

ESCUELA DE INGENIERÍA DE TELECOMUNICACIÓN Y ELECTRÓNICA



PROYECTO FIN DE CARRERA

IDENTIFICACIÓN DE TUMORES CEREBRALES USANDO DATOS ESPECTROGRÁFICOS

Autor: Ruymán Hernández López
Tutores: Gustavo Iván Marrero Callicó
Himar Fabelo Gómez
Fecha: Junio 2017

ESCUELA DE INGENIERÍA DE TELECOMUNICACIÓN Y ELECTRÓNICA



PROYECTO FIN DE CARRERA

IDENTIFICACIÓN DE TUMORES CEREBRALES USANDO DATOS ESPECTROGRÁFICOS

HOJA DE FIRMAS

Alumno

Fdo.: Ruymán Hernández López

Tutor

Tutor

Fdo.: Gustavo Iván Marrero Callicó

Fdo.: Himar Fabelo Gómez

Fecha: Junio 2017

ESCUELA DE INGENIERÍA DE TELECOMUNICACIÓN Y ELECTRÓNICA



PROYECTO FIN DE CARRERA

IDENTIFICACIÓN DE TUMORES CEREBRALES USANDO DATOS ESPECTROGRÁFICOS

HOJA DE EVALUACIÓN

Calificación: _____

Presidente

Fdo.:

Vocal

Fdo.:

Secretario/a

Fdo.:

Fecha: Junio 2017

*A mi hermana, a mi madre y a mi padre,
los tres pilares fundamentales sobre los que se erige mi formación.*



Agradecimientos

He decidido dedicar este trabajo a las personas con las que me siento más profundamente agradecido por plantar la semilla que ha originado mi crecimiento intelectual. Gracias infinitas a mis padres por inculcarme estos valores y a mi hermana por enseñarme a modelarlos; estos son los conocimientos más importantes que he adquirido en mi formación. Aunque cuando empecé esta carrera ni siquiera sabía escribir con el teclado, papá, tú fuiste la primera persona a la que ví implementando un algoritmo en una computadora. Tan solo era un niño, y ni tan siquiera entendía que introducir ese código que hacía rebotar el puntito en los límites de la pantalla, era programar. Es solo un ejemplo de la influencia imprescindible que has ejercido sobre mí a la hora de iniciarme en la comprensión de la tecnología en general. Mamá, no soy un médico, pero tal vez alguien pueda aprovechar una pequeña parte de lo que cuento en este trabajo para mejorar la vida de las personas. Fuiste tú la que cultivaste esta faceta de querer ayudar a otras personas. Fayna, tú eres mi gran maestra. Cada vez que te explico algo, lo meditamos y lo hablamos, eres tú quien me está enseñando a mí. Aunque no seas consciente de ello, no te imaginas lo que me has enseñado en la vida.

También quiero mostrar mi gratitud al resto de mi familia y a mis amigos. Se hace inviable mencionarlos a todos en un espacio tan reducido, pero a quienes me han apoyado en todo momento, tanto en el ámbito académico como en cualquier otro, en la distancia o en la cercanía, en los malos momentos y en los buenos; todos los que han estado siempre sin esperar nada a cambio sabrán que estas palabras van para ellos, muchas gracias por estar a mi lado. Gracias también a mis familiares y amigos que han dejado de existir y dejaron algo en mi interior. No me olvido de ustedes. Gracias a Ronia y a los demás animales que han pasado por mi vida porque también han contribuido a mi crecimiento.

Además, dar las gracias a todos los que han plantado alguna otra faceta de mi intelecto a lo largo de mi vida. A los profesores que me motivaron desde mi formación básica, a los profesores de secundaria que me inspiraron y me mostraron auténticos tesoros del conocimiento, a los profesores de la carrera que me instruyeron para adquirir los conocimientos más complejos, a los maestros de las destrezas no académicas y a mis tutores, Gustavo e Himar, por ayudarme y concederme la oportunidad de aplicar mis conocimientos en este proyecto.

Por último, aunque resulte un tanto áspero, siento la necesidad de dar las gracias a los que me han puesto barreras y obstáculos en el camino, ya que estas desavenencias me han hecho endurecerme y seguir adelante frente a todo.



Índice

Índice de figuras	XI
Índice de tablas.....	XVII
MEMORIA.....	XXI
RESUMEN	XXIII
ABSTRACT.....	XXV
ACRÓNIMOS.....	XXVII
CAPÍTULO 1 INTRODUCCIÓN	1
1.1 Introducción.....	1
1.1.1. Datos espectrográficos e imágenes hiperespectrales	3
1.1.2. Clasificadores.....	5
1.2 Antecedentes	7
1.3 Objetivos	8
1.4 Peticiónario	9
1.5 Organización del documento	9
1.5.1. Memoria.....	9
1.5.2. Presupuesto.....	10
1.5.3. Pliego de condiciones	11
CAPÍTULO 2 TUMORES CEREBRALES.....	13
2.1 Introducción.....	13
2.2 Anatomía del cerebro humano	14
2.2.1. Estructura del cerebro	14
2.2.2. Tipos de células del cerebro	17
2.2.2.1. Células microgliales	18
2.2.2.2. Astrocitos	18
2.2.2.3. Oligodendrocitos	20
2.2.2.4. Células ependimarias o ependimocitos.....	20

2.3	Tumores cerebrales	21
2.3.1.	Astrocitomas	23
2.3.2.	Oligodendrogliomas	24
2.3.3.	Oligoastrocitomas o gliomas mixtos.....	25
2.3.4.	Otros tumores gliales.....	25
2.3.5.	Tumores secundarios	26
2.4	Imágenes diagnósticas	26
2.5	Resumen	27
CAPÍTULO 3 CLASIFICACIÓN BASADA EN INFORMACIÓN ESPECTRAL.....		29
3.1	Introducción	29
3.2	Espectroscopía	30
3.3	Imágenes hiperespectrales.....	34
3.3.1.	Concepto de imagen hiperespectral.....	34
3.3.2.	La imagen hiperespectral aplicada a la detección de tumores	37
3.3.3.	Técnicas de análisis hiperespectral.....	40
3.4	Preprocesado de datos hiperespectrales.....	42
3.5	Clasificación de datos hiperespectrales	45
3.5.1.	Algoritmos de clasificación no supervisados.....	49
3.5.1.1.	K-Means	50
3.5.1.2.	ISODATA	50
3.5.2.	Algoritmos de clasificación supervisados.....	51
3.5.2.1.	Spectral Angle Mapper (SAM).....	52
3.5.2.2.	Multinomial Logistic Regression (MLR)	53
3.5.2.3.	Linear Discriminant Analysis (LDA)	55
3.5.2.4.	Markov Random Fields (MRFs)	62
3.5.2.5.	Random Forest (RF).....	65
3.5.2.6.	Artificial Neural Networks (ANNs).....	69
3.5.2.7.	Support Vectors Machines (SVMs)	78
3.6	Métodos de validación de algoritmos de clasificación	85
3.7	Resumen	89

CAPÍTULO 4	BASE DE DATOS DE LAS MUESTRAS ESPECTROGRÁFICAS.....	91
4.1	Introducción.....	91
4.2	Adquisición de los datos espectrográficos.....	92
4.2.1.	Muestras de la campaña 1	93
4.2.1.1.	Eliminación de muestras de la campaña 1	109
4.2.2.	Muestras de la campaña 2	111
4.3	Resumen.....	125
CAPÍTULO 5	MÉTODO DESARROLLADO	127
5.1	Introducción.....	127
5.2	Cadena de procesado de los datos espectrográficos.....	128
5.2.1.	Etapa de datos de entrada	128
5.2.2.	Etapa de preprocesado	129
5.2.3.	Etapa de clasificación	131
5.3	Proceso de clasificación	131
5.4	Métricas de validación	134
5.5	Árboles de clasificación.....	136
5.5.1.	Árbol de clasificación de la campaña 1	137
5.5.2.	Árbol de clasificación de la campaña 2	139
5.6	Sistemas de clasificación multiclase.....	141
5.7	Optimización de parámetros en los sistemas multiclase	142
5.7.1.	Optimización de parámetros en la clasificación de la campaña 1.....	144
5.7.2.	Optimización de parámetros en la clasificación de la campaña 2	146
5.7.3.	Optimización de parámetros en la clasificación de la fusión de las campañas	147
5.8	Resumen.....	147
CAPÍTULO 6	RESULTADOS EXPERIMENTALES	151
6.1	Introducción.....	151
6.2	Firmas espectrales de las muestras preprocesadas.....	152
6.3	Resultados en sistemas de árboles jerárquicos	158
6.3.1.	Resultados en sistemas de árboles jerárquicos (campaña 1)	159

6.3.1.1.	Diseño preliminar y comparativa	159
6.3.1.2.	Mejoras en el diseño	165
6.3.2.	Resultados en sistemas de árboles jerárquicos (campaña 2)	169
6.3.3.	Resumen de resultados más relevantes con sistemas de árboles jerárquicos.....	174
6.4	Resultados en sistemas de clasificación multiclase.....	177
6.4.1.	Resultados en sistemas de clasificación multiclase (campaña 1)	178
6.4.1.1.	Resultados de la optimización de parámetros del <i>kernel</i> RBF (campaña 1)	181
6.4.2.	Resultados en sistemas de clasificación multiclase (campaña 2 - transmitancia)	191
6.4.2.1.	Resultados de la optimización de parámetros del <i>kernel</i> RBF (campaña 2)	194
6.4.3.	Resultados en sistemas de clasificación multiclase (campaña 2 - absorbancia)	200
6.4.4.	Resultados en sistemas de clasificación multiclase (fusión de las campañas).....	203
6.4.4.1.	Resultados de la optimización de parámetros del <i>kernel</i> RBF (fusión de las campañas)	209
6.4.5.	Resumen de resultados más relevantes con sistemas multiclase	215
6.5	Resumen	220
CAPÍTULO 7 CONCLUSIONES Y LÍNEAS FUTURAS.....		221
7.1	Introducción	221
7.2	Análisis de los objetivos alcanzados	222
7.3	Conclusiones.....	223
7.4	Líneas futuras	226
7.5	Aplicabilidad del proyecto.....	227
BIBLIOGRAFÍA.....		229
PRESUPUESTO		239
P.1	Introducción	241
P.2	Recursos humanos	241
P.3	Recursos hardware.....	243
P.4	Recursos software	244
P.5	Material fungible	244
P.6	Redacción del proyecto	245
P.7	Aplicación de impuestos	245

P.8	Presupuesto total	246
PLIEGO DE CONDICIONES.....		247
PC.1	Introducción	249
PC.2	Recursos hardware	249
PC.3	Recursos software.....	249
PC.4	Recursos biológicos	250

Índice de figuras

Figura 1.1: Resonancia magnética de paciente afectado por un glioblastoma multiforme.	2
Figura 1.2: Firma espectral estelar de la luz de una estrella captada por un espectrógrafo [5].	3
Figura 1.3: Tipos de imágenes según la cantidad de bandas espectrales.....	4
Figura 1.4: Representación esquemática del cubo hiperespectral e ilustración de la firma espectral.	5
Figura 2.1: Vista esquemática anterior del cerebro.....	15
Figura 2.2: Lóbulos del cerebro.	15
Figura 2.3: Estructura de una neurona	17
Figura 2.4: Células gliales del Sistema Nervioso Central [13].....	21
Figura 3.1: Dispersión de luz en un prisma triangular.....	30
Figura 3.2: Diagrama de niveles de energía en una molécula.....	31
Figura 3.3: Radiación incidente, absorbida y transmitida por un compuesto químico.	32
Figura 3.4: Espectros de absorción de tres pigmentos fotosintéticos cloroplastídicos.	33
Figura 3.5: Espectro electromagnético.....	34
Figura 3.6: Imagen multidimensional de 4 bandas.	35
Figura 3.7: Comparación entre un cubo hiperespectral y una imagen RGB [32].....	36
Figura 3.8: Dimensiones espaciales (x, y) y dimensión espectral del cubo hiperespectral.....	39
Figura 3.9: Procedimiento de análisis hiperespectral.	41
Figura 3.10: Ejemplo de clasificación binaria de imágenes hiperespectrales (valores ficticios).....	47
Figura 3.11: Etapa de entrenamiento de un clasificador supervisado.	51
Figura 3.12: Etapa de verificación o test, de un clasificador supervisado.	51
Figura 3.13: Representación en dos dimensiones del concepto de ángulo espectral.	52
Figura 3.14: Las muestras de dos dimensiones son proyectadas en una dimensión espacial menor (línea), con lo que sus proyecciones sobre la línea son escalares.	56
Figura 3.15: Discriminante de Fisher para un problema biclase.	56
Figura 3.16: Interpretación geométrica de la proyección del vector x	57
Figura 3.17: Proyección de muestras bidimensionales en diferentes ejes de proyección.....	57
Figura 3.18: Ejemplo biclase y 2-dimensional de la proyección de las medias en ejes diferentes. ...	58
Figura 3.19: Nuevo eje de proyección.	59
Figura 3.20: Discriminante de Fisher para una problema de 3 clases.	60
Figura 3.21: Ejemplos de distribuciones que presentan problemas en clasificadores LDA.	61
Figura 3.22: Ejemplo de distribución de muestras de dos clases en el que la información de discriminación puede estar en la media o la varianza.....	61
Figura 3.23: Cálculo de la energía de un vóxel.....	64
Figura 3.24: Ejemplo de clasificación realizada por un árbol de decisión en teledetección.	68

Figura 3.25: Ejemplo de red neuronal.	70
Figura 3.26: Modelo de neurona de McCulloch y Pitts (MCP).	71
Figura 3.27: Ejemplo de una red sencilla <i>feed-forward</i>	72
Figura 3.28: Ejemplo de una red <i>feed-back</i>	73
Figura 3.29: <i>Perceptrón</i>	74
Figura 3.30: Esquema de funcionamiento del clasificador SVM.	79
Figura 3.31: Ejemplo de datos definidos en $-1, 1 \times -1, 1 \in \mathbb{R}^2$ en el que se ilustra una posible superficie de decisión no lineal.	82
Figura 3.32: Ejemplo de aplicación de SVM en neuroimagen.	85
Figura 3.33: Ejemplo de un mapa temático de clasificación a partir de la verdad terrena asociada a la imagen a clasificar.	87
Figura 4.1: Espectrómetro Varian FT-IR 600 [108].	93
Figura 4.2: Firmas espectrales de muestras asociadas al paciente 115 [4].	94
Figura 4.3: Firma espectral de la muestra de astrocitoma pilocítico (Grado I).	97
Figura 4.4: Firmas espectrales, media y desviación típica de las muestras de astrocitoma difuso (Grado II).	98
Figura 4.5: Firmas espectrales, media y desviación típica de las muestras de oligodendroglioma (Grado II).	99
Figura 4.6: Firmas espectrales, media y desviación típica de las muestras de astrocitoma anaplásico (Grado III).	100
Figura 4.7: Firmas espectrales, media y desviación típica de las muestras de oligodendroglioma anaplásico (Grado III).	101
Figura 4.8: Firmas espectrales, media y desviación típica de las muestras de “GBM mixed astro/oligo” (Grado IV).	102
Figura 4.9: Firmas espectrales, media y desviación típica de las muestras de “GBM with gliosarcomatous” (Grado IV).	103
Figura 4.10: Firmas espectrales, media y desviación típica de las muestras de “GBM astro” (Grado IV).	104
Figura 4.11: Firmas espectrales, media y desviación típica de las muestras de “IV gliosarcoma” (Grado IV).	105
Figura 4.12: Firmas espectrales, media y desviación típica de las muestras de glioblastoma multiforme (Grado IV).	106
Figura 4.13: Firmas espectrales, media y desviación típica de las muestras de sustancia blanca (Tejido sano).	107
Figura 4.14: Firmas espectrales, media y desviación típica de las muestras de sustancia gris (Tejido sano).	108
Figura 4.15: Espectrómetro PerkinElmer FT-IR System, Spectrum BX.	111

Figura 4.16: Firmas espectrales, media y desviación típica de las muestras de tumor glioneuronal (Grado I).....	115
Figura 4.17: Firmas espectrales, media y desviación típica de las muestras de astrocitoma difuso (Grado II).	116
Figura 4.18: Firmas espectrales, media y desviación típica de las muestras de oligodendroglioma (Grado II).	117
Figura 4.19: Firmas espectrales, media y desviación típica de las muestras de astrocitoma anaplásico (Grado III).....	118
Figura 4.20: Firmas espectrales, media y desviación típica de las muestras de “GBM with oligo” (Grado IV).	119
Figura 4.21: Firmas espectrales, media y desviación típica de las muestras de glioblastoma multiforme (Grado IV).....	120
Figura 4.22: Firmas espectrales de las muestras de glioblastoma multiforme (Grado IV) por paciente.	121
Figura 4.23: Firmas espectrales, media y desviación típica de las muestras de adenocarcinoma pulmonar (Metástasis).....	122
Figura 4.24: Firmas espectrales, media y desviación típica de las muestras de carcinoma gástrico (Metástasis).....	123
Figura 4.25: Firmas espectrales, media y desviación típica de las muestras de epilepsia con esclerosis del hipocampo (no tumoral).	124
Figura 4.26: Firmas espectrales de las muestras no tumorales (pacientes 17 y 21).	125
Figura 5.1: Etapas de la cadena de procesamiento empleada en este proyecto.	128
Figura 5.2: Proceso de clasificación.....	132
Figura 5.3: Árbol de clasificación de la campaña 1.....	137
Figura 5.4: Árbol de clasificación de la campaña 2.	140
Figura 6.1: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de grado I (campaña 1 y campaña 2) y sus respectivas medias.	153
Figura 6.2: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de grado II (campaña 1 y campaña 2) y sus respectivas medias.....	154
Figura 6.3: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de grado III (campaña 1 y campaña 2) y sus respectivas medias.	155
Figura 6.4: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de grado IV (campaña 1 y campaña 2) y sus respectivas medias.....	156
Figura 6.5: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de metástasis (campaña 2) y sus respectivas medias.....	157
Figura 6.6: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras no tumorales (campaña 1 y campaña 2) y sus respectivas medias.	158

Figura 6.7: <i>Clasificador 1</i> del árbol jerárquico de la campaña 1.	160
Figura 6.8: <i>Clasificador 2.1</i> del árbol jerárquico de la campaña 1.	161
Figura 6.9: <i>Clasificador 2.2</i> del árbol jerárquico de la campaña 1.	162
Figura 6.10: <i>Clasificador 3</i> del árbol jerárquico de la campaña 1.	163
Figura 6.11: <i>Clasificador 1</i> del árbol jerárquico de la campaña 2.	169
Figura 6.12: <i>Clasificador 2</i> del árbol jerárquico de la campaña 2.	170
Figura 6.13: <i>Clasificador 3</i> del árbol jerárquico de la campaña 2.	171
Figura 6.14: <i>Clasificador 4.1</i> del árbol jerárquico de la campaña 2.	172
Figura 6.15: <i>Clasificador 4.2</i> del árbol jerárquico de la campaña 2.	173
Figura 6.16: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 1).	182
Figura 6.17: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 2).	183
Figura 6.18: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 3).	184
Figura 6.19: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 4).	185
Figura 6.20: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 5).	186
Figura 6.21: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 6).	187
Figura 6.22: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 7).	188
Figura 6.23: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 8 - ajuste fino).	190
Figura 6.24: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 1).	196
Figura 6.25: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 2).	197
Figura 6.26: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 3 - ajuste fino).	198
Figura 6.27: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test barrido 1.	210
Figura 6.28: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test barrido 2.	211
Figura 6.29: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 3 - ajuste fino).	212

