the possibility of cancer with supervised Learning Algorithms," *European Journal of Molecular & Clinical Medicine*, vol. 08, no. 1, p. 2021.

[69] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer," *Annals of Medicine and Surgery*, vol. 62, pp. 53–64, Feb. 2021, doi: 10.1016/J.AMSU.2020.12.043.

[70] "El IUMA ha logrado una herramienta hiperespectral que ayuda al diagnóstico para la detección de tumores cerebrales de algo grado | ULPGC - Universidad de Las Palmas de Gran Canaria." https://www.ulpgc.es/noticia/iuma-ha-logrado-herramienta-hiperespectral-que-ayuda-al-diagnostico-deteccion-tumores (accessed Jun. 01, 2021).

[71] R. Candia B and G. Caiozzi A, "Intervalos de confianza," *Revista Medica de Chile*, vol. 133, no. 9, pp. 1111–1115, Sep. 2005, doi: 10.25237/REVCHILANESTV43N02.11.

[72] "Estrategia de uno contra el resto para la clasificación de clases múltiples – Acervo Lima." https://es.acervolima.com/estrategia-de-uno-contra-el-resto-para-la-clasificacion-de-clases-multiples/ (accessed Aug. 21, 2022).

[73] "Precision, Recall, F1, Accuracy en clasificación - IArtificial.net." https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/ (accessed Aug. 21, 2022).

[74] H. Fabelo *et al.*, "Spatio-spectral classification of hyperspectral images for brain cancer detection during surgical operations," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, doi: 10.1371/journal.pone.0193721.

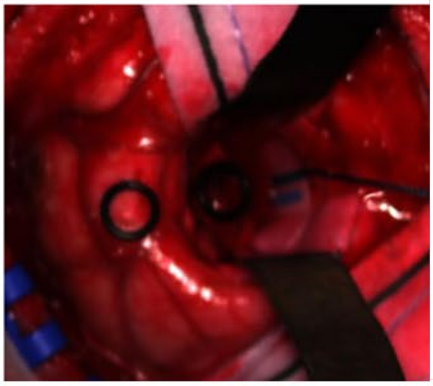
[75] B. S. Sorg, B. J. Moeller, O. Donovan, Y. Cao, and M. W. D. D.V.M., "Hyperspectral imaging of hemoglobin saturation in tumor microvasculature and tumor hypoxia development," *https://doi.org/10.1117/1.2003369*, vol. 10, no. 4, p. 044004, Jul. 2005, doi: 10.1117/1.2003369.