Figura 6.30: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 4 - ajuste fino).....	213
Figura 7.1: Promedio de sensibilidad de las clases finales en los sistemas más relevantes del proyecto (campana 1).....	224
Figura 7.2: Promedio de sensibilidad de las clases finales en los sistemas más relevantes del proyecto (campana 2).	225
Figura 7.3: Promedio de sensibilidad de las clases finales en los sistemas más relevantes del proyecto (campana fusión).	226

Índice de tablas

Tabla 3.1: Definición de los rangos espectrales (en nanómetros de la radiación electromagnética (desde la región UV hasta la región MIR).	38
Tabla 3.2: Matriz de confusión del ejemplo ilustrado en la Figura 3.33.	87
Tabla 3.3: Matriz de confusión para la clasificación biclase de ámbito médico.	88
Tabla 4.1: Relación de muestras proporcionadas por el Hospital Universitario de Southampton (campaña 1).	96
Tabla 4.2: Muestras válidas de la campaña 1 empleadas en este proyecto.....	110
Tabla 4.3: Relación de muestras proporcionadas por el Hospital Universitario de Southampton (campaña 2). Todas las muestras de esta campaña han sido empleadas en este proyecto.....	113
Tabla 5.1: Matriz de confusión considerando una clase concreta bajo test.	135
Tabla 6.1: Número de muestras de entrenamiento en el diseño preliminar del árbol de clasificación (Campaña 1).	159
Tabla 6.2: Resultados del árbol de clasificación en los PFCs actual y precedente en el <i>Clasificador 1</i> (campaña 1).	160
Tabla 6.3: Resultados del árbol de clasificación en los PFCs actual y precedente en el <i>Clasificador 2.1</i> (campaña 1).	161
Tabla 6.4: Resultados del árbol de clasificación en los PFCs actual y precedente en el <i>Clasificador 2.2</i> (campaña 1).	162
Tabla 6.5: Resultados del árbol de clasificación en los PFCs actual y precedente en el <i>Clasificador 3</i> (campaña 1).	163
Tabla 6.6: Resultados del árbol de clasificación en el <i>Clasificador 1</i> para 70% y 30% de muestras de entrenamiento (campaña 1).	166
Tabla 6.7: Resultados del árbol de clasificación en el <i>Clasificador 2.1</i> para 70% y 30% de muestras de entrenamiento (campaña 1).	167
Tabla 6.8: Resultados del árbol de clasificación en el <i>Clasificador 2.2</i> para 70% y 30% de muestras de entrenamiento (campaña 1).	167
Tabla 6.9: Resultados del árbol de clasificación en el <i>Clasificador 3</i> para 70% y 30% de muestras de entrenamiento (campaña 1).	168
Tabla 6.10: Resultados del árbol de clasificación en el <i>Clasificador 1</i> para el 70% de muestras de entrenamiento (campaña 2).	169
Tabla 6.11: Resultados del árbol de clasificación en el <i>Clasificador 2</i> para el 70% de muestras de entrenamiento (campaña 2).	170

Tabla 6.12: Resultados del árbol de clasificación en el <i>Clasificador 3</i> para el 70% de muestras de entrenamiento (campaña 2).....	171
Tabla 6.13: Resultados del árbol de clasificación en el <i>Clasificador 4.1</i> para el 70% de muestras de entrenamiento (campaña 2).....	172
Tabla 6.14: Resultados del árbol de clasificación en el <i>Clasificador 4.2</i> para el 70% de muestras de entrenamiento (campaña 2).....	173
Tabla 6.15: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad, en cada uno de los clasificadores para <i>trainings</i> del 30% y 70% (campaña 1).	175
Tabla 6.16: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad, en cada uno de los clasificadores para un training del 70% (campaña 2).	176
Tabla 6.17: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> lineal (campaña 1 - absorbancia).....	179
Tabla 6.18: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF (campaña 1 - absorbancia).	180
Tabla 6.19: Pares de medias de sensibilidad y especificidad con la clase "grado IV" bajo test, que mejores resultados presentan, y los parámetros que los proporcionan.....	189
Tabla 6.20: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF y parámetros óptimos (campaña 1 - absorbancia).	191
Tabla 6.21: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> lineal (campaña 2 - transmitancia).	192
Tabla 6.22: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF (campaña 2 - transmitancia).	193
Tabla 6.23: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (para la clase "grado IV"), con tipo de <i>kernel</i> RBF y parámetros óptimos de la campaña 1 (campaña 2 - transmitancia).	195
Tabla 6.24: Pares de medias de sensibilidad y especificidad con la clase "grado IV" bajo test, que mejores resultados presentan, y los parámetros que los proporcionan.....	199
Tabla 6.25: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF y parámetros óptimos (campaña 2 - transmitancia).	200
Tabla 6.26: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> lineal (campaña 2 - absorbancia).	201
Tabla 6.27: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF (campaña 2 - absorbancia).	202

Tabla 6.28: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de kernel RBF y parámetros óptimos obtenidos para esta campaña en transmitancia (campaña 2 - absorbanca).	203
Tabla 6.29: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> lineal (campaña fusión - absorbanca).	204
Tabla 6.30: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> lineal (campaña fusión - transmitancia).	205
Tabla 6.31: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF (campaña fusión - absorbanca y transmitancia).	206
Tabla 6.32: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF y parámetros óptimos obtenidos para la campaña 1 en absorbanca (campaña fusión - absorbanca).	207
Tabla 6.33: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF y parámetros óptimos obtenidos para la campaña 2 en transmitancia (campaña fusión - absorbanca).	208
Tabla 6.34: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de kernel RBF y parámetros óptimos obtenidos para la campaña 1 en absorbanca (campaña fusión - transmitancia).	209
Tabla 6.35: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad (por clase), con tipo de <i>kernel</i> RBF y parámetros óptimos obtenidos para la fusión de las campañas en absorbanca (campaña fusión - transmitancia).	215
Tabla 6.36: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad; en las clasificaciones realizadas por los sistemas multiclase con tipo de <i>kernel</i> lineal.	218
Tabla 6.37: Promedio y desviación típica de las métricas: <i>overall accuracy</i> , sensibilidad y especificidad; en las clasificaciones realizadas por los sistemas multiclase con tipo de <i>kernel</i> RBF y los parámetros óptimos obtenidos en sus respectivos estudios.	219
Tabla P . 1: Factor de corrección en función de horas trabajadas.	242
Tabla P . 2: Coste total de recursos humanos.	242
Tabla P . 3: Coste total de recursos hardware.	243
Tabla P . 4: Coste total de recursos software.	244
Tabla P . 5: Coste total del Proyecto Fin de Carrera.	246

MEMORIA



Resumen

En la presente memoria se sintetiza un trabajo de investigación realizado como Proyecto Fin de Carrera (PFC), que tiene como principal objetivo desarrollar un sistema clasificador que permita diferenciar entre muestras tumorales y no tumorales a partir de datos espectrográficos.

Un diagnóstico temprano de las enfermedades ocasionadas por tumores cerebrales puede ser vital para evitar que un tumor benigno evolucione en un cáncer mucho más agresivo, y una correcta *resección* (eliminación) del tumor evita que este se vuelva a reproducir. Sin embargo, aumentar la zona de resección con un margen de seguridad adicional puede generar daños cerebrales importantes. Es por ello que se hacen necesarias técnicas que determinen de forma precisa y mínimamente invasiva los límites del tumor.

En la actualidad, los avances tecnológicos en la captura de imágenes han aumentado considerablemente la resolución espectral de estas, de este modo es posible identificar el tipo de sustancia de forma automática. En el ámbito médico es posible realizar diagnósticos mediante imágenes de alta resolución espectral (*imágenes hiperespectrales*), por tanto, esta línea de investigación se centra en identificar correctamente los tipos de tejidos cerebrales.

En este trabajo se realiza un estudio sobre diferentes tipos de tumores cerebrales, espectroscopía y sobre algoritmos de clasificación, prestando especial atención a las técnicas utilizadas comúnmente en la clasificación de imágenes hiperespectrales. La idea es implementar sistemas de clasificación usando la tecnología más apropiada para clasificar diferentes tipos de muestras tumorales y sanas. En concreto, se emplea una base de datos proporcionada por el Hospital Universitario de Southampton (Reino Unido) que consiste en respuestas espectrales de absorción y transmitancia para diferentes tipos de tejidos extraídos a pacientes con tumores cerebrales en intervenciones quirúrgicas. Este tipo de datos suponen el caso límite de resolución espacial (un único píxel) pero en contraste, presentan gran resolución y rango espectral.

Al final del este trabajo se dispondrá de varios sistemas diseñados a través de diferentes métodos, capaces de diferenciar entre tejidos sanos y tumorales, de distintos tipos y grados.



Abstract

In this Final Project, a research study is synthetised, which has as main goal, the development of a classifier system that is able to discriminate between samples of tumour and non-tumour tissue from spectrographic data.

Early diagnosis of illnesses produced by brain tumours could be of vital importance in order to avoid that a benign tumour progresses to a malignant tumour, and a proper resection of the tumour can avoid it to be reproduced again. However, increasing the region of resection with an additional security margin can cause important brain damage. So minimally invasive technics that accurately determine the tumour limits are needed.

Nowadays, due to the technological advances, the spectral resolution in capturing images has increased dramatically, allowing the identification of different types of substances automatically. In the medical field, it is possible to make a diagnosis by using high spectral resolution imaging (*hyperspectral imaging*). This research line focuses on identifying correctly the different types of brain tissues.

In this project, a survey is carried out about different types of brain tumour, spectroscopy and classification algorithms, with particular focus on technics commonly used in classification of hyperspectral images. The idea is to implement classification systems in the most suitable technology to classify different types of tumours and healthy samples. In particular, a database provided by the University Hospital Southampton (United Kingdom) has been used, which consists in absorption and transmittance spectral responses of different types of tissues obtained from patients with brain tumours during removal surgeries. This type of data is the limit case in spatial resolution (a single pixel), but on contrast, it provides a great spectral resolution and range.

At the end of this project, several systems, designed using different methods, able to discriminate between both tumour and healthy tissues, of different types and grades, will be provided.



Acrónimos

ANN	<i>Artificial Neural Networks</i>
AVIRIS	<i>Airborne Visible Infrared Imaging Spectrometer</i>
CAT	<i>Computerized Axial Tomography</i>
COIT	<i>Colegio Oficial de Ingenieros de Telecomunicación</i>
CPU	<i>Central Processing Unit</i>
CQP	<i>Convex Quadratic Programming</i>
DSI	<i>Diseño de Sistemas Integrados</i>
EITE	<i>Escuela de Ingeniería de Telecomunicación y Electrónica</i>
ERM	<i>Empirical Risk Minimization</i>
FS	<i>Feature Selection</i>
GHMRF	<i>Gaussian Hidden Markov Random Field</i>
HELICoiD	<i>HypErspectraL Imaging Cancer Detection</i>
HMRF	<i>Hidden Markov Random Field</i>
HYDICE	<i>Hyperspectral Digital Imagery Collection Experiment</i>
ISODATA	<i>Interactive Self-Organizing Data Analysis Techniques</i>
IUMA	<i>Instituto Universitario de Microelectrónica Aplicada</i>
KBr	<i>Bromuro de potasio</i>
LDA	<i>Linear Discriminant Analysis</i>
MAP	<i>Máximo a posteriori</i>
MCP	<i>McCulloch-Pitts</i>
MIR	<i>Mid-Infrared</i>
MLR	<i>Multinomial Logistic Regression</i>
MNF	<i>Minimum Noise Fraction</i>
MRF	<i>Markov Random Field</i>
MRI	<i>Magnetic Resonance Imaging</i>
ND	<i>Nivel Digital</i>
NHS	<i>National Health Service</i>
NIR	<i>Near-infrared</i>
OA	<i>Overall Accuracy</i>

PCA	<i>Principal Component Analysis</i>
PET	<i>Positron Emission Tomography</i>
PFC	<i>Proyecto Fin de Carrera</i>
RBF	<i>Radial Basis Function</i>
RF	<i>Random Forest</i>
SAM	<i>Spectral Angle Mapper</i>
SMLR	<i>Sparse Multinomial Logistic Regression</i>
SNC	<i>Sistema Nervioso Central</i>
SNP	<i>Sistema Nervioso Periférico</i>
SNR	<i>Signal to Noise Ratio</i>
SPECT	<i>Single-Photon Emission Computed Tomography</i>
SRM	<i>Structural Risk Minimization</i>
SVM	<i>Support Vectors machines</i>
TAC	<i>Tomografía Axial Computarizada</i>
UI	<i>Ultrasonic Imaging</i>
ULPGC	<i>Universidad de Las Palmas de Gran Canaria</i>
UV	<i>Ultraviolet spectrum</i>
VIS	<i>Visible spectrum</i>
WHO	<i>World Health Organization</i>

Capítulo 1

Introducción

1.1 Introducción

Debido al aumento en las últimas décadas de la incidencia y la mortalidad por tumor cerebral en la población mundial, el número de trabajos de investigación relacionados con el diagnóstico y el tratamiento de éstos ha crecido exponencialmente. Un diagnóstico temprano de este tipo de enfermedades puede ser vital para evitar que un tumor benigno evolucione en un cáncer mucho más agresivo.

Los tumores cerebrales se generan debido a un crecimiento anómalo de células derivadas de componentes cerebrales, en el caso de tumores primarios, o de células tumorales localizadas en otras partes del organismo, si se trata de metástasis. Según su evolución, extensión y acotación se pueden agrupar en cuatro grados de peligrosidad según la organización mundial de la salud (*World Health Organization, WHO*). Esta clasificación abarca desde los tumores de grado I, tumores benignos de crecimiento lento, a los tumores de grado IV, cánceres malignos con un crecimiento muy rápido [1].

El 77% de los cánceres cerebrales malignos pertenece al grupo de tumores denominados *gliomas*. Este tipo de tumores afectan a las células *gliales* encargadas del soporte de las neuronas y del procesamiento cerebral de la información. Dependiendo del

tipo de células gliales afectadas (*astrocitos, oligodendrocitos, ependimocitos*, etc.), existen diferentes tipos de gliomas, siendo el *glioblastoma multiforme* (*astrocitoma* de grado IV) el más agresivo y con menor probabilidad de supervivencia de todos ellos [2, 3] (Figura 1.1).

Los síntomas y el tratamiento de este tipo de tumores dependen en gran parte de la edad del paciente, del tipo del tumor y de su localización en el cerebro. Sin embargo, estos tumores tienden a infiltrarse en el tejido cerebral sano, por lo que su extirpación quirúrgica es compleja y en ocasiones imposible. Es por ello que es de vital importancia la investigación y el estudio de técnicas no invasivas que faciliten una diagnosis lo más exacta posible.

Una correcta *resección* (eliminación) del tumor cerebral elimina el tejido maligno y evita que el tumor se vuelva a reproducir. Aumentar la zona de resección con un margen adicional de seguridad maximiza las probabilidades de supervivencia del paciente, pero en el caso de tumores cerebrales, una resección excesiva puede generar daños importantes de todo tipo (motrices, cognitivos, visuales, etc.). Es por ello que se hacen necesarias técnicas que determinen de forma precisa y mínimamente invasiva los límites del glioma.

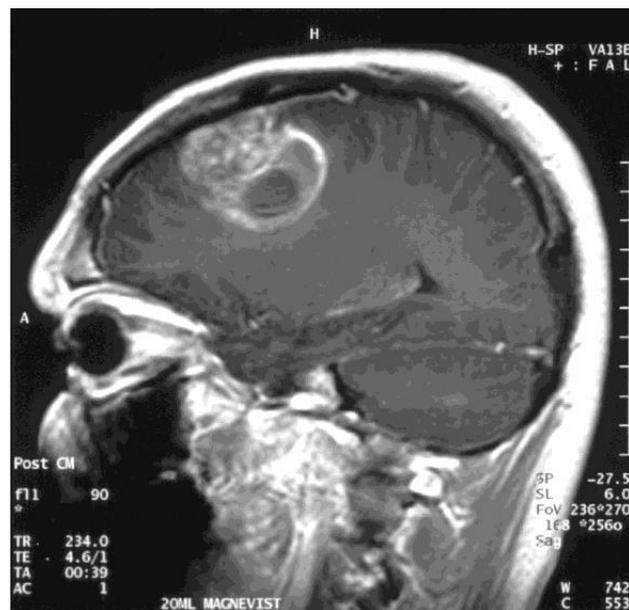


Figura 1.1: Resonancia magnética de paciente afectado por un glioblastoma multiforme.

Actualmente existen diversas técnicas que emplean imágenes diagnósticas para ayudar a los médicos a detectar o tratar enfermedades y lesiones neurológicas, como es el caso de las *Resonancias Magnéticas* o las *Tomografías Axiales Computarizadas* [4]. Sin embargo, este *Proyecto Fin de Carrera (PFC)* propone la utilización de una nueva técnica

de diagnóstico basada en el procesamiento de imágenes hiperespectrales de los tejidos cerebrales.

1.1.1. Datos espectrográficos e imágenes hiperespectrales

Los dispositivos encargados de la captura del espectro de un determinado cuerpo físico se llaman espectrógrafos. Estos dispositivos descomponen la luz blanca en las longitudes de onda de sus componentes formando el espectro del cuerpo físico objeto de estudio. El conjunto de valores de dicho espectro dan lugar al dato espectrográfico o firma espectral (Figura 1.2), el cual es registrado para su posterior análisis.

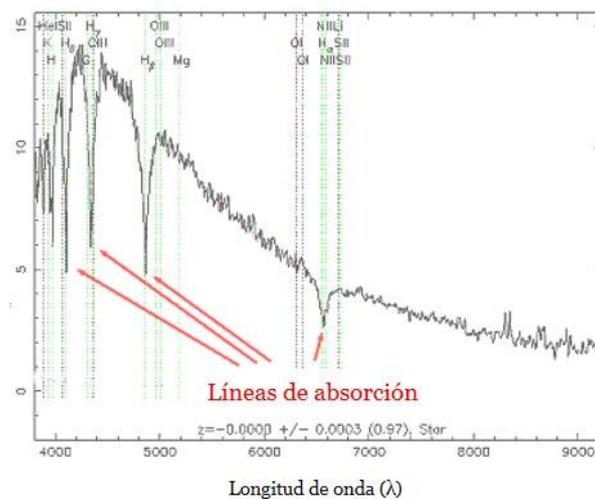


Figura 1.2: Firma espectral estelar de la luz de una estrella captada por un espectrógrafo [5].