# Anexo: Imágenes de la base de datos

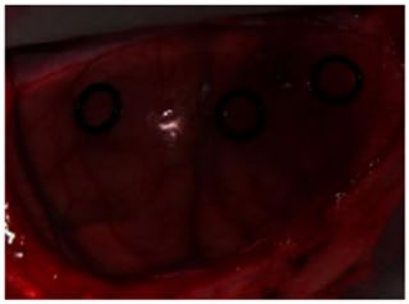
Op4C2

Op4C3

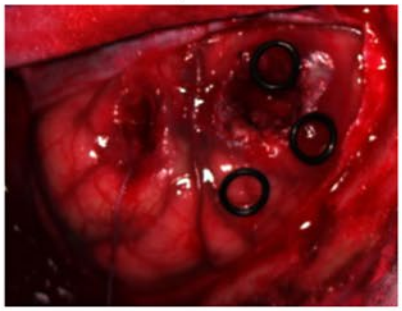
Op5C1



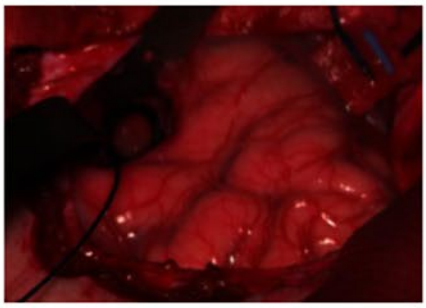
Op8C1



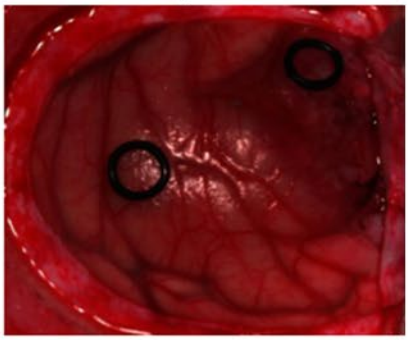
Op8C2



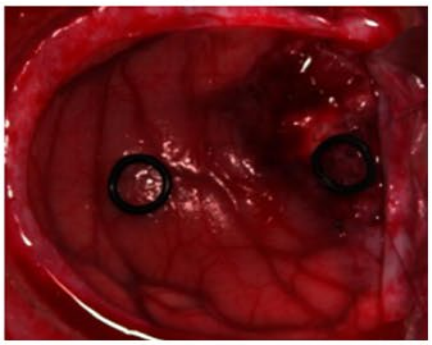
Op10C3



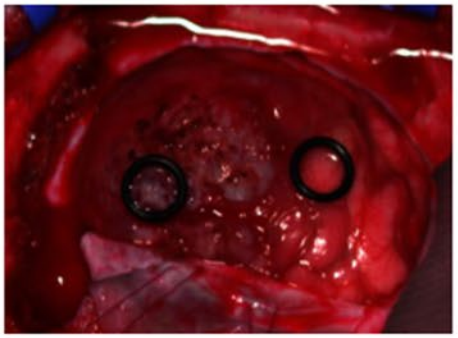
Op12C1



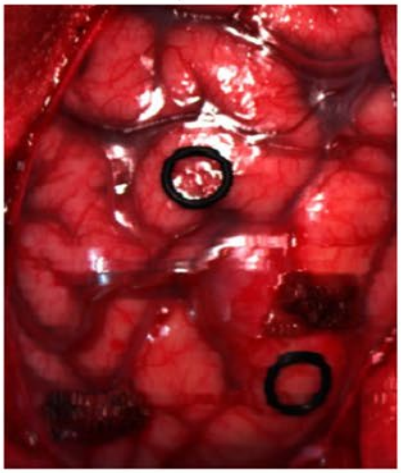
Op12C2



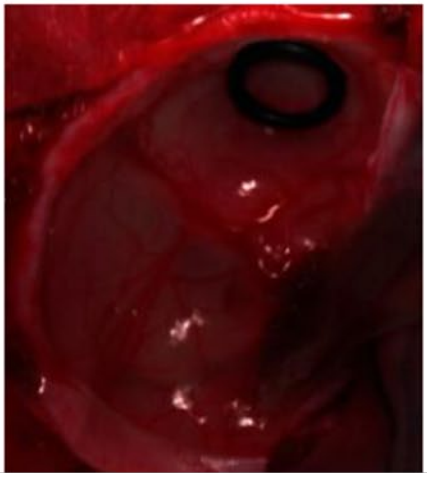
Op15C1



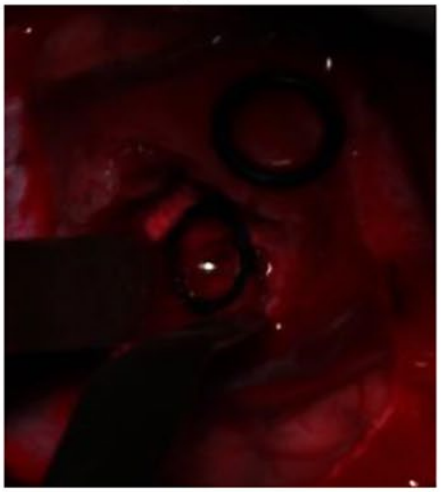
Op7C1



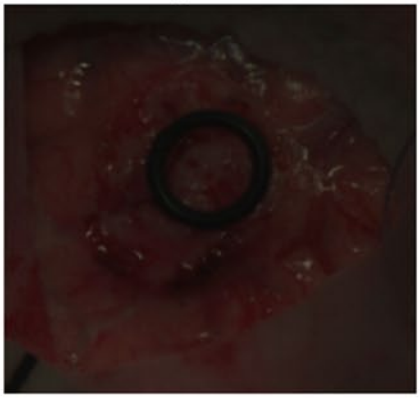
Op13C1



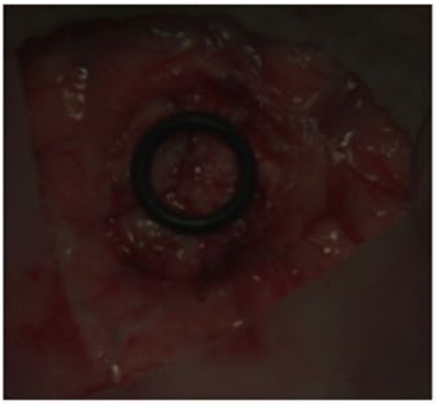
Op14C1



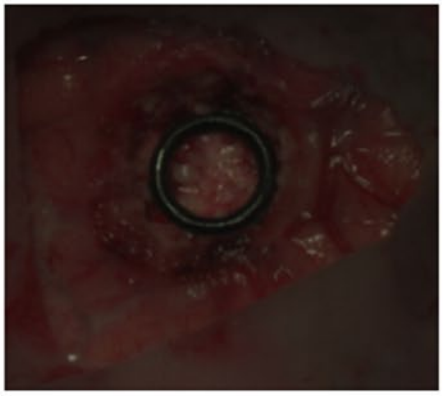
Op16C1



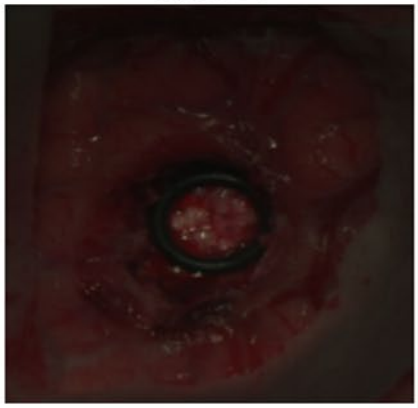
Op16C2



Op16C3

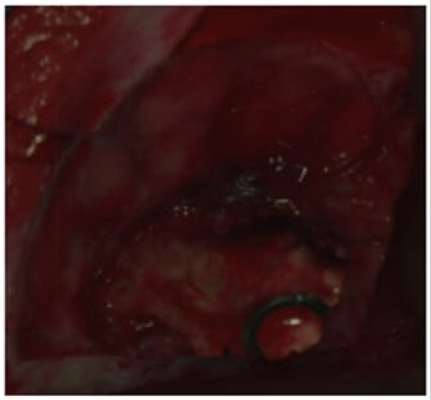


Op16C4

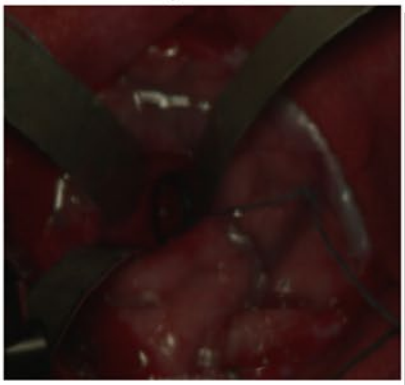


Op16C5

Op17C1