Por otro lado, en los últimos años los dispositivos encargados de la captura de imágenes han experimentado un gran desarrollo tecnológico. Actualmente es posible encontrar cámaras con una gran resolución espacial, por ejemplo, en el ámbito doméstico ya son habituales las cámaras fotográficas de varias decenas de *megapíxeles*. Las cámaras fotográficas comerciales permiten obtener imágenes mediante sensores que capturan la luz de unas pocas longitudes de onda pertenecientes al espectro visible; habitualmente las longitudes correspondientes a las longitudes de onda del color rojo, verde y azul (imagen *RGB*).

No obstante, esta revolución tecnológica también ha permitido aumentar la resolución espectral de las cámaras. Se han desarrollado sensores más sofisticados capaces de capturar la radiación de una mayor cantidad de longitudes de onda que se extienden más allá del espectro visible. Este tipo de imágenes se denominan *multiespectrales*, cuando las

imágenes constan de decenas de longitudes de onda o bandas espectrales; hiperespectrales, cuando la cantidad de bandas espectrales asciende a cientos; y *ultraespectrales*, en el caso de imágenes compuestas por miles de bandas (Figura 1.3).

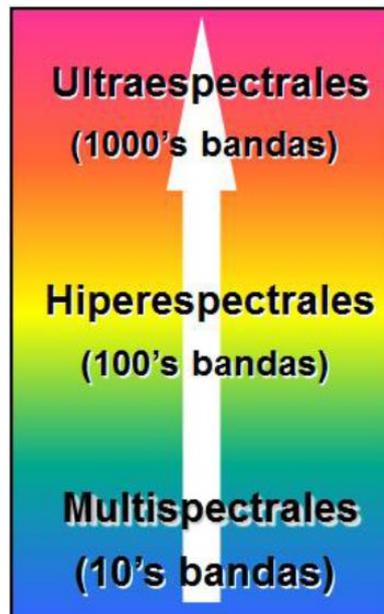


Figura 1.3: Tipos de imágenes según la cantidad de bandas espectrales.

El trabajo realizado en este PFC ofrece un conjunto de herramientas, resultados y conclusiones, obtenidos a nivel de *píxel* (puesto que se desarrolla empleando únicamente datos espectrográficos), extrapolables al estudio a nivel de imagen hiperespectral. La imagen hiperespectral que se obtiene por medio de este tipo de sensor, es realmente un cubo hiperespectral de tres dimensiones, donde cada *píxel* está formado por un vector con los valores de radiación para cada una de las longitudes de onda medidas por dicho sensor (Figura 1.4). Por consiguiente, si una imagen RGB se divide en tres imágenes, una por cada una de las longitudes de onda de los tres colores, de igual modo, una imagen hiperespectral está formada por cientos de imágenes, una por cada una de las longitudes de onda captadas.

Si la diferencia entre cada una de estas longitudes de onda es de unos pocos nanómetros se puede obtener por cada *píxel* una respuesta espectral “casi continua” del material fotografiado para ese *píxel*. Esta respuesta o firma espectral es única para cada material y puede ser utilizada para identificar sustancias específicas en una imagen. Además, cuando varios materiales están presentes en el mismo *píxel*, la respuesta espectral obtenida puede expresarse a nivel macroscópico como la suma de las firmas espectrales de cada sustancia, por lo que mediante técnicas matemáticas se puede calcular incluso la proporción de cada uno de los materiales presentes [3, 6]

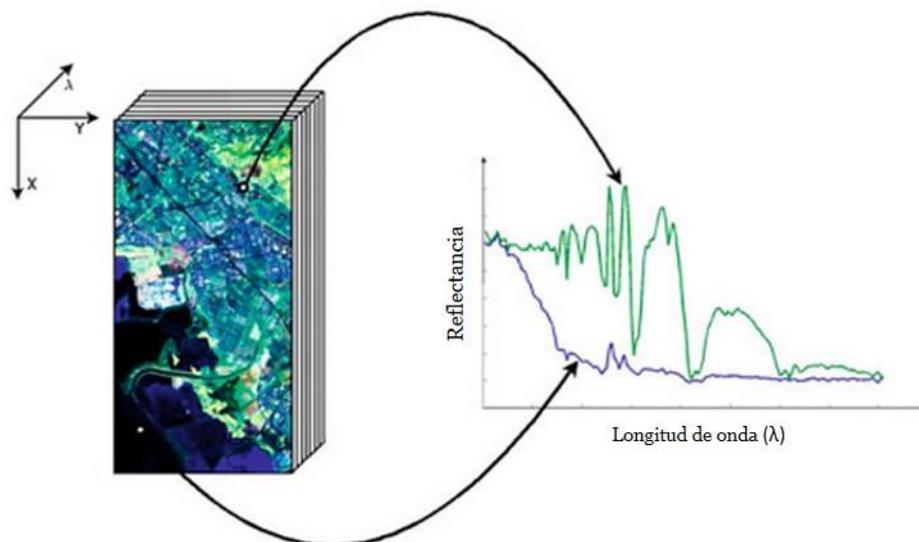


Figura 1.4: Representación esquemática del cubo hiperespectral e ilustración de la firma espectral.

Debido a las características únicas de las firmas espectrales, las imágenes hiperespectrales se pueden emplear para detectar, diferenciar o clasificar sustancias. A pesar del coste computacional tan elevado que supone el procesamiento de este tipo de imágenes, son de especial interés en aplicaciones de teledetección o biomédicas, como la clasificación de células sanas o tumorales en muestras de tejidos [3, 7].

1.1.2. Clasificadores

La línea de investigación en la que se ha trabajado está dedicada a la aplicación de técnicas avanzadas de clasificación en imágenes hiperespectrales provenientes de tumores cerebrales. No obstante, este campo engloba una amplia línea de investigación altamente novedosa y de gran impacto social, como puede verse en las múltiples aplicaciones que se pueden encontrar en el campo de la teledetección aplicada a problemas como la monitorización de recursos naturales, en el estudio de ecosistemas y cambio climático global, en la detección de agentes contaminantes en aguas y atmósfera, en aplicaciones militares, etc. El proceso de clasificación constituye uno de los principales elementos estudiados en este trabajo [6, 7].

Para la identificación y clasificación de sustancias en imágenes hiperespectrales se recurre a técnicas que buscan píxeles representativos de cada uno de los materiales que se pretenden identificar. Sin embargo, el hecho de disponer de un número de bandas elevado

suele implicar el sacrificio de la resolución espacial de la imagen, obteniendo incluso en ocasiones la información espectral de un solo píxel. En estos casos no se pueden emplear las técnicas de búsqueda de píxeles representativos y se utilizan clasificadores que permitan explotar solo la información espectral, como es el caso de las máquinas de vectores soporte (*Support Vectors Machines, SVMs*) o de las redes neuronales artificiales (*Artificial Neural Networks, ANNs*).

Esta situación es la estudiada en el presente proyecto, pues se dispone de una base de datos de imágenes espectrográficas con la mínima resolución espacial, estando formadas las imágenes por un único píxel (dato espectrográfico). Esto se debe a que han sido obtenidas mediante un espectrómetro en vez de una cámara hiperespectral. Cada una de las muestras tiene una resolución espectral muy alta, disponiendo en concreto de dos conjuntos de muestras obtenidas en dos campañas diferentes. De la primera campaña se obtienen respuestas de absorción de 2906 longitudes de onda diferentes, mientras que la segunda campaña proporciona respuestas en *transmitancia* para 1501 longitudes de onda. Estas longitudes de onda pertenecen al espectro visible, al infrarrojo cercano y se adentran incluso en el infrarrojo medio.

Esta alta resolución espectral permite introducir las respuestas de absorción y transmitancia como parámetros característicos de clasificación. La creación de los sistemas basados en el clasificador SVM desarrollado en este trabajo se divide en dos etapas: entrenamiento y validación. En la etapa de entrenamiento se introducen los parámetros de un conjunto de muestras en el que se encuentran muestras pertenecientes a todas las clases a identificar. Con este conjunto el clasificador “aprende” a clasificar cada una de las clases. En nuestro caso, la clasificación de tumores, estas clases o clústeres corresponden a cada uno de los tipos o grados del tumor y a las muestras sanas o no tumorales. El correcto funcionamiento del clasificador se comprueba mediante la etapa de validación, donde se introduce otro conjunto diferente de muestras y se verifica su correcta identificación.

Tras la validación del clasificador, cualquier muestra nueva introducida es asignada a cada una de las clases. La fiabilidad del resultado obtenido depende de la precisión del clasificador desarrollado, por lo que al tratarse de muestras médicas y de diagnósticos críticos es necesario alcanzar precisiones muy altas, cercanas al 100%.

1.2 Antecedentes

El trabajo desarrollado en este proyecto está basado en el análisis y el estudio de las diferentes técnicas de procesamiento automático de clasificación supervisada mediante imágenes hiperespectrales. En concreto, de imágenes hiperespectrales de muestras de tumores cerebrales humanos. Este trabajo forma parte de las líneas de investigación financiadas por la Comisión Europea a través del proyecto europeo *HELICoiD* (*HypErspectraL Imaging Cancer Detection*).

Para este Proyecto Fin de Carrera se cuenta con datos espectrográficos suministrados por el Hospital Universitario de Southampton perteneciente al servicio nacional de salud (*National Health Service, NHS*) de Reino Unido, y a su vez, participante del proyecto HELICoiD. Estas muestras espectrales se adquirieron a partir de varias craneotomías realizadas a pacientes con tumores cerebrales y fueron capturadas mediante dos espectrómetros diferentes, para obtener la respuesta o firma espectral de un único píxel perteneciente a cada muestra.

Como se ha mencionado en el apartado anterior, las muestras se obtuvieron en dos campañas. La resolución espectral de los datos espectrográficos es muy elevada, estando comprendida entre 400 nm y 6000 nm, con una resolución de 1,93 nm (2906 bandas) para la primera campaña; y entre 1000 nm y 4000 nm, con una resolución de 2 nm (1501 bandas) para la segunda. Las muestras se corresponden a tejidos de diferentes tipos y grados de tumores cerebrales, así como también a tejidos sanos o no tumorales.

La primera campaña sirvió como soporte muestral a otro PFC titulado: “*Clasificación de tumores cerebrales mediante espectroscopia utilizando Máquinas de Vectores Soporte (SVM)*” [4]. Sin embargo, cuando se procedió a trabajar con las muestras de la segunda campaña, empleando los sistemas desarrollados en dicho PFC, se observó que los algoritmos que se habían implementado no ofrecían buenos resultados de clasificación. Por esta razón el presente Proyecto Fin de Carrera, “*Identificación de tumores cerebrales usando datos espectrográficos*”, continuó la línea de investigación comenzada en el proyecto citado, basándose en la idea de que no solo fuera posible la correcta clasificación de los datos espectrográficos de la primera campaña de muestras, sino también la correcta clasificación de la segunda campaña, así como la clasificación de una fusión de ambas campañas.

Adicionalmente, aunque este proyecto se centra exclusivamente en el terreno de los datos espectrográficos, también se pretende que los resultados y las conclusiones obtenidas en él sean extrapolables y de utilidad para estudios en el campo de las imágenes hiperespectrales en líneas futuras de investigación.

1.3 Objetivos

Este PFC tiene como principal objetivo desarrollar un sistema clasificador que permita diferenciar de forma automática entre muestras tumorales y no tumorales a partir de datos espectrográficos, así como identificar el tipo y el grado del tumor a partir de la información contenida en la firma espectral del tejido a clasificar. Además, el sistema permitirá evaluar el porcentaje de acierto de diferentes sistemas de clasificación con el fin de seleccionar el modelo más adecuado. Esto se puede desglosar en los siguientes objetivos concretos que se han alcanzado a lo largo del proyecto:

- Estudio sobre los tipos de tumores cerebrales relevantes para el desarrollo de este proyecto y su clasificación en relación al riesgo que entrañan a la salud del paciente.
- Estudio del estado del arte sobre datos espectrográficos e imágenes hiperespectrales en el ámbito médico.
- Estudio del estado del arte de los sistemas de clasificación más adecuados para esta aplicación.
- Estudio de técnicas preprocesado y los diferentes tipos de *kernels* empleados por los clasificadores.
- Desarrollo de algoritmos capaces de identificar correctamente muestras tumorales y no tumorales. Además, los clasificadores implementados deben ser capaces de distinguir las muestras de tumores primarios de las muestras de tumores secundarios, y dentro de los tumores primarios, distinguir los diferentes grados de tumor.
- Validación de los algoritmos de clasificación diseñados empleando las muestras proporcionadas por el Hospital Universitario de Southampton.
- Comparación de las prestaciones de los sistemas de clasificación a tenor de los resultados obtenidos al clasificar cada una de las campañas por separado y la fusión de ambas campañas.
- Proporcionar una alternativa fiable a las técnicas utilizadas actualmente, que ofrezca posibilidades de mejorar la detección de tumores cerebrales.

La investigación y desarrollo de sistemas de clasificación mediante imágenes hiperespectrales 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 clasificadores para imágenes hiperespectrales, en el caso límite de disponer únicamente de un píxel de dicha imagen con una resolución muy alta (dato espectrográfico).

1.4 Peticionario

La realización de este Proyecto Fin de Carrera se lleva a cabo a petición de la *Escuela de Ingeniería de Telecomunicación y Electrónica (EITE)* de la *Universidad de Las Palmas de Gran Canaria (ULPGC)* como requisito indispensable para la obtención del título de *Ingeniero de Telecomunicación* por parte del estudiante encargado en su realización, una vez haya superado esta totalidad de los créditos de los que consta dicha titulación.

1.5 Organización del documento

Este Proyecto Fin de Carrera está estructurado en tres bloques principales: Memoria, Presupuesto y Pliego de Condiciones. Además, se adjunta un CD que complementa la información contenida en el presente documento.

1.5.1. Memoria

La memoria supone la parte descriptiva de este Proyecto Fin de Carrera y está dividida en 7 capítulos y la bibliografía.

Capítulo 1: Introducción. En este capítulo introductorio de la memoria se presenta una breve introducción del tema que se aborda y las razones que motivan la realización del proyecto, se detallan los antecedentes, así como los objetivos del proyecto, su petionario y la estructura del documento.

Capítulo 2: Tumores cerebrales. Se presenta un estudio simplificado de la anatomía del cerebro y de los tipos de tumores cerebrales. Además, se realiza una breve introducción a las principales técnicas que emplean imágenes para el diagnóstico médico.

Capítulo 3: Clasificación basada en información espectral. Se presenta un estudio del estado del arte sobre espectroscopía, imágenes hiperespectrales y se describen los principales algoritmos de clasificación, prestando especial atención en los algoritmos utilizados comúnmente para la clasificación de imágenes hiperespectrales. Además, se presentan las métricas empleadas para la validación de estos sistemas de clasificación.

Capítulo 4: Base de datos de las muestras espectrográficas. En este capítulo se presenta una descripción de las muestras espectrales proporcionadas por el Hospital Universitario de Southampton, así como de los espectrógrafos que fueron empleados para su obtención.

Capítulo 5: Método desarrollado. Se exponen los métodos de clasificación desarrollados a partir de sistemas de clasificación basados en algoritmos de SVM. Adicionalmente, se presentan métodos de optimización de los parámetros del *kernel* RBF, se explica el tratamiento que han recibido de las muestras y el modo en el que se establecen los conjuntos de entrenamiento empleados en la creación de los modelos SVM.

Capítulo 6: Resultados experimentales. En este capítulo se reflejan los resultados que se han ido obteniendo a lo largo del desarrollo de este proyecto. Estos resultados son producto de los múltiples experimentos realizados por medio de los diferentes sistemas de clasificación desarrollados, empleando como muestras fuente de los algoritmos, tanto las campañas 1 y 2 por separado como la fusión de ambas. Adicionalmente se presentan los resultados de las pruebas de normalización de las muestras y la adaptación del espectro de las muestras de la campaña 1 para fusionarlas con la campaña 2.

Capítulo 7: Conclusiones y líneas futuras. En el último capítulo de la memoria se detallan las conclusiones que se han ido obteniendo a lo largo del presente PFC. Además, se exponen las posibles líneas de trabajo que podrían tomar como puntos de referencia las conclusiones y los resultados conseguidos en este proyecto.

Bibliografía. En la bibliografía se detallan las referencias empleadas para la elaboración del proyecto, así como para la redacción de esta memoria.

1.5.2. Presupuesto

En el presupuesto se cuantifican los costes del desarrollo del presente Proyecto Fin de Carrera según el baremo establecido por el *Colegio Oficial de Ingenieros de*

Telecomunicación (COIT) de 2008. A partir de esta fecha, se liberaliza el mercado y no es obligatorio seguir este baremo, que queda a nivel de recomendación y que usamos en este proyecto como referencia.

1.5.3. Pliego de condiciones

En este apartado se detallan las condiciones bajo las cuales se han desarrollado las diferentes etapas del presente Proyecto Fin de Carrera.

Capítulo 2

Tumores cerebrales

2.1 Introducción

El presente Proyecto Fin de Carrera se centra en el estudio y la clasificación de tumores cerebrales mediante la utilización de técnicas de espectroscopia. Por tanto, se hace indispensable reflejar en este capítulo el estudio preliminar realizado para la familiarización de los términos asociados a los tumores cerebrales en el ámbito médico.

Por una parte, se realiza un estudio simplificado de la anatomía y funcionalidad del cerebro humano. Se expone someramente la estructura del cerebro y se muestran diferentes tipos de células que se pueden encontrar en este.

Además, se presentan los tumores primarios distinguiendo los diferentes grados de tumores cerebrales, haciendo énfasis en los tumores gliales por ser el principal objeto de estudio de este proyecto.

Para finalizar se describen los tumores secundarios de interés en este PFC, y se introducen las principales técnicas de diagnóstico mediante imagen.

2.2 Anatomía del cerebro humano

El *cerebro humano* es el órgano mayor del *Sistema Nervioso Central (SNC)* y el centro de control para todo el cuerpo. Está localizado en la cavidad craneal quedando protegido por el cráneo. Las funciones del cerebro incluyen la iniciación y coordinación de los movimientos, la temperatura, el tacto, la vista, el oído, el sentido común, el razonamiento, la resolución de problemas, las emociones y el aprendizaje. Este órgano procesa la información sensorial, y coordina y mantiene las funciones del organismo, lo que requiere una comunicación constante entre los miles de millones de células nerviosas del cerebro (neuronas) a través de impulsos eléctricos [4, 8, 9].

El cerebro en conjunto con el diencefalo, el cerebelo y el tronco encefálico conforman el encéfalo, y el encéfalo junto a la médula espinal constituyen el Sistema Nervioso Central.