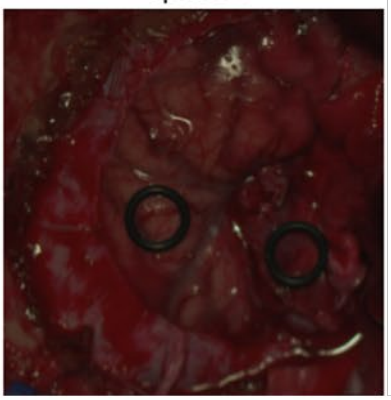


Op18C1

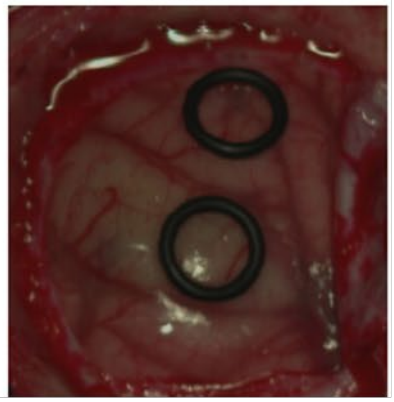


Op18C2

Op19C1

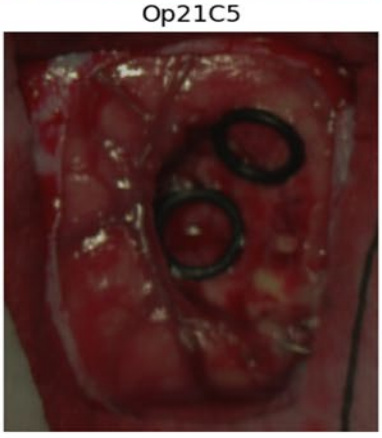


Op20C1

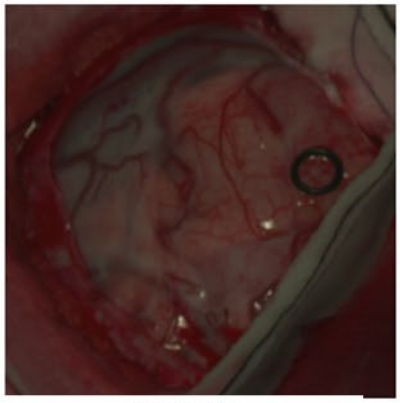


Op21C1

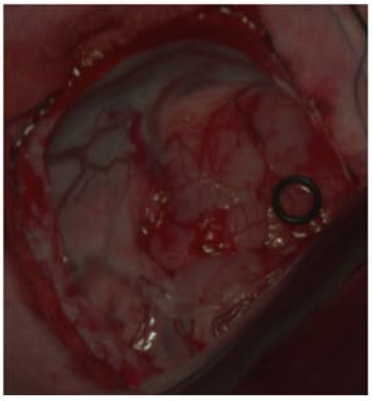
Op21C5



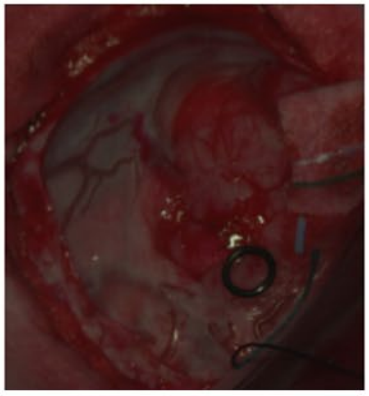
Op22C1



Op22C2



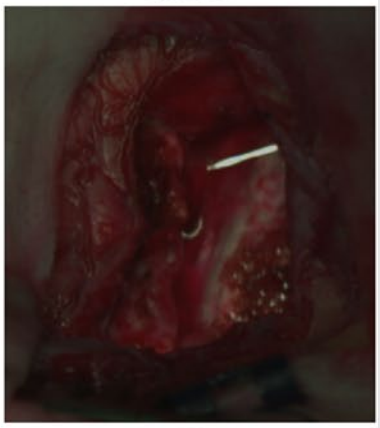
Op22C3



Op21C2

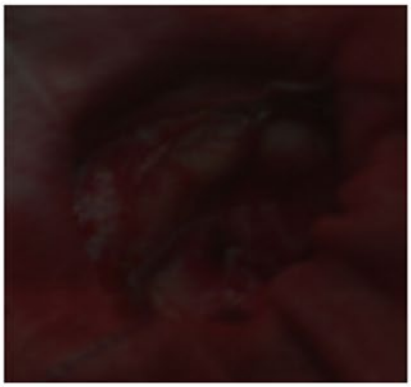
Op28C4

Op28C5

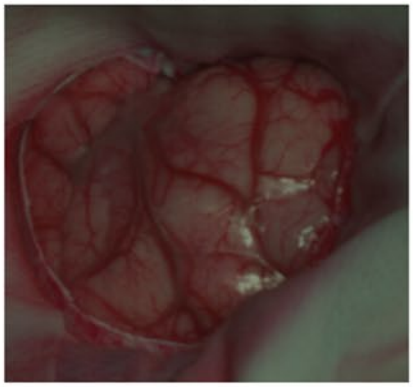


Op25C2

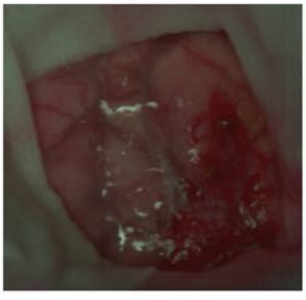
Op26C2



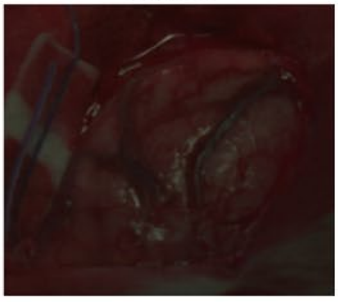
Op27C2



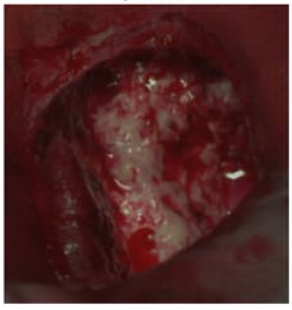
Op28C3



Op29C2



Op29C4



Op30C2

# Pliego de condiciones

En este apartado se expondrá el pliego de condiciones necesario para el desarrollo de este TFG. Lo dividiremos en dos partes *software* y *hardware.*