2.2.1. Estructura del cerebro

En la literatura anatómica existen dos formas de describir las regiones del cerebro; la primera se basa en la diferenciación de estas regiones durante la formación del cerebro en el desarrollo embrionario; y la segunda, utiliza las regiones del esquema médico del cerebro adulto, las que se desarrollan partiendo de aquellas del cerebro primitivo. En este proyecto utilizaremos la segunda forma, es decir usando el esquema médico del cerebro adulto [10].

El cerebro está formado por dos hemisferios cerebrales que constituyen el 83 % de la masa total cerebral. Estos hemisferios están separados parcialmente por una hendidura o fisura profunda, en cuyo fondo se encuentran unidos en la línea media a través de un tracto de fibras comisurales denominado *cuero calloso*. Cada hemisferio tiene una capa externa de unos 2-4 mm de grosor denominada corteza cerebral, la cual se encuentra extraordinariamente plegada formando unas crestas de tejido, las *circunvoluciones*, separadas por ranuras llamadas *surcos*. Los surcos más profundos, llamados *cisuras*, separan grandes regiones del cerebro. La *cisura interhemisférica longitudinal* separa ambos hemisferios y otro gran surco, la *cisura transversal* separa el cerebro del cerebelo (Figura 2.1) [4, 10, 11].

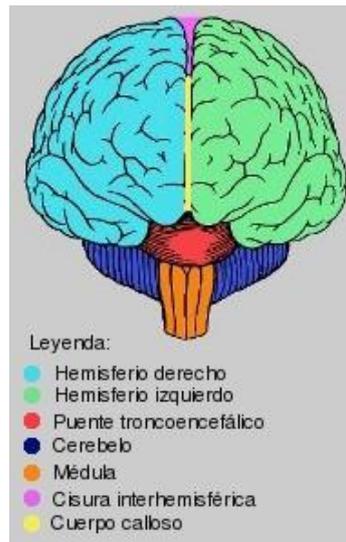


Figura 2.1: Vista esquemática anterior del cerebro.

Cada hemisferio se divide a su vez en cuatro lóbulos: frontal, parietal, occipital y temporal. Sin embargo, esta división no se debe a la estructura real de la corteza cerebral sino a los huesos del cráneo que la protegen (Figura 2.2) [9].

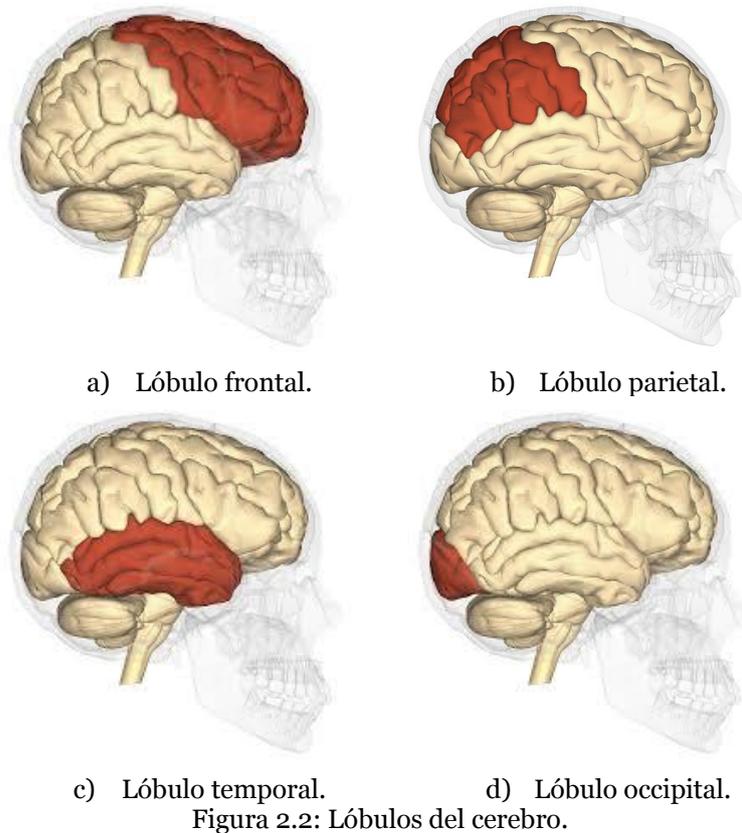


Figura 2.2: Lóbulos del cerebro.

Las diferentes áreas de la corteza cerebral se encuentran involucradas en distintas funciones comportamentales y cognitivas.

El lóbulo frontal (Figura 2.2 a)) controla varias funciones como la resolución de problemas, pensamiento creativo, juicio, intelecto, atención, comportamiento, reacciones físicas, pensamiento abstracto, movimientos coordinados, músculos coordinados y personalidad.

En el lóbulo parietal (Figura 2.2 b)) se pueden encontrar:

- *El córtex motor*: permite que el cerebro controle el movimiento del cuerpo. Se localiza en la parte media superior del cerebro.
- *El córtex sensorial*: se localiza en la parte frontal del lóbulo parietal y recibe información desde la médula espinal sobre la posición de varias partes del cuerpo y cómo se mueven. Esta región también se puede usar para transmitir información del sentido del tacto, incluyendo dolor o presión, la cual afecta a diferentes porciones del cuerpo.

El lóbulo temporal (Figura 2.2 c)) controla la memoria visual, auditiva y comprensión del habla. Este lóbulo incluye áreas que ayudan a controlar capacidades de habla y escucha, comportamiento y lenguaje. Por ejemplo, el área de Wernicke es una porción del lóbulo temporal que se encuentra alrededor del córtex auditivo en la cual se formula y entiende el habla.

El último lóbulo es el occipital (Figura 2.2 d)), se encuentra en la parte posterior de la cabeza y controla la visión, por lo que una lesión en esta zona puede provocar dificultades para leer.

Adicionalmente, existe un quinto lóbulo denominado ínsula que se encuentra ubicado profundamente en la superficie lateral recubierto parcialmente por secciones de los lóbulos frontal, parietal y temporal.

Por otra parte, en el cerebro se encuentran dos tipos de sustancias o materias comúnmente conocidas como sustancia gris y sustancia blanca. La sustancia gris está compuesta por los somas y cuerpos neuronales, que no poseen *mielina* (son los causantes del color blanquecino) y *oligodendrocitos* (tipo de células cerebrales que se verán en el

siguiente apartado). La sustancia blanca generalmente no contiene cuerpos celulares y se encarga de conducir el impulso nervioso dentro del Sistema Nervioso Central.

En el encéfalo la sustancia blanca está distribuida en el interior, mientras que la corteza encefálica se compone de materia gris [12].

2.2.2. Tipos de células del cerebro

Desde el punto de vista neurológico en el cerebro existen dos tipos de células: las *neuronas* y las células *gliales*.

Las neuronas son la unidad básica del sistema nervioso. Estas se componen de un cuerpo celular, *soma*, con múltiples prolongaciones, *dendritas* y una prolongación extensa, *axón* (Figura 2.3). Cada neurona recibe información a través de sus dendritas, la procesa en el soma y la envía a lo largo del axón hasta la sinapsis con otra neurona o célula receptora. La sinapsis es una unión intercelular especializada que permite a las células nerviosas comunicarse unas con otras [4].

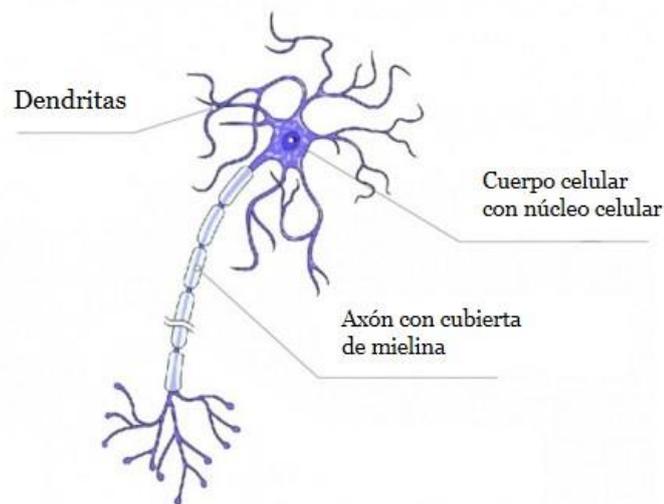


Figura 2.3: Estructura de una neurona

Las neuronas pueden procesar y transmitir información mediante señales bioeléctricas o químicas. Las señales bioeléctricas son producidas por el movimiento de iones a través de la membrana celular mientras que las químicas son moléculas que actúan como transmisores nerviosos.

Por otro lado, las células de la glía o *neuroglía* sirven como elemento de soporte, tanto físico como metabólico, para las neuronas. También tienen una gran importancia en el desarrollo y en los procesos de regeneración del sistema nervioso y permiten la comunicación e integración de las redes neuronales. En general, son células pequeñas de prolongaciones cortas y ramificadas y a pesar de ser consideradas principalmente células de sostén del tejido nervioso, existe una dependencia funcional muy importante entre neuronas y células gliales. Cada neurona presenta un recubrimiento glial complementario a sus interacciones con otras neuronas, de manera que sólo se rompe el entramado glial para dar paso a la sinapsis. Además, este tipo de células son hasta 10 veces más abundantes que las neuronas y el origen más común de tumores cerebrales (*gliomas*).

Generalmente, las células gliales se clasifican en *microglía* y *macroglía*. Mientras que la microglía se encuentra únicamente en el Sistema Nervioso Central, la macroglía está presente además en el *Sistema Nervioso Periférico (SNP)*. Las células de la macroglía del SNC pueden subclasificarse aún más en astrocitos, oligodendrocitos y células ependimarias o ependimocitos.

A continuación, se presentan las principales características de las células microgliales y de las células de la macroglía presentes en el SNC, al ser las células causantes de los tumores cerebrales estudiados en este proyecto.

2.2.2.1. Células microgliales

Las células microgliales forman el sistema inmunitario del Sistema Nervioso Central. Las neuronas mantienen a este tipo de células en estado de reposo y solo se “activan” tras un estímulo, por ejemplo, pequeñas microrroturas de vasos sanguíneos, daños de tejido, golpes o agentes infecciosos. En las zonas de lesión, las células microgliales se dividen, aumentan de tamaño y adquieren facultades fagocitarias para eliminar las células dañadas o tejido alterado.

2.2.2.2. Astrocitos

También conocidos como *astroglía*, son las células más numerosas en el SNC. Los astrocitos tienen funciones básicas para el mantenimiento de la función neuronal entrelazándose alrededor de las neuronas para formar el sostén de estas, formando así una barrera filtradora entre las neuronas y la sangre. Tienen formas estrelladas, presentan

largas prolongaciones y cuando existe una destrucción neuronal liberan el factor de crecimiento para ayudar a reconstruir las neuronas dañadas.

Las principales funciones de los astrocitos son las siguientes:

- Dan soporte físico a las neuronas.
- Regulan la composición del fluido extracelular.
- Tienen una función nutricional. Algunas de sus prolongaciones contactan con vasos sanguíneos y otras con neuronas. De este modo pueden captar los nutrientes de la sangre y transmitirlos a las neuronas (principalmente glucosa).
- Protegen al SNC de la entrada de sustancias extrañas al formar parte de la *barrera hematoencefálica*.
- Intervienen en las reacciones inmunológicas cerebrales, concretamente cuando se produce algún daño cerebral. En esa zona se eliminan los restos de las células y proliferan astrocitos para reparar la zona dañada.
- Tienen un papel importante durante el desarrollo ontogénico¹. Sirven para guiar a las neuronas a la posición cerebral en la que tienen que quedar situadas.

Se han identificado dos tipos de astrocitos:

- **Astrocito tipo I o protoplasmático.** Este tipo de astrocitos se encuentran principalmente en la sustancia gris. Presentan abundancia de citoplasma, un núcleo grande y muchas prolongaciones muy ramificadas que suelen extenderse hasta las paredes de los vasos sanguíneos. De esta manera, los astrocitos tipo I participan en la regulación de las uniones estrechas de las células endoteliales de los capilares (células que forman las paredes de los vasos sanguíneos) y venas que conforman la barrera hematoencefálica. Esta barrera limita el paso de sustancias hasta las neuronas, por lo que estos astrocitos controlan de esta forma el medio ambiente de la neurona. Los más superficiales emiten prolongaciones hasta contactar con la piamadre encefálica y medular, conformando la membrana pial-glial.

¹ El desarrollo ontogénico describe el desarrollo de un individuo, referido en especial al período embrionario [111].

- **Astrocito tipo II o fibroso.** Estos astrocitos se encuentran sobre todo en la sustancia blanca. Tienen prolongaciones más largas y menos ramificadas que los astrocitos tipo I. Además, suelen encapsular las sinapsis químicas.

2.2.2.3. Oligodendrocitos

Los oligodendrocitos u oligodendroglía son células más pequeñas que presentan menor cantidad de prolongaciones, y menos ramificadas, que los astrocitos. Se encargan de formar la vaina de mielina que envuelve los axones de las neuronas, pero también realizan la función de sostén y unión de células. La mielina es una sustancia *lipoproteica* (compuesta por agua, lípidos y proteínas) que actúa como aislante para favorecer a conducción del impulso nervioso.

Estas células se localizan tanto en la sustancia gris como en la blanca del SNC. Su cuerpo celular es pequeño y el citoplasma es muy denso. Asimismo, tienen un núcleo esférico y más pequeño que el de los astrocitos.

Se han identificado dos tipos de oligodendrocitos: los oligodendrocitos *interfasciculares* que se encargan de la producción de la vaina de mielina y aislamiento del axón y los oligodendrocitos *satelitales*, de los cuales aún no se ha precisado su función.

2.2.2.4. Células endimarias o endimocitos

Las células endimarias son células de revestimiento que recubren *los ventrículos cerebrales* (cavidades interconectadas entre sí por las que circula el líquido cefalorraquídeo) y el canal de entrada de la médula espinal que contiene al líquido cefalorraquídeo. De los dos tipos que existen: *endimocitos* y *tanicitos*, los tanicitos pueden considerarse una variedad especializada que envía prolongaciones hacia neuronas y vasos sanguíneos del hipotálamo.

En la Figura 2.4 se muestran diferentes células gliales del SNC.

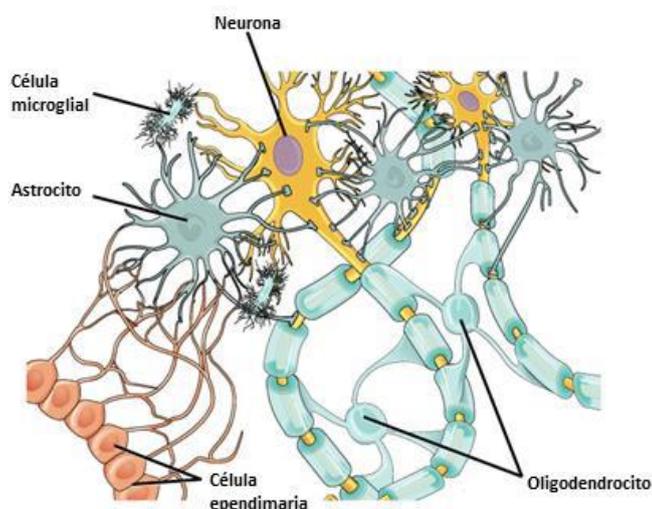


Figura 2.4: Células gliales del Sistema Nervioso Central [13].

2.3 Tumores cerebrales

El cáncer o tumor cerebral se trata de una de las enfermedades más incidentes en el sistema nervioso, provocada por un grupo de células que se multiplican sin control y de manera autónoma, invadiendo localmente y a distancia otros tejidos.

El cáncer es provocado por anomalías en el material genético de las células que provocan un crecimiento no controlado en el número de células, e incluso diseminándose a otras partes del cuerpo (*metástasis*). Estas células cancerosas también son capaces de inducir el crecimiento de tejido conectivo y vasos sanguíneos alrededor del mismo para favorecer su crecimiento. Los tumores se pueden clasificar en función de la localización del organismo en la que se originan. De esta manera podemos diferenciar a los tumores cerebrales *primarios*, aquellos cuyas células afectadas se derivan de componentes cerebrales; de los tumores cerebrales *secundarios* o metástasis cerebrales, aquellos que se desarrollan en el cerebro como consecuencia de la diseminación de un tumor maligno localizado en otra parte del organismo [14]. Estos últimos se pueden diseminar hacia el cerebro a través del sistema circulatorio, a través de la vía linfática y del líquido cefalorraquídeo.

Los tumores primarios se pueden diferenciar según su malignidad en tumores benignos y malignos. Los tumores benignos son de crecimiento lento, no se propagan a otros tejidos y rara vez reinciden tras ser extirpados, mientras que los tumores malignos son de

crecimiento rápido, se propagan a otros tejidos, reinciden con frecuencia tras ser extirpados y provocan la muerte en un periodo variable de tiempo si no se realiza tratamiento [15, 16].

Los tumores del sistema nervioso central derivan de diversos tejidos, como los *neuroepiteliales*, meníngeos y otros como los linfomas, quistes y tumores de la glándula pituitaria. No obstante, los tumores neuroepiteliales gliales o gliomas, como su nombre indica, afectan a las células gliales y son los tumores cerebrales más frecuentes y agresivos, causando globalmente un 2% en muertes por cáncer [4].

Los síntomas y el tratamiento de este tipo de tumores dependen en gran parte de la edad del paciente, del tipo de tumor y de su localización en el cerebro. Sin embargo, estos tumores tienden a infiltrarse en el tejido cerebral sano, por lo que su extirpación quirúrgica es compleja y en ocasiones imposible. Por tanto, es de vital importancia la investigación y el estudio de técnicas no invasivas que faciliten una diagnosis lo más exacta posible.

Una correcta resección del tumor cerebral elimina el tejido maligno y evita que el tumor se vuelva a producir. Aumentar la zona de resección con un margen adicional de seguridad maximiza las probabilidades de supervivencia del paciente, pero en el caso de tumores cerebrales, una resección excesiva puede generar daños importantes de todo tipo en el paciente (motrices, cognitivos, visuales, etc.). Es por ello que se hace necesaria una técnica que determine de forma precisa y mínimamente invasiva los límites del tumor.

Antes de comenzar con la clasificación de este tipo de tumores se presenta la graduación histológica utilizada. Entre los diferentes tipos de clasificación existentes, en este proyecto se ha seguido la graduación establecida por la Organización Mundial de la Salud en la edición del año 2007 de *Histological Typing of Tumours of the Central Nervous System* por ser la más empleada [17].

- **Grado I:** tumores de crecimiento lento, no malignos y asociados a un largo periodo de supervivencia.
- **Grado II:** tumores de crecimiento relativamente lento pero que pueden evolucionar en tumores de alto grado. Pueden ser malignos o benignos.
- **Grado III:** tumores malignos que con frecuencia evolucionan en tumores de alto grado.
- **Grado IV:** tumores malignos muy agresivos que se reproducen rápidamente.

Mientras que los tumores cerebrales primarios reciben el nombre según el tipo de célula en que se originan [14], los secundarios se denominan a razón del origen de la metástasis. Y en este caso, las metástasis más comunes son: pulmonares, mamarias, por melanoma, colorrectales o renales. En los siguientes subapartados se listan los principales tipos de gliomas, así como tumores secundarios, objetos de estudio en este proyecto.

2.3.1. Astrocitomas

Los astrocitomas son el grupo de tumores intracraneales primarios cuya célula predominante deriva de los astrocitos que se han vuelto inmortales y crecen en número de forma descontrolada. Este tipo de tumores constituyen aproximadamente el 80% de los tumores neuroepiteliales.

Astrocitomas de grado I:

- **Astrocitoma pilocítico.** Se presenta típicamente en niños y personas jóvenes. Es habitualmente un tumor bien delimitado, de aspecto sólido o quístico.
- **Astrocitoma subependimario de células gigantes.** Tumor generalmente circunscrito originado en la pared de los ventrículos laterales.

Astrocitomas de grado II:

- **Astrocitoma difuso, astrocitoma o astrocitoma de grado II.** Es el glioma de bajo grado más frecuente, afecta a niños y adultos entre 20 y 40 años.
- **Astrocitoma pilomixoide.** Variante de astrocitoma pilocítico que afecta a niños con edad media de 10 meses. La recaída y la diseminación es más frecuente que en el astrocitoma pilocítico.
- **Xantoastrocitoma pleomórfico.** Tumor poco frecuente que aparece en niños y jóvenes, bien delimitado, parcialmente quístico y localizado frecuentemente en el lóbulo temporal.

Astrocitomas de grado III:

- **Astrocitoma anaplásico.** Tumor difuso, infiltrante que se caracteriza por la poca diferenciación de las células que lo componen y por su potencial de crecimiento. Puede proceder de tumores de bajo grado en vía de transformación a glioblastomas, por lo que tienen gran tendencia a evolucionar en astrocitomas de grado IV.

Astrocitomas de grado IV:

- **Glioblastoma, GBM o astrocitoma de grado IV.** Es el tumor primario cerebral más frecuente y agresivo. Es más frecuente en adultos, siendo la edad media al diagnóstico de 62 años. La mayor parte de ellos surgen sin evidencia de una lesión previa pero algunos se desarrollan a partir de astrocitomas de menor grado. Es un tumor de rápido crecimiento compuesto por una mezcla heterogénea de células astrocitarias, con una gran proliferación vascular, necrosis y frecuente mitosis².
- **Glioblastoma de células gigantes.** Variedad histológica del glioblastoma que presenta una prevalencia de células gigantes y multinucleadas.
- **Gliosarcoma.** Se caracteriza por presentar un componente glial y otro sarcomatoso. Los sarcomas son cánceres que se originan en ciertos tejidos, tales como los huesos o los músculos.

2.3.2. Oligodendrogliomas

Grupo de tumores intracraneales del Sistema Nervioso Central de crecimiento lento y que rara vez producen metástasis a otros tejidos. La célula predominante en estos tumores deriva de los oligodendrocitos que, al igual que los astrocitos en los astrocitomas, se han vuelto inmortales y se reproducen de forma descontrolada.

Oligodendrogliomas de grado II:

- **Oligodendroglioma.** Este tipo de tumor normalmente aparece en adultos, con un pico de incidencia entre los 35 y los 45 años. Es un tumor infiltrante, difuso, bien diferenciado, de localización habitual en la sustancia blanca de los hemisferios cerebrales y predominantemente en el lóbulo frontal.

Oligodendrogliomas de grado III:

- **Oligodendroglioma anaplásico.** Tumor poco frecuente localizado generalmente en los lóbulos frontal y temporal. Con frecuencia presenta calcificaciones, necrosis, degeneración quística y hemorragias intratumorales.

² Proceso de reproducción de las células.

2.3.3. Oligoastrocitomas o gliomas mixtos

Presentan características histopatológicas³ y moleculares propias tanto de los astrocitomas como de los oligodendrogliomas.

Oligoastrocitomas de grado II:

- **Oligoastrocitoma.** Presenta un crecimiento de células moderado con escasas o nulas mitosis. En este tumor se distinguen células astrocitarias y células oligodendrogliales, que pueden estar en áreas claramente separadas o entremezcladas, que es lo más frecuente. Los oligoastrocitomas de bajo grado o grado II son indistinguibles de los oligodendrogliomas desde el punto de vista de la *neuroimagen*.

Oligoastrocitomas de grado III:

- **Oligoastrocitoma anaplásico.** Se caracteriza morfológicamente por un patrón mixto de células de aspecto oligodendrogliar entre células de aspecto astrocitario, con signos de malignidad en alguno de sus dos componentes.

2.3.4. Otros tumores gliales

- **Ependinomas.** Los tumores ependimarios habitualmente empiezan en las células que revisten los espacios llenos de líquido del cerebro y que rodean la médula espinal.
- **Astroblastoma.** Es un tumor glial poco frecuente y de origen desconocido que aparece principalmente en niños, adolescentes y adultos jóvenes. Los astroblastomas suelen aparecer típicamente en los hemisferios cerebrales, como lesiones bien definidas, a menudo de aspecto quístico y con captación de contraste. La organización mundial de la salud no ha establecido todavía su grado histológico de modo definitivo.
- **Gliomatosis cerebri.** Tumor glial difuso poco frecuente, caracterizado por una gran capacidad infiltrativa. Las células afectadas pueden ser astrocitarias, oligodendrogliales o mixtas.

³ La Histopatología es la rama de la Patología que trata el diagnóstico de enfermedades a través del estudio de los tejidos [112].

2.3.5. Tumores secundarios

Los tumores secundarios con los que se contaron a la hora de desarrollar este proyecto tienen como tumores origen, o tumores primarios, el adenocarcinoma pulmonar y el carcinoma gástrico. Los carcinomas son el tipo de cáncer más frecuente y se trata de tumores que se inician en la piel o en las células superficiales de los órganos internos (pulmón, mama, colón, útero, etc.), células de tipo epitelial o glandular. Si el carcinoma afecta a zonas de la superficie de los órganos que tienen una función secretora, entonces reciben el nombre de adenocarcinomas [18].

- **Adenocarcinoma pulmonar.** Se trata de un tipo de carcinoma pulmonar no *microcítico* (células no pequeñas) que suele encontrarse en partes externas del pulmón por lo que frecuentemente afecta a la pleura y pared torácica [19, 20]. Es el principal cáncer de pulmón en los no fumadores, aunque es más frecuente en fumadores.
- **Carcinoma gástrico.** Se produce con la formación de células cancerosas en los tejidos del estómago. Puede propagarse al esófago o al intestino delgado, y finalmente infiltrarse en los ganglios linfáticos propagándolo a cualquier parte del organismo [21]. La infección con la bacteria *helicobacter pylori* es el factor de riesgo primario asociado con el cáncer gástrico [22].

2.4 Imágenes diagnósticas

Las pruebas y procedimientos de diagnóstico son herramientas vitales que ayudan a los médicos a confirmar o descartar la presencia de un trastorno neurológico u otra enfermedad. Hoy en día, tras décadas de investigación, es posible observar el cerebro humano sin necesidad de realizar una autopsia después de la muerte del paciente.

Uno de los grandes avances en el diagnóstico médico ha sido el uso de imágenes diagnósticas debido a que son pruebas poco o nada invasivas. Los cambios más significativos en las imágenes diagnósticas en los últimos 20 años han sido los adelantos en la resolución espacial (tamaño, intensidad y claridad) de las imágenes anatómicas y las reducciones en el tiempo necesario para enviar señales y recibir datos del área de donde se obtiene la imagen. Estos avances permiten a los médicos ver simultáneamente la estructura cerebral y los cambios en la actividad cerebral a medida que ocurren. Los científicos

continúan mejorando los métodos que proporcionarán imágenes anatómicas más nítidas e información funcional más detallada [4].

Las principales pruebas diagnósticas con imágenes para tumores cerebrales son las siguientes: *Tomografía Axial Computarizada (Computerized Axial Tomography, CAT)*, *Resonancia Magnética (Magnetic Resonance Imaging, MRI)*, *Tomografía por Emisión de Positrones (Positron Emission Tomography, PET)*, *Tomografía por Emisión de Fotón Único (Single-Photon Emission Computed Tomography, SPECT)* e *Imágenes Ultrasónicas (Ultrasonic Imaging, UI)*. No obstante, este proyecto está orientado a la utilización de una nueva técnica de diagnóstico basada en el procesamiento de imágenes hiperespectrales de tejidos cerebrales. Como se comentó en el Capítulo 1, este proyecto se centra concretamente en el estudio y análisis de datos espectrográficos de tumores cerebrales.

2.5 Resumen

La tasa de mortalidad entre los pacientes que padecen tumores cerebrales está directamente relacionada con el tipo de tumor que estos presentan. Entre las clasificaciones de tumores que se pueden realizar, destacan las que contemplan el origen del tumor (tumor primario o metástasis) y las que definen el grado de peligrosidad, que va desde tumores benignos de crecimiento lento a los cánceres malignos con un crecimiento muy rápido. Un diagnóstico temprano de este tipo de enfermedades es vital para evitar que un tumor benigno evolucione a un cáncer mucho más agresivo. Por tanto, es necesario investigar nuevas técnicas que faciliten su diagnóstico.

Capítulo 3

Clasificación basada en información espectral

3.1 Introducción

En este capítulo se refleja el estudio realizado en este Proyecto Fin de Carrera, en relación al estado del arte de los diversos métodos de clasificación de datos. Se hace especial énfasis en algunos tipos de datos que contienen la información relevante en un rango de su espectro de frecuencia.

Antes de presentar los métodos de clasificación, se introducen detalladamente los conceptos de espectroscopía e imagen hiperespectral. Es necesario familiarizarse con este tipo de datos espectrales, principalmente porque son la base de estudio en este trabajo, y además, porque se requiere su conocimiento para poder entender las técnicas que se emplean en la clasificación de estos.

Para finalizar, se estudia la imagen hiperespectral aplicada a la detección de tumores, técnicas de análisis y algunas técnicas de preprocesado. Además, se exponen las diferencias entre los sistemas de clasificación supervisados, no supervisados y semi-supervisados; se explica la base teórica de algunos de estos algoritmos de clasificación, y en el caso de los

supervisados, se enuncian algunas de las limitaciones que presentan y ejemplos de las aplicaciones en las que han sido empleados.

3.2 Espectroscopía

La espectroscopia es un término general para la ciencia que trata de las distintas interacciones de la radiación con la materia. Las interacciones de interés estudiadas inicialmente se producían entre la radiación electromagnética y la materia. Más tarde el concepto se amplió para comprender cualquier medida en función de la longitud de onda (λ) o de la frecuencia. Por tanto, la espectroscopia puede referirse también a una respuesta a un campo alternante o frecuencia variante (ν).

En un principio se refería al uso de la luz visible dispersada según su longitud de onda, por ejemplo por un prisma (Figura 3.1). Una extensión adicional del alcance de la definición añadió la energía (E) como variable, al establecerse la relación $E = h\nu$ para los fotones, donde (h) es la constante de Planck. Más aún, tomando en cuenta la dualidad partícula-onda, en la actualidad se incluyen también interacciones entre la materia y otras formas de energía como son las ondas acústicas y los haces de partículas como iones o electrones. Por otro lado, el término *espectrometría* hace referencia a la medición, por tanto a la obtención de resultados medibles de las interacciones que estudia teóricamente la espectroscopía. Cuando hablamos de *espectrofotometría* se hace referencia a la medida de la intensidad de la interacción entre la radiación y las sustancias mediante un detector fotoeléctrico o con otro tipo de dispositivo electrónico. El gráfico de la respuesta como función de la longitud de onda o la frecuencia se conoce como espectro [23, 24].

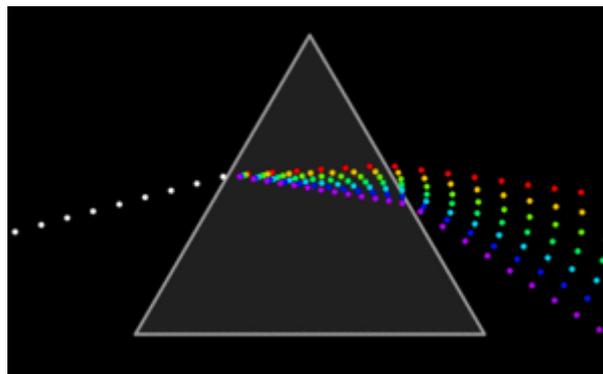


Figura 3.1: Dispersión de luz en un prisma triangular.

La espectrometría es por tanto la técnica espectroscópica que nos permite tasar la concentración o la cantidad de especies químicas determinadas. En estos casos, el

instrumento que realiza tales medidas es un espectrómetro o *espectrógrafo*. Este dispositivo a menudo se usa en física y química analítica para la identificación de sustancias mediante el espectro emitido o absorbido por las mismas.

Por otra parte, la espectrometría se usa mucho en astronomía y detección remota. La mayoría de los telescopios grandes tienen espectrómetros, que son usados para medir la composición química y propiedades físicas de los objetos astronómicos, o para medir sus velocidades a partir del efecto *Doppler* de sus líneas espectrales.

El fundamento de la espectroscopía se debe a la capacidad de las moléculas para absorber radiaciones. Las longitudes de onda de las radiaciones que una molécula puede absorber y la eficiencia con que se absorben dependen de la estructura atómica y de las condiciones del medio (pH, temperatura, fuerza iónica, constante dieléctrica), por lo que dicha técnica constituye un valioso instrumento para determinación y caracterización de biomoléculas [25].

Las biomoléculas pueden absorber energía luminosa y almacenarla en forma de energía interna. Cuando la luz (considerada como energía) es absorbida por una molécula se origina un salto desde un estado energético basal o fundamental, E_1 , a un estado de mayor energía (estado excitado), E_2 , y solo se absorberá la energía que permita el salto al estado excitado (Figura 3.2). Cada molécula tiene una serie de estados excitados (o bandas) que la distingue del resto de moléculas. Como consecuencia, la absorción que a distintas longitudes de onda presenta una molécula, esto es, su espectro de absorción, constituye una señal de identidad de la misma. Por último la molécula en forma excitada libera la energía absorbida hasta el estado energético fundamental.

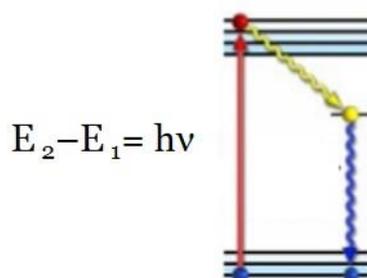


Figura 3.2: Diagrama de niveles de energía en una molécula.

En espectroscopía el término *luz* no solo se aplica a la forma visible de radiación electromagnética, sino también a las formas de radiación que son invisibles.

Cuando un rayo de luz de una determinada longitud de onda de intensidad (I_o) incide perpendicularmente sobre un compuesto químico que absorbe luz o *cromóforo*⁴, el compuesto absorberá una parte de la radiación incidente (I_a) y dejará pasar el resto (I_t), de forma que se cumple (Figura 3.3):

$$I_o = I_a + I_t \quad (3.1)$$

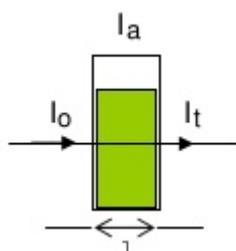


Figura 3.3: Radiación incidente, absorbida y transmitida por un compuesto químico.

La transmitancia (T) de una sustancia es la relación entre la cantidad de luz transmitida que llega al detector una vez que ha atravesado la muestra, (I_t), y la cantidad de luz que incidió sobre ella, (I_o), y se representa normalmente en tanto por ciento:

$$T(\%) = \frac{I_t}{I_o} 100 \quad (3.2)$$

La transmitancia nos da una medida física de la relación de intensidad incidente y transmitida al pasar por la muestra. La relación entre $T(\%)$ y la concentración no es lineal, pero asume una relación logarítmica inversa.

La absorbancia (A) es un concepto más relacionado con la muestra puesto que nos indica la cantidad de luz absorbida por la misma, y se define como:

$$A = \log\left(\frac{1}{T}\right) = -\log\left(\frac{I_t}{I_o}\right) \quad (3.3)$$

Cuando la intensidad incidente y transmitida son iguales ($I_o = I_t$), la transmitancia es del 100% e indica que la muestra no absorbe a una determinada longitud de onda, y entonces (A) vale $\log(1) = 0$.

⁴ Sustancia de los tejidos que absorben la radiación. En química es cualquier molécula o parte de una molécula responsable por el color del material. Cuando la luz alcanza un cromóforo, la excitación de un electrón hace que sean emitidos fotones de un color específico [113].

La cantidad de luz absorbida dependerá de la distancia que atraviesa la luz a través de la solución del cromóforo y de la concentración de este.

El espectro de absorción es una representación gráfica que indica la cantidad de luz absorbida a diferentes valores de λ . A partir de una solución diluida de un compuesto, cuya absorbancia máxima entra dentro del rango de medida del espectrógrafo, se verá el valor de absorbancia a diferentes longitudes de onda frente a un “blanco” que contenga el disolvente de la solución de la muestra a caracterizar. Esto es, se mide primero la absorbancia del disolvente (conocido como blanco) y al que se le asigna el valor de cero de manera que la intensidad incidente y transmitida sean iguales, y por tanto la absorbancia es cero. A continuación, se analiza la muestra impregnada en ese disolvente con lo que se lee la absorbancia de esta.

El espectro de absorción de un cromóforo depende, fundamentalmente, de la estructura química de la molécula. A modo de ejemplo, en la Figura 3.4 se presentan los espectros de absorción de tres pigmentos fotosintéticos cloroplastídicos obtenidos por medio de un *espectrofotómetro*⁵.

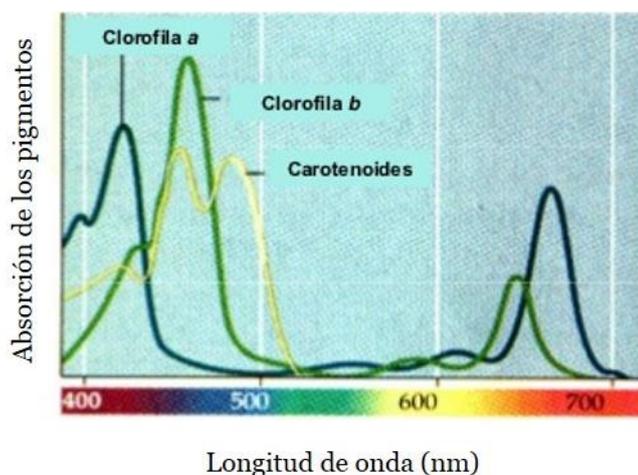


Figura 3.4: Espectros de absorción de tres pigmentos fotosintéticos cloroplastídicos.

A lo largo de esta memoria se hace referencia a las respuestas espectrales de las muestras, tanto en absorbancia así como en transmitancia, indistintamente con los términos firmas espectrales y datos espectrográficos.