## 1.1. Recursos software

- vnirCameraController_v2: software creado por compañeros del IUMA para la captura de los cubos HS con la cámara Hyperspec® VNIR.
- Python 3.8: lenguaje de programación que permite realizar cálculos, realizar distintos modelos de clasificación y procesar imágenes entre otras muchas funciones. Dentro del lenguaje las librerías:
    - Matplotlib: librería de gráficos
    - statsmodels: librería de *datascience* (contiene los modelos de regresión lineal)
    - sklearn: libreria de librería (contiene los modelos de entrenamiento supervisado y no supervisado)
    - numpy : optimización para el manejo de datos
    - scipy: módulo de IO para leer las imágenes HS en formato mat
    - mpl_toolkits: funciones complementarias a matplotlib
    - pickle : librería para lectura y escritura de modelos
- Microsoft Office 365: conjunto de programas utilizados para la redacción, póster y defensa del TFG.
- Mendeley: aplicación para gestionar, definir y redactar las referencias bibliográficas utilizadas en este TFG.

## 1.2. Recursos hardware

- Cámara Hyperspec® VNIR: cámara HS que trabaja en el rango espectral VNIR de 400 a 1000 nm.
- Cámara Hyperspec® VNIR cámara HS que trabaja en el rango espectral VNIR de 400 a 1700 nm.
- Sistema de iluminación que cubre el rango espectral de 400 nm a 2200 nm
- Ordenador de sobremesa

# Presupuesto

En este capítulo se especifican los costes asociados a la elaboración del presente TFG. Los costes, por tanto, se encuentran desglosados en los siguientes grupos:

- Mano de obra
- Recursos *software*
- Recursos hardware
- Otros gastos asociados a la tramitación.