⁵ Instrumento empleado en espectrofotometría. Si la región de interés está restringida a un rango cercano al espectro visible, el estudio se denomina *espectrofotometría* [113].

3.3 Imágenes hiperespectrales

3.3.1. Concepto de imagen hiperespectral

La observación de un determinado objeto se basa en la captación, por parte de un instrumento de medida o sensor, de la radiación electromagnética proveniente de la interacción entre el objeto y la fuente de radiación. La radiación electromagnética recibe varias denominaciones dependiendo de la longitud de onda que la caracteriza, como puede apreciarse en la Figura 3.5. Para medir la radiación emitida o reflejada por una determinada superficie, es preciso cuantificar la cantidad de flujo energético que procede de la misma. Para ello, se utiliza la medida de la radiancia, que depende de factores como la percepción del brillo, reflectancia, ángulos de observación, entre otros [26].

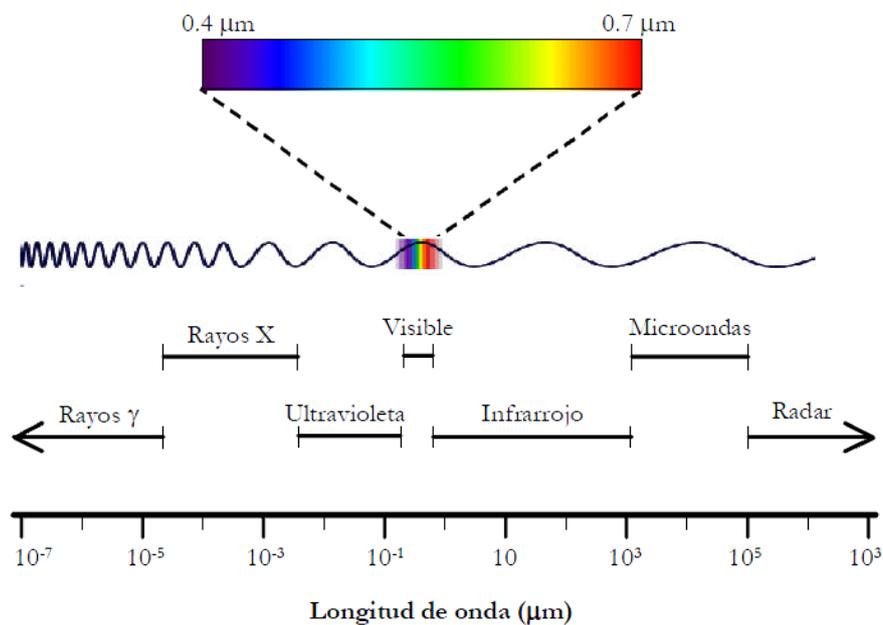


Figura 3.5: Espectro electromagnético.

Las técnicas de detección espectral se basan en el hecho de que todos los materiales presentes en el mundo real reflejan, absorben y emiten energía electromagnética de forma distinta en diferentes longitudes de onda [27].

En la actualidad existe un amplio conjunto de instrumentos o sensores capaces de medir singularidades espectrales en diferentes longitudes de onda [28]. La disponibilidad de estos instrumentos ha facilitado una redefinición del concepto de imagen digital a través de la extensión de la idea de píxel.

El valor asociado a cada píxel viene definido por un valor numérico denominado *nivel digital (ND)*. El nombre se justifica por tratarse de un valor numérico, no visual, pero que puede fácilmente traducirse a una intensidad visual o nivel de gris mediante cualquier convertidor digital-analógico. Así, en un esquema puramente espacial, un píxel está constituido por un único valor discreto, mientras que, en un esquema espectral, un píxel consta de un conjunto de valores. Estos valores pueden ser entendidos como vectores N-dimensionales [29], siendo N el número de bandas espectrales en las que el sensor mide información.

La ampliación del concepto de píxel da lugar a lo que se conoce como *imagen multidimensional*, como aparece en el ejemplo ilustrativo de la Figura 3.6. En dicha figura, el orden de magnitud N permite realizar una distinción a la hora de hablar de imágenes multidimensionales. De esta manera, cuando el valor de N es reducido, típicamente unas cuantas bandas espectrales [30], se habla de *imágenes multipespectrales*; mientras que, cuando el orden de magnitud de N es de cientos de bandas [31], se habla de *imágenes hiperespectrales* o *cubos hiperespectrales*.

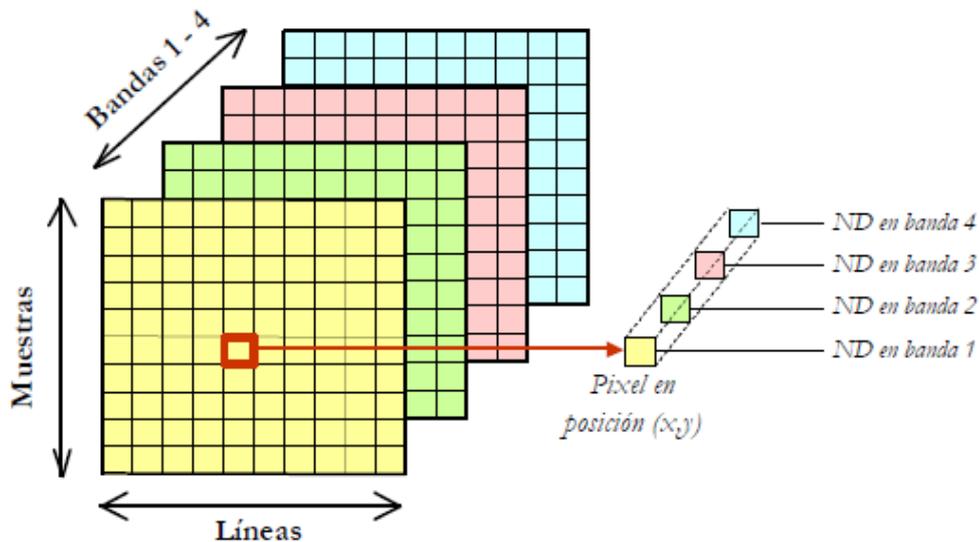


Figura 3.6: Imagen multidimensional de 4 bandas.

Mientras que generalmente el ojo humano percibe la luz visible en tres bandas (rojo, verde y azul), las imágenes hiperespectrales dividen el espectro en muchas más bandas. Esta técnica de dividir las imágenes en bandas puede ser extendida más allá del espectro visible tal como se ilustra en la Figura 3.7. Como puede observarse, el cubo hiperespectral representa la respuesta en reflectancia para un número elevado de longitudes de onda

mientras que en una imagen RGB, solo se tiene la respuesta en intensidad para tres longitudes de onda (correspondientes al rojo, azul y verde).

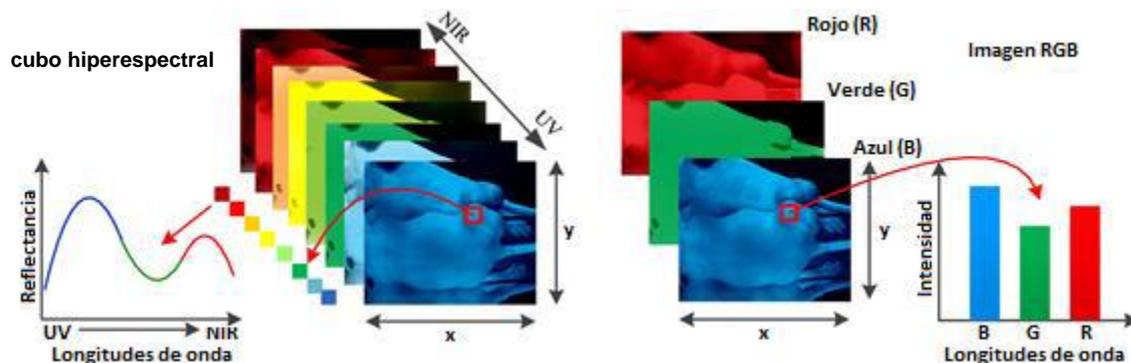


Figura 3.7: Comparación entre un cubo hiperespectral y una imagen RGB [32].

Como se ha comentado en la introducción del Capítulo 1 de esta memoria, la resolución de una imagen hiperespectral está definida por dos tipos distintos de resoluciones: la resolución espacial y la resolución espectral. La resolución espacial viene dada por el número de píxeles de la imagen, y en el caso límite de tener un único píxel en lugar de una imagen, se dispone del dato espectroscópico en lugar de la imagen hiperespectral.

Por otro lado, la resolución espectral hace referencia al número de longitudes de onda para las que se tiene valores y a la distancia entre estas longitudes. La imagen hiperespectral tiene mayor resolución espectral y cubre un rango más amplio de longitudes de onda. En caso de disponer de un solo píxel, este queda definido en la imagen RGB por tres valores de intensidad, mientras que en el cubo hiperespectral queda representado por su firma espectral. Si la distancia entre estas longitudes de onda es de unos pocos nanómetros, se puede obtener para cada píxel una respuesta espectral “casi continua” del material fotografiado en ese píxel. Esta respuesta o firma espectral es única para cada material y puede ser utilizada para identificar sustancias específicas en una imagen. Por ejemplo, estas firmas espectrales pueden ser utilizadas para que un cirujano pueda visualizar áreas de tumor en el tejido cerebral haciendo uso de una cámara hiperespectral.

Debido a la ventaja de disponer de imágenes bidimensionales a lo largo de un amplio rango del espectro electromagnético, las imágenes hiperespectrales han sido aplicadas a numerosas áreas. Ingenieros han construido sensores hiperespectrales y sistemas para aplicaciones en astronomía, agricultura, imágenes biomédicas, geociencia, física y vigilancia [32].

Las principales desventajas de estas técnicas se deben al coste y a la complejidad, para analizar los datos hiperespectrales. Para ello, se requiere de computadores rápidos, detectores sensibles y capacidades altas de almacenamiento. La capacidad de almacenamiento es fundamental debido a que los cubos hiperespectrales son grandes conjuntos de datos que potencialmente exceden de cientos de *megabytes*. Todos estos factores incrementan significativamente el costo de la adquisición y procesado de datos de imágenes hiperespectrales.

3.3.2. La imagen hiperespectral aplicada a la detección de tumores

Aunque la tecnología hiperespectral fue desarrollada inicialmente para detección remota, su aplicación se ha extendido al campo de la ingeniería biomédica como una poderosa herramienta analítica para estudios biológicos y biomédicos [33].

El campo de la imagen hiperespectral es una modalidad emergente para aplicaciones médicas. Este campo presenta un gran potencial aplicable a la diagnosis no invasiva (la captación de la imagen no implica la rotura física del tejido) de enfermedades y a la orientación quirúrgica. La luz captada por el tejido biológico se ve sometida a múltiples dispersiones debido a la inhomogeneidad de estructuras biológicas y a la absorción, principalmente en hemoglobina, melanina y agua, puesto que se propaga a través del tejido [34, 35]. Una de las ventajas importantes de esta técnica es que, para cada píxel de la imagen, se puede obtener información espectral de la reflexión, absorción o *fluorescencia*⁶. Se sabe que la absorción, la fluorescencia y las características de dispersión del tejido cambian a medida que la enfermedad evoluciona [36], lo que puede ser útil para detectar los cambios bioquímicos que no pueden ser identificados con los métodos tradicionales de imágenes en escala de grises o color [33]. Por tanto, la absorción, transmisión o fluorescencia de la luz captada por el tejido empleando la imagen hiperespectral, nos aporta información cuantitativa en la diagnosis de la patología del tejido [36-39].

De acuerdo a la teoría electromagnética, diferentes componentes bioquímicos normalmente tienen diferentes firmas espectrales [40]. Estas firmas son generadas por interacciones entre los materiales y las ondas electromagnéticas, tales como transiciones de electrones, vibraciones o rotaciones atómicas y moleculares. Los cambios biológicos y

⁶ La célula en estados de enfermedad, generalmente tiene velocidades de metabolismo diferentes o estructuras alteradas, por lo tanto hay variaciones en los espectros de emisión de fluorescencia [114].

patológicos en tejidos y órganos están íntimamente relacionados con el espectro de manera que, las características espectrales en diferentes regiones del espectro proporcionan firmas espectrales diferenciables, haciendo posible que los cambios patológicos sean distinguibles.

En años recientes, se han hecho posibles múltiples aplicaciones en el campo de la medicina gracias al avance en cámaras hiperespectrales, métodos de análisis de imagen y en la potencia de computación. Estas aplicaciones principalmente cubren las regiones del espectro ultravioleta (*UV*), visible (*VIS*) e infrarrojo cercano (*NIR*). En la Tabla 3.1 se muestra el rango espectral desde la región ultravioleta hasta el infrarrojo medio (*MIR*) (200-25000 nm) [41].

Tabla 3.1. Definición de los rangos espectrales (en nanómetros de la radiación electromagnética (desde la región UV hasta la región MIR).

Región espectral	Rango espectral (nm)
Ultravioleta (UV)	200-400
Visible (VIS)	400-780
Infrarrojo cercano (NIR)	780-2500
Infrarrojo medio (MIR)	2500-25000

La luz visible penetra solamente de 1 a 2 mm bajo la piel, de ese modo se obtiene información *subpapilar*⁷ [42], mientras que la luz de la región NIR penetra a mayor profundidad en el tejido que la radiación del rango VIS o del rango MIR [43]. Por tanto, la aplicación de radiación en el rango NIR es preferible en orientación quirúrgica debido a su profundidad de penetración en el tejido, ya que puede ayudar al cirujano a ver a través del tejido conjuntivo⁸ con el fin de visualizar estructuras anatómicas críticas de interés, que no son visibles a simple vista, y a detectar moléculas con un determinado espectro conocido [32]. De esta manera, expandiendo el rango de radiación más allá del espectro visible, se puede obtener información adicional que ayuda a caracterizar aún más las células de interés [44].

La imagen hiperespectral también puede ser entendida como un conjunto de datos tridimensional, dos dimensiones espaciales y una dimensión espectral. La Figura 3.8 muestra el concepto del cubo hiperespectral captado por un sistema de imagen hiperespectral, haciendo énfasis en estas tres dimensiones. Debido a su estructura

⁷ La región subpapilar se encuentra debajo de las papilas dérmicas.

⁸ Es el tejido que conecta un tejido con otro, forma una continuidad con tejido epitelial, músculo y tejido nervioso, lo mismo que con otros componentes de este tejido para conservar al cuerpo integrado desde el punto de vista funcional [115].

intrínseca, el cubo hiperespectral puede ser visto como un cubo tridimensional o como una pila de múltiples imágenes bidimensionales en la que la cara del cubo es una función de las coordenadas espaciales y la profundidad es una función de la longitud de onda [33]. Un sistema de imagen hiperespectral puede proporcionar una solución espacial, a partir de la imagen hiperespectral obtenida, de manera que dicha solución proporcione información diagnóstica sobre la fisiología, la morfología y la composición del tejido [32].

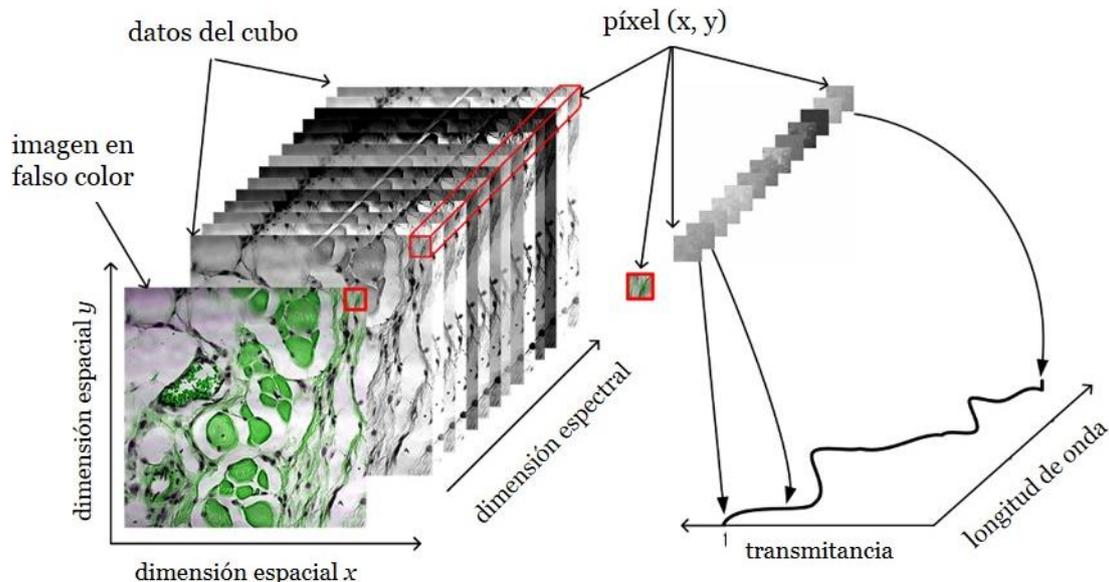


Figura 3.8: Dimensiones espaciales (x, y) y dimensión espectral del cubo hiperespectral.

Por consiguiente, la tecnología de la imagen espectral se extiende al campo de la ingeniería biomédica para estimar el estado de los tejidos biológicos, ya que puede aprovecharse la ventaja de la relación espacial de los diferentes espectros en una vecindad. Esta tecnología abre un nuevo abanico de posibilidades en las ciencias biológicas, de manera que los científicos pueden ser capaces de identificar y cuantificar las relaciones entre las moléculas biológicamente activas, observando organismos vivos de manera no invasiva, realizando análisis histopatológicos y de fluorescencia, y mejorando el entendimiento biológico de las enfermedades [33].

El análisis mediante imágenes facilita la adquisición de información diagnósticamente útil a partir del conjunto de datos hiperespectrales médicos de los tejidos, a nivel celular y molecular, por tanto, es crucial en la investigación, diagnosis y tratamiento de las enfermedades. Un cubo hiperespectral con alta resolución espacial y espectral puede aportar mayor información diagnóstica. Sin embargo, esta alta resolución también dificulta la realización del análisis automático de los datos hiperespectrales. En concreto, la

complejidad radica en los siguientes aspectos: (1) mayor redundancia en los datos debido a la alta correlación en bandas adyacentes, (2) variabilidad de firmas hiperespectrales y (3) la alta dimensionalidad. Con la abundante información espectral y espacial disponible, los métodos avanzados de clasificación de imágenes para datos hiperespectrales requieren extraer, separar y clasificar la información espectral relevante. El objetivo no es solamente diferenciar entre los diferentes tejidos (tejidos sanos y tejidos enfermos) y proporcionar mapas diagnósticos, sino que además se precisa descomponer mezclas en el espectro de componentes moleculares puros y correlacionar estos marcadores biológicos moleculares con estados patológicos. Aunque los métodos de análisis hiperespectral han sido intensivamente investigados en áreas de detección remota, su desarrollo y aplicación en el dominio médico le siguen a una gran distancia. Los primeros pasos en el análisis hiperespectral generalmente involucran el preprocesado, la extracción y selección de características, técnicas de *unmixing*⁹ y la clasificación de los datos [45].

3.3.3. Técnicas de análisis hiperespectral

El análisis hiperespectral se basa en la capacidad de los sensores hiperespectrales para adquirir imágenes digitales en una gran cantidad de canales espectrales muy cercanos entre sí, obteniendo para cada píxel, una firma espectral característica de cada material [28]. Este proceso facilita la identificación y cuantificación de los materiales presentes en la escena [46, 47].

La Figura 3.9 ilustra el procedimiento de análisis hiperespectral mediante un sencillo diagrama de ejemplo. Como puede apreciarse, la capacidad de captación de este sensor es mucho más avanzada que la de otros dispositivos y permite la obtención de píxeles formados por 224 valores espectrales, a partir de los cuales puede obtenerse una firma espectral característica que será utilizada en el proceso de análisis.

⁹ *Spectral unmixing* consiste en la estimación de la fracción de área del píxel cubierta por cada material presente en la escena [116].

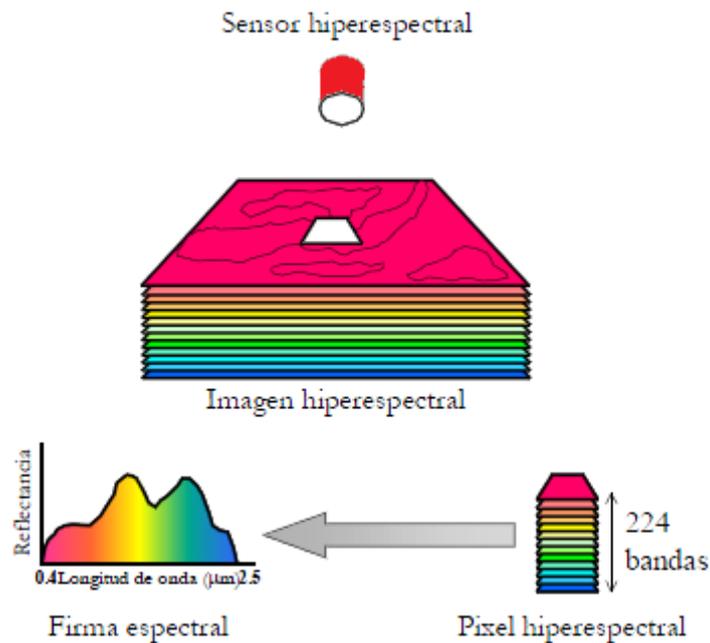


Figura 3.9: Procedimiento de análisis hiperespectral.

Como ya sabemos, el potencial de estas imágenes es la capacidad de obtener una gran cantidad de información, permitiendo distinguir clases y objetivos de una manera más detallada. Sin embargo, esta gran ventaja se convierte también en una desventaja cuando no se dispone de suficiente capacidad computacional para tratar y almacenar estos cientos de bandas. Nos enfrentamos entonces a los problemas de alta dimensionalidad de los datos, y de redundancia que se comentaban en el apartado 3.3.2. La alta dimensionalidad podemos apreciarla si nos hacemos una idea del tamaño total de una imagen de este tipo, multiplicando el tamaño del píxel en bits, por el tamaño de una banda individual y por el número de bandas totales. La redundancia de la información, es decir, la repetición de muchos patrones espectrales, puede llegar a ser bastante significativa, resultando en muchos casos un inconveniente cuando se quieren utilizar métodos estadísticos de clasificación. De ahí, que las aproximaciones o técnicas geométricas y no paramétricas sean más apropiadas en muchos casos.

Por tanto, se hace necesario realizar un conjunto de actividades y técnicas de procesamiento tanto hardware como software capaz de encargarse de la complejidad intrínseca de los datos hiperespectrales de manera efectiva (gran dimensionalidad) [48]. A pesar de que la cadena de procesamiento de datos hiperespectrales no resulta un procedimiento fácil de definir de manera consistente, se puede aplicar una serie de

recomendaciones en cuanto a la definición de una cadena de procesamiento adecuada para datos hiperespectrales [49], la cual viene dada por dos partes claramente diferenciadas.

La primera parte de esta cadena es un proceso específico para obtener las imágenes hiperespectrales como muestras libres de ruido, proporcionando estos datos listos para su procesamiento e identificación de objetivos [45]. Esta parte se divide en los siguientes pasos:

- Creación del cubo hiperespectral a partir de la imagen obtenida por la cámara.
- Calibración del dato.
- Filtrado de ruido.
- Reducción espectral para reducir la dimensionalidad de la imagen, con el fin de eliminar la redundancia espectral y que su procesado sea más fácil.
- Normalización de las muestras.
- Extracción de las muestras del cubo hiperespectral empleando funciones de contorno.

Una vez que los datos de interés han sido obtenidos y preprocesados para ser empleados correctamente, existe la necesidad de extraer información relevante de los conjuntos de datos recogidos. Por tanto, la segunda parte de la cadena de procesamiento de datos se divide en los siguientes pasos adicionales:

- Extracción de características para reducir la alta dimensionalidad y estudiar la naturaleza de cada tejido.
- Clasificación.

Sobre esta parte es importante destacar que cualquier cadena de procesamiento de datos en cualquier ámbito científico tiene que ser flexible y adecuarse no solo a su aplicación sobre distintos escenarios, sino también a los distintos tipos de resolución espacial y espectral que proporcionan los diversos instrumentos existentes.

3.4 Preprocesado de datos hiperespectrales

Cuando se pretenden abordar problemas complejos como es el caso de la clasificación de imágenes hiperespectrales, es necesario emplear técnicas de pre-procesado que reduzcan la cantidad de datos a procesar sin pérdida notable de la información. Se trata de métodos

que por sí solos no llegan a ninguna conclusión, pero que sirven como paso previo para que los algoritmos de clasificación puedan implementarse con una mayor eficacia [4].

En muchos problemas es necesario emplear estas técnicas debido al fenómeno de Hughes, que establece que la bondad conseguida con un clasificador aumenta con la dimensionalidad de los datos hasta cierto punto, a partir del cual decrece conforme se incorporan nuevas [50]. En una clasificación, el error mínimo esperado que se puede alcanzar se conoce como el error de Bayes [51]. Este error es una función que decrece con la dimensionalidad de los datos.

Al añadir una nueva característica se añade información, por lo que es lógico suponer que la clasificación será tan buena como cuando esta información no se había introducido. En la práctica, cuando se añade una nueva característica a los datos el error de Bayes disminuye, pero al mismo tiempo las desviaciones del error de clasificación aumentan. Este incremento se debe al hecho de que se necesitan calcular más parámetros partiendo del mismo número de ejemplos. Si el incremento de las desviaciones en la clasificación del error es mayor que el decremento del error de Bayes, entonces el uso de la característica adicional degrada la regla de decisión. Este fenómeno se denomina efecto Hughes o maldición de la dimensionalidad. Además, cuando la dimensionalidad de los datos y la complejidad de la regla de decisión aumentan, el efecto Hughes puede llegar a ser aún más grave [28].

Para resumir, el rendimiento de un clasificador supervisado decrece con la dimensionalidad de los datos a menos que el número de muestras sea infinito [50].

A continuación, se presentan algunas de las principales técnicas de preprocesado que se emplean para reducir la carga computacional del proceso de clasificación.

- **Análisis de Componentes Principales (Principal Component Analysis, PCA).** Este método aprovecha la elevada correlación existente entre bandas consecutivas de una imagen hiperespectral. Obtiene un conjunto reducido de bandas (*autovectores*) poco correlacionadas entre sí (ortogonales en el caso ideal) que contienen la mayor parte de la información de la imagen original. Los autovectores quedan ordenados en orden decreciente del porcentaje de varianza de la imagen original [52, 53].
- **Fracción Mínima de Ruido (Minimum Noise Fraction, MNF).** El análisis de componentes principales visto en el apartado anterior maximiza la

varianza. Esto no siempre lleva a obtener la calidad deseada de la imagen (por ejemplo, que el ruido de esta sea mínimo). La transformación MNF justamente trata de resolver ese problema a partir de maximizar la relación señal a ruido (SNR) en vez de la varianza. Para separar la señal del ruido lo primero que realiza este algoritmo es una transformación PCA, quedando así el ruido aislado en las última bandas [54].

- ***Selección de características (Feature Selection, FS)***. Selecciona las bandas que, dado un conjunto de patrones de entrenamiento, obtienen mayor información para realizar una clasificación más eficiente. En la literatura existen diversos métodos [55].

Hasta ahora, todas las técnicas que se han presentado tienen como objetivo reducir el número de bandas espectrales de las muestras a clasificar, por tanto, son técnicas de preprocesado espectral. Sin embargo, en la clasificación de imágenes hiperespectrales también es habitual emplear técnicas de preprocesado espacial. Entre estas últimas se encuentran diversas técnicas de filtrado, la morfología matemática que permite analizar estructuras espaciales dentro de una imagen o el análisis de texturas.

También es habitual aplicar algoritmos de *unmixing* y extracción de *endmembers* [45]. El *unmixing* hiperespectral es un problema de separación de fuentes (materiales de la escena) que son dependientes estadísticamente hablando y deben combinarse en una función no lineal. Existen diferentes estrategias (espectrales frente a técnicas híbridas) que están siendo comparadas en busca de una solución eficiente teniendo en cuenta la alta dimensionalidad de los datos.

En el caso que nos ocupa en este PFC, las muestras son datos espectroscópicos, por lo que las técnicas de preprocesado espacial, *unmixing* y extracción de *endmembers* carecen de sentido.

Por otro lado, el número de bandas que presentan estos datos espectrográficos es significativamente mayor al de las imágenes hiperespectrales. Sin embargo, al tratarse de muestras de un único píxel, no se ha contemplado el uso de algoritmos de reducción dimensional, ya que se ha pretendido trabajar con la muestra en su espectro completo.

3.5 Clasificación de datos hiperespectrales

La forma más simple de abordar el problema de clasificación de píxeles en una imagen hiperespectral es considerar que los píxeles de interés están compuestos por un único material, utilizando las técnicas convencionales de clasificación de patrones [56] pero con mayor precisión, debido al elevado número de bandas espectrales disponibles.

En la práctica, el uso de sensores hiperespectrales permite una mejor determinación de la composición interna de cada píxel, que raramente estará compuesto por un único material, puesto que el fenómeno de la mezcla es muy habitual en el mundo real, independientemente de cual sea la escala espacial considerada [57].

Existe un conjunto de técnicas de clasificación de patrones que realizan la interpretación de una escena obtenida en base a la asignación de una etiqueta o clasificación individual a cada uno de los píxeles de la misma. Estas técnicas ofrecen resultados interesantes en determinadas aplicaciones, particularmente en las que se destacan a continuación.

- **Clasificación temática.** Las técnicas de clasificación han sido utilizadas de forma satisfactoria en aplicaciones que tienen como objetivo la obtención de un mapa temático en el que cada píxel de la imagen hiperespectral está debidamente etiquetado como perteneciente a una clase concreta [58]. Puede existir una clase adicional denominada “*fondo*” o “*resto*” que representa a los píxeles que no han sido clasificados en ninguna de las clases anteriores. El resultado ideal se obtiene cuando todas las clases, incluyendo la clase “*fondo*”, son mutuamente excluyentes entre sí. La tarea clave en este tipo de aplicaciones suele ser la determinación del número de clases y la caracterización de las mismas en términos de datos de entrenamiento o información de verdad terrena¹⁰ (*ground-truth*).
- **Detección de targets.** Las técnicas de clasificación también han sido utilizadas de forma muy extensa en aplicaciones de detección de objetivos o *targets* en imágenes hiperespectrales [59]. En este tipo de aplicaciones, el objetivo

¹⁰ El término verdad terrena determina la exactitud en el proceso de clasificación a partir de un conjunto de entrenamiento, realizado durante las técnicas de aprendizaje supervisado y validación. El proceso de recopilación de datos objetivos adecuados para el testeo del clasificador se llama verificación sobre el terreno [117].

fundamental es la identificación de un material u objeto específico entre todos los píxeles de la imagen.

Conceptualmente, los dos problemas mencionados pueden considerarse como un problema de clasificación binario:

- En la detección de targets, los píxeles son clasificados en dos clases, denominadas “objeto” y “fondo”, dependiendo de si contienen o no el target buscado.
- En la clasificación temática, hay varias clases posibles asociadas a distintos objetos. El objetivo es, en última instancia, determinar la existencia o no de cada uno de los objetos considerados en cada píxel, situación que puede expresarse como un problema de clasificación binario [60].

El problema de clasificación binaria puede ser formulado matemáticamente del siguiente modo. Sea R el espacio N -dimensional formado por todos los píxeles de la imagen hiperespectral y $u = (u_1, u_2, \dots, u_N)^T$ un vector N -dimensional, asociado a un píxel concreto. La clasificación binaria consiste en dividir el espacio R en dos regiones, R_o y R_f , de forma que u es clasificado como “objeto” si $u \in R_o$ y como “fondo” si $u \in R_f$. Este problema puede ilustrarse de forma gráfica utilizando un “*diagrama de dispersión*” en dos bandas poco correlacionadas de la imagen hiperespectral.

Como puede apreciarse en el diagrama de dispersión de la Figura 3.10, la situación ideal de un problema de clasificación se produce cuando la separación entre objeto y fondo está claramente definida en agrupaciones o *clusters* claramente diferenciables.

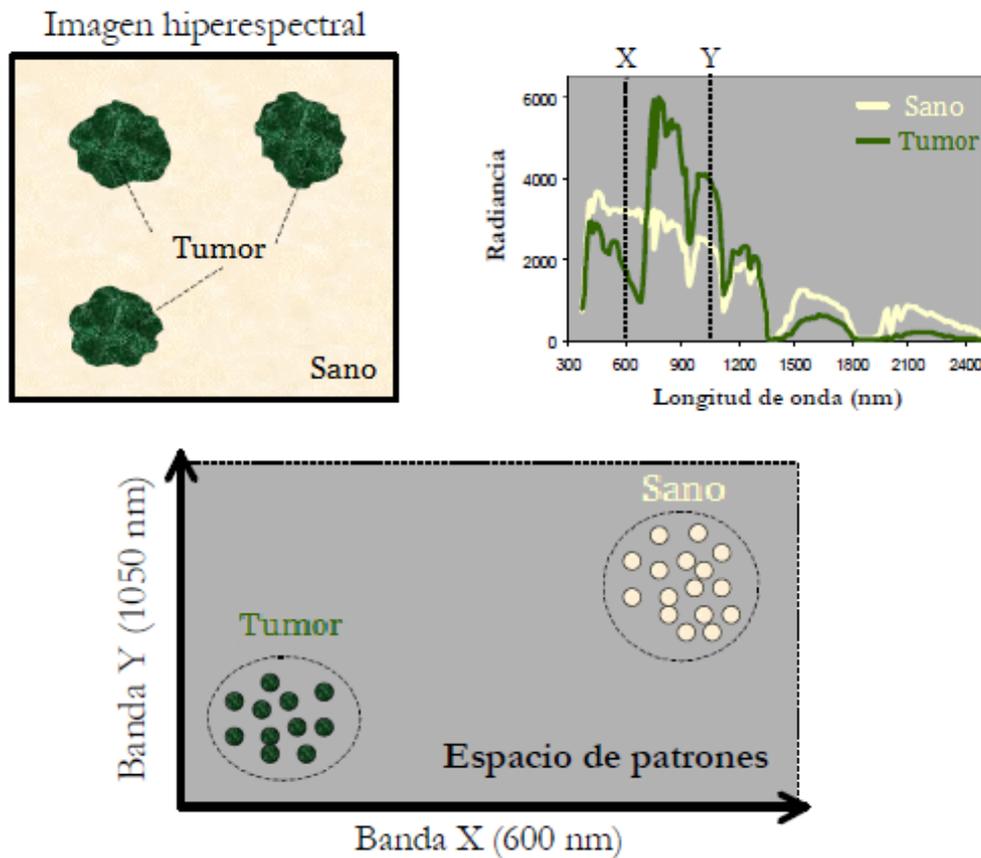


Figura 3.10: Ejemplo de clasificación binaria de imágenes hiperespectrales (valores ficticios).

Los algoritmos de clasificación pueden dividirse en dos grandes categorías [58]: algoritmos supervisados y algoritmos no supervisados.

- **Algoritmos de clasificación no supervisados.** Presuponen que no existe ningún tipo de conocimiento a priori sobre las clases existentes. El objetivo de estas técnicas es identificar, de forma automatizada, clases o agrupaciones de píxeles, utilizando para ello una métrica de similitud. Este tipo de algoritmos solamente requiere el conjunto de muestras a clasificar. El problema principal de estos clasificadores es encontrar el número correcto de clases para identificarlas y para ello se buscan regularidades presentes en los datos, se extraen rasgos o agrupan patrones según similitud. Las técnicas empleadas suelen denominarse métodos de agrupamiento o *clustering*.
- **Algoritmos de clasificación supervisados.** Parten de un cierto conocimiento sobre las clases existentes a partir del cual pueden derivarse criterios de clasificación. Esta aproximación suele venir dada por un paso previo en el que se seleccionan firmas espectrales características de las clases existentes. Se dice

entonces que estas firmas espectrales están etiquetadas para cada una de las clases. Estas muestras se emplean para generar el modelo de un clasificador que permita clasificar nuevas muestras. En este caso, el principal problema es obtener un conjunto grande de muestras completamente etiquetado que permita un error de clasificación pequeño, pero esto puede suponer un gran esfuerzo o coste.

De la categoría de los algoritmos de clasificación supervisados se deriva otro tipo de algoritmos:

- **Algoritmos de clasificación semi-supervisados.** Estos algoritmos usan un conjunto de muestras que contiene tanto datos etiquetados como no etiquetados. Por lo general, suele contener un conjunto de muestras etiquetadas pequeñas respecto a la cantidad de muestras no etiquetadas. Se obtienen nuevas muestras a partir de las muestras etiquetadas mediante algoritmos autónomos que consiguen que el problema de clasificación sea solventado. Por tanto, se emplea cuando el proceso de etiquetado requiere un esfuerzo muy superior a la obtención de muestras, y se desean resultados más precisos que los proporcionados por clasificadores no supervisados.

La clasificación consiste en encontrar una regla de decisión tal que, dada una muestra externa, esta sea asignada a su correspondiente clase. El caso más sencillo es una clasificación supervisada binaria en la que únicamente existen dos clases (ejemplo de la Figura 3.10), con los valores $+1$, si se trata de tejido sano; y -1 , si es tejido tumoral. La búsqueda de una regla de decisión adecuada puede interpretarse como la estimación de una función g , la cual asigna a cada punto del espacio de observación, un valor en el espacio de las clases (-1 y $+1$). Esta búsqueda se lleva a cabo usando un conjunto de datos de entrenamiento (*conjunto de muestras de entrenamiento*) formado por N muestras etiquetadas, distribuidas generalmente mediante una distribución de probabilidad desconocida $P(x, y)$.