El valor de los costes asociados a la compra del material necesario para el desarrollo de este trabajo coincide con el precio de venta al público en el momento de su compra.

## 1.1. Mano de obra

En este apartado, la mano de obra cuenta como un ingeniero técnico junior atendiendo a la naturaleza de las tareas realizadas asociadas a este TFG. Con ello, la siguiente tabla especifica los costes asociados a la mano de obra para unas determinadas horas de trabajo:

*Tabla 5. Presupuesto mano de obra.*

| Concepto | Coste (€)/horas | Horas | Coste total |
|---|---|---|---|
| Ingeniero técnico | 8,9 | 300 | 2670,00 |

## 1.2. Recursos software

En este apartado se muestran los costes asociados a recursos *software* asociados a este trabajo. Para una mejor definición se muestra el tipo de licencia y su coste asociado:

*Tabla 6. Presupuesto recursos software.*

| Concepto | Tipo de licencia | Coste 8€) |
|---|---|---|
| Anaconda para Windows lenguaje de programación Python | Gratuito | 0,00 |
| Microsoft Office 365 | Universitaria | 0,00 |
| Mendeley | Gratuito | 0,00 |
| **Total** | | 0,00 |

## 1.3. Recursos hardware

En este apartado se presentan los costes asociados a recursos *hardware*. La siguiente tabla presenta los elementos hardware utilizados, una estimación de amortización, su coste por unidad y coste total. Para obtener el coste de utilización de los recursos hardware se ha usado un porcentaje de amortización de 7% como dato de referencia de otros proyectos. El coste total de los recursos hardware será de dos mil novecientos cinco euros:

*Tabla 7. Presupuesto hardware.*

| Concepto | Cantidad | Costes por unidad (€) | Amortización (%) | Coste de utilización (€) |
|---|---|---|---|---|
| Cámara Hyperspec® VNIR de 400 a 1.000nm | 1 | 40.000,00 | 7 | 2.800,00 |
| Cámara Hyperspec® VNIR de 400 a 1.700nm | 1 | 40.000,00 | 7 | 2.800,00 |
| Sistema de iluminación MI-150 | 1 | 400,00 | 7 | 400,00 |
| Ordenador de sobremesa | 1 | 1.500,00 | 7 | 105,00 |
| **Total** | | | | 6.105,00 |

## 1.4. Otros gastos

En este apartado se presentan los gastos asociados a la redacción de la memoria y la tramitación de la documentación. Para los gastos de tramitación supondremos el coste presentado en la Tabla 5 donde se presentan los costes de mano de obra, donde se cuenta con horas para la redacción del documento. Para la tramitación de la documentación se usará como referencia un 20% del precio de la asignatura (Ecuación 12), siendo este un total de treinta y ocho euros y cincuenta céntimos.

*Ecuación 12. Costes de tramitación.*

$$Precio\ de\ la\ asignatura * 20\% = Gastos\ de\ tramitación$$

## 1.5. Presupuesto total

Se presenta finalmente el presupuesto total de todos los costes presentados anteriormente aplicando el Impuesto General Indirecto Canario (IGIC) del 7%. El presupuesto que se puede apreciar en la Tabla 8, llega a un total de nueve mil cuatrocientos treinta euros con cuarenta y cinco céntimos.

*Tabla 8. Presupuesto total.*

| Concepto | Coste (€) |
|---|---|
| Mano de obra | 2.670,00 € |
| Recursos software | 0,00 € |
| Recursos hardware | 6.105,00 € |
| Otros gastos | 38,50 € |
| **Total** | **8.813,50 €** |
| IGIC | 7% |
| **Total con IGIC** | **9.430,45 €** |

Firma del presupuesto:

Kenya Espino Gutiérrez