Lo que se busca con la función g es que se clasifique correctamente un conjunto de datos que no pertenezcan al conjunto de entrenamiento. Esta función suele tomar la forma:

$$g(x) = \text{sign}(f(x)) \quad (3.4)$$

A partir de esta función, las muestras nuevas (*conjunto de muestras de test*) se asignarán a la clase $+1$ (tejido sano) si $f(x) > 0$ o a la clase -1 (tejido tumoral) si $f(x) < 0$. Para un resultado óptimo, tanto las muestras de entrenamiento como las de test deben tener una misma distribución de probabilidad $P(x, y)$. La mejor función $f(x)$ será aquella que tenga la menor esperanza del error de clasificación, es decir, la que requiere un menor coste.

$$C(f) = \int l(f(x), y) \partial P(x, y) \quad (3.5)$$

En la ecuación (3.5) se representa el coste, donde l es la función de coste que representa la penalización que se produce al tomar, a partir de la observación de x , una decisión basada en el valor de $f(x)$ siendo el valor correcto y .

Por otro lado, tanto los datos de entrenamiento como los de test son vectores en los que cada dimensión constituye una característica o atributo de la muestra. Así por ejemplo, si el conjunto a clasificar fuera coches, cada ejemplo x será un vector que contenga las características del coche (longitud, anchura, velocidad máxima, etc.). En este proyecto estos atributos corresponderán a la respuesta en absorbancia o transmitancia para cada una de las bandas del espectro [61, 62].

A continuación, se describen algunos de los principales algoritmos de clasificación no supervisados y supervisados, poniéndose especial énfasis en el algoritmo supervisado SVM, por ser el elegido en el desarrollo de este proyecto.

3.5.1. Algoritmos de clasificación no supervisados

Entre las técnicas de clasificación de píxeles de forma no supervisada en imágenes hiperespectrales existentes, destaca el método *K-Means* [63] y el método *ISODATA* (*Interactive Self-Organizing Data Analysis Techniques*) [64]. No obstante, la literatura reciente demuestra que los resultados obtenidos por estas dos técnicas no han sido demasiado satisfactorios, salvo en aplicaciones muy concretas [65].

3.5.1.1. K-Means

El algoritmo *K-Means* supone la existencia de K clases (parámetro que debe ser determinado a priori) y realiza una agrupación de los píxeles de la imagen en dichas clases utilizando los vectores de datos contenidos en cada píxel. Pertenece al grupo de los métodos puramente estadísticos basados en los espectros promedio de dichas clases.

El algoritmo procede del siguiente modo:

1. Se introduce el número de clases que se desea obtener y se seleccionan los centros de gravedad, *centroides* o medias de las clases para su comparación con el resto de los datos. Normalmente estos son seleccionados aleatoriamente de entre los datos de entrada.
2. Para cada muestra se calcula la similitud del mismo a todos los centros de gravedad, utilizando técnicas de mínima distancia, siendo finalmente incluido en el que mejor se ajusta, es decir, el que presenta menor distancia a los centroides de cada clase.
3. Se recalcula el centro de gravedad de la clase y se vuelven a clasificar todos los píxeles. Todos los píxeles serán clasificados si se limita la desviación estándar o la distancia máxima de búsqueda.

El principal inconveniente consiste en que es necesario encontrar una técnica automática que introduzca el número de clases para evitar tanto una segmentación excesiva como una segmentación demasiado pobre. También se deberá comprobar que la carga computacional de comparar cada píxel con los centros de gravedad en cada iteración no sea excesiva [63, 66].

3.5.1.2. ISODATA

El método ISODATA también necesita la inicialización de un parámetro K relativo al número de clases deseadas en una etapa previa a la ejecución del algoritmo. Sin embargo, este método necesita información relativa al número mínimo de firmas pertenecientes a una clase. Si el valor inicial de K es bajo, la dispersión entre clases diferentes puede ser muy alta; por el contrario, si el valor inicial de K es alto, la distancia entre clases puede ser muy pequeña, provocando la división de una misma clase en varias clases similares entre sí [64, 65].

3.5.2. Algoritmos de clasificación supervisados

En los algoritmos de clasificación supervisados se puede diferenciar dos etapas fundamentales en el proceso de clasificación: la etapa de entrenamiento y la etapa de verificación o test. En la etapa de entrenamiento se introduce en el sistema clasificador un conjunto de muestras etiquetadas para obtener la función de clasificación o clasificador entrenado (Figura 3.11), mientras que en la etapa de verificación o test se introduce un conjunto de muestras diferente que proporcionarán el coste o error de clasificación (Figura 3.12). Es necesario destacar que en ambos conjuntos de muestras (entrenamiento y test), la clase a la que pertenece realmente cada muestra es conocida a priori.

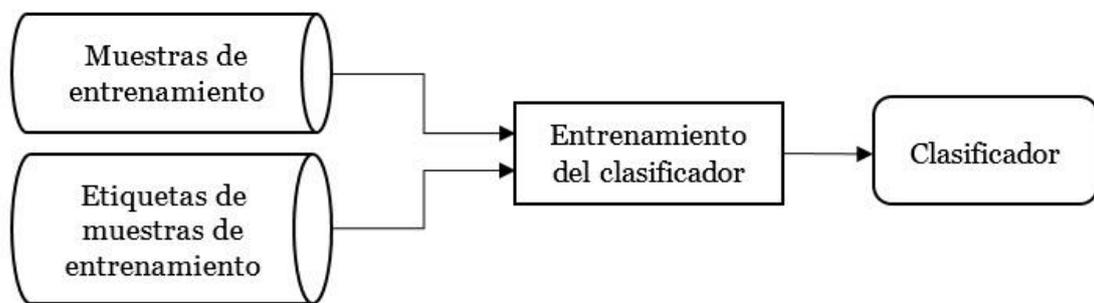


Figura 3.11: Etapa de entrenamiento de un clasificador supervisado.

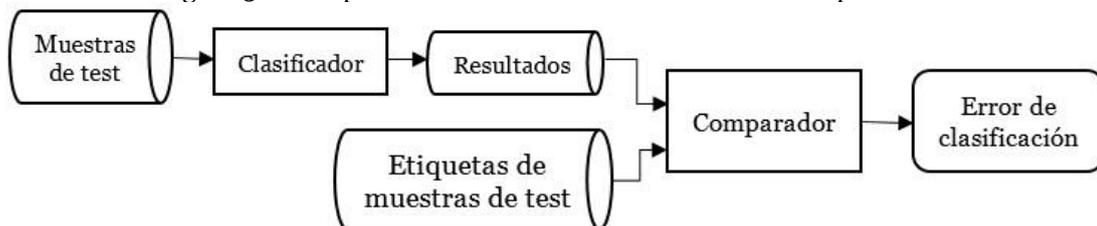


Figura 3.12: Etapa de verificación o test, de un clasificador supervisado.

Dentro de las técnicas de clasificación supervisadas, destacan los filtros de similitud (*matched filters*) y el método de *mapeo de ángulo espectral* (*Spectral Angle Mapper, SAM*), ambos basados en estadísticas de primer orden¹¹. También pueden encontrarse en esta categoría otros clasificadores como las técnicas de *vecino más cercano* (*Nearest Neighbour*), *distancia mínima* (*Minimum Distance*), *clasificación paralelepípedica* (*Parallelepiped*), *máxima probabilidad* (*Maximum Likelihood*), *regresión logística multinomial* (*Multinomial Logistic Regression, MLR*), *análisis discriminante lineal* (*Linear Discriminant Analysis, LDA*), *campos aleatorios de Markov* (*Markov Random*

¹¹ Al tratarse de estadísticas de orden inferior solo se analiza un píxel y su vecino más inmediato. Además, las estadísticas de primer orden se basan en el vector de densidad de probabilidad (obtenido a partir de la imagen original) [118].

Fields, MRFs), bosques aleatorios (*Ramdon Forest, RF*), redes neuronales artificiales (*Artificial Neural Networks, ANN*), o la técnica escogida para el diseño de los sistemas de clasificación empleados en el presente proyecto: *máquinas de vectores soporte (Support Vector Machines, SVM)* [26]. A continuación se describen algunos de estos algoritmos de clasificación supervisados.

3.5.2.1. Spectral Angle Mapper (SAM)

El algoritmo de mapeo del ángulo espectral se basa en la medida de la similitud espectral entre dos firmas espectrales. Esta similitud α se obtiene considerando cada espectro como un vector en un espacio N-dimensional y empleando la regla del producto escalar (3.6), donde n es el número de bandas espectrales y, e_1 y e_2 son los dos espectros que se están comparando [67, 68].

$$\alpha = \cos^{-1} \left(\frac{\langle e_1 e_2 \rangle}{\|e_1\| \|e_2\|} \right) = \cos^{-1} \left(\frac{\sum_{i=1}^n e_1 e_2}{(\sum_{i=1}^n e_1^2)^{1/2} \cdot (\sum_{i=1}^n e_2^2)^{1/2}} \right) \quad (3.6)$$

Un ángulo α pequeño indica una elevada similitud entre los espectros, mientras que ángulos elevados indican baja similitud. En la Figura 3.13 se muestra el concepto de ángulo espectral para dos bandas concretas. Obsérvese que a medida que el ángulo se hace más pequeño, los espectros son más parecidos. Esta medida es además independiente de la magnitud de los vectores, lo que permite aislar la medida de las condiciones de iluminación.

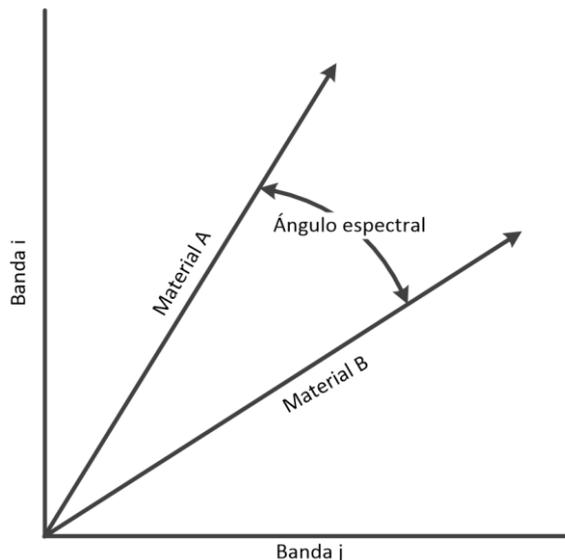


Figura 3.13: Representación en dos dimensiones del concepto de ángulo espectral.

El algoritmo comienza calculando la media de todas las muestras de una agrupación. A continuación, se calcula el ángulo espectral de cada una de las muestras de la clase respecto a la media calculada previamente, obteniéndose un vector con los distintos valores de ángulo espectral. El valor del mayor ángulo espectral se compara con un umbral establecido previamente, si resulta ser mayor a este umbral, la muestra que provocó ese ángulo se elimina, se recalcula la media con el resto de muestras y se vuelve a repetir el proceso. El algoritmo finaliza cuando ningún ángulo espectral, entre la media y cada una de las muestras de la agrupación, supera el valor del umbral establecido.

3.5.2.2. Multinomial Logistic Regression (MLR)

En estadística, la regresión logística multinomial generaliza el método de regresión logística para problemas *multiclase*, es decir, con más de dos posibles resultados discretos. Se trata de un modelo que se utiliza para predecir las probabilidades de los diferentes resultados posibles de una distribución categórica como variable dependiente, dado un conjunto de variables independientes [69]. Este clasificador probabilístico ha emergido recientemente como una técnica factible para el análisis de datos hiperespectrales [70].

Fundamentos teóricos de MLR

En un problema de aprendizaje supervisado de m clases en el que el conjunto de muestras etiquetadas de entrenamiento (*training*) es $D_L = \{(x_1, y_1), \dots, (x_L, y_L)\}$, donde $x_1 \in \mathbb{R}^d$ es el vector características d -dimensional e y_1 la etiqueta de la clase correspondiente. En una codificación "1-de- m ", $y_1 = [y_1^{(1)}, \dots, y_1^{(m)}]$ es un vector binario, tal que $y_i^{(c)} = 1$ y $y_i^{(j)} = 0$, para $j \neq c$ indica que la muestra i pertenece a la clase c . En MLR, las probabilidades a *posteriori* de las clases son modeladas como:

$$\log P(y_i^{(c)} = 1|x) = x^T w^{(c)} - \log \sum_{k=1}^m e^{x^T w^{(k)}} \text{ para } c = 1, \dots, m \quad (3.7)$$

donde $w^{(c)} \in \mathbb{R}^d$ es el vector de pesos de la clase c . Hay que tener en cuenta que, puesto que $\sum_{c=1}^m P(y^{(c)} = 1|x) = 1$, uno de los vectores de pesos es redundante; se puede escoger arbitrariamente $w^{(m)} = 0$, y considerar el vector $(d(m-1))$ -dimensional $w = [(w^{(1)})^T, \dots, (w^{(m-1)})^T]^T$. Se puede lograr la estimación de w maximizando la *log-verosimilitud* (*log-likelihood*), $l(w)$, (con $Y \equiv \{y_1, \dots, y_L\}$).

$$l(w) \equiv \log P(Y|w) = \sum_{i=1}^L \left[\sum_{c=1}^m y_i^{(c)} x_i^T w^{(c)} - \log \left(\sum_{j=1}^m e^{(x_i^T w^{(j)})} \right) \right] \quad (3.8)$$

En la presencia de una $P(w)$ a *priori*, se busca la estimación de un *máximo a posteriori* (MAP), $\hat{w} = \arg \max_w \{l(w) + \log p(w)\}$. Realmente, si los datos de entrenamiento son separables, $l(w)$ es ilimitado, y uno previo es crucial [71].

Limitaciones

Se debe mencionar que existe un problema con \hat{w} puesto que, aunque se trata de un problema convexo [72], es muy difícil su computación debido a que $l(w)$ es una función no cuadrática y el término $\log(p(w))$ es *non-smooth*¹². El algoritmo *sparse MLR (SMLR)* soluciona este problema con la complejidad $O((d(m-1))^3)$. Sin embargo, la mayoría de los conjuntos de datos hiperespectrales escapan al alcance de este algoritmo puesto que, sus análisis se hace insostenible a medida que incrementa el número de clases.

Ejemplos de aplicación

La naturaleza de los algoritmos MLR tiene sus ventajas en el campo de la segmentación y clasificación supervisada de imágenes hiperespectrales. Esto es debido a que dichos algoritmos son capaces de ayudar a entrenar un sistema con muestras de entrenamiento no etiquetadas. Ello implica una gran importancia en aplicaciones de *remote sensing (teledetección)*, puesto que la adquisición de datos de entrenamiento etiquetados es muy costosa en términos de financiación y tiempo. Por tanto, existen estudios que emplean estas técnicas en el campo de la agronomía para averiguar los diferentes tipos de cultivo.

Por otro lado, también se pueden encontrar algoritmos MLR en múltiples aplicaciones fuera del ámbito de la imagen hiperespectral, tales como la clasificación de textos académicos o en investigación sobre gastroenteritis por salmonela.

¹² Las funciones *non-smooth* incluyen a las funciones no diferenciables y a las funciones discontinuas.

3.5.2.3. Linear Discriminant Analysis (LDA)

LDA se trata de un método fundamental en el análisis de datos, es una generalización del discriminante lineal propuesto originalmente por R. Fisher en 1936 para discriminar entre diferentes tipos de flores. Este método estadístico se emplea en el reconocimiento de patrones y *machine learning (aprendizaje automático)*, para encontrar una combinación lineal de características que permitan caracterizar o separar dos o más clases. La combinación resultante puede ser usada como un clasificador lineal, o más habitualmente, para reducir la dimensionalidad antes de la clasificación [73].

Consiste en encontrar el *hiperplano* de proyección que minimiza la varianza de la distancia *intraclase* y maximiza la distancia entre las medias proyectadas de la clase. De manera similar a *PCA (Principal Component Analysis)*, estos dos objetivos pueden ser solventados resolviendo un problema de *autovalor* con su correspondiente autovector para definir el hiperplano de interés. El hiperplano puede ser empleado para clasificación, reducción dimensional y para la interpretación de la importancia de las características dadas.

Fundamentos teóricos de LDA

La idea que subyace de este método es determinar un *subespacio*¹³ de dimensión menor, en comparación al de las muestras originales, en el que los datos del problema original sean “separables” (Figura 3.14). La *separabilidad* se define en términos de las medidas estadísticas valor medio y varianza. Una de las ventajas de LDA es que la solución puede ser obtenida resolviendo un sistema de autovalor generalizado. Esto permite un procesamiento rápido y masivo de muestras. El algoritmo se propuso originalmente para resolver problemas *biclases* aunque luego el algoritmo se generalizó para ser empleado en problemas multiclase.

¹³ subconjunto de un espacio vectorial, V , que satisface por sí mismo la definición de espacio vectorial con las mismas operaciones que V .

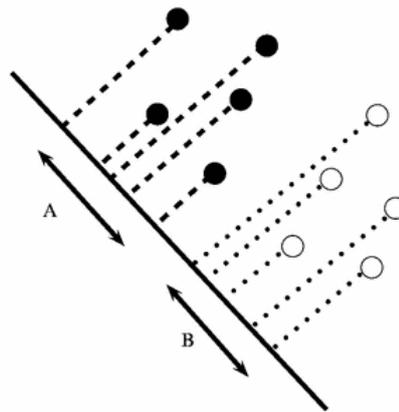


Figura 3.14: Las muestras de dos dimensiones son proyectadas en una dimensión espacial menor (línea), con lo que sus proyecciones sobre la línea son escalares.

El clasificador LDA estándar permite encontrar una matriz de transformación lineal, w^T , que reduce un vector m -dimensional de características, x , a un vector l -dimensional $y = w^T x \in \mathbb{R}^l$, donde $l < m$. Este espacio de características de baja dimensionalidad es seleccionado para satisfacer un criterio dado de maximización de separabilidad entre las distribuciones de las clases. En la Figura 3.15 se ilustra un ejemplo para el problema biclase con muestras bidimensionales.

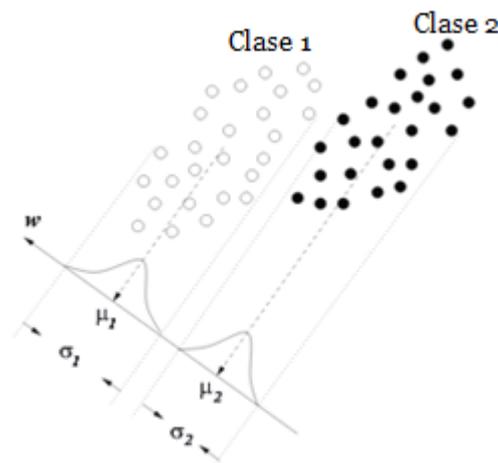


Figura 3.15: Discriminante de Fisher para un problema biclase.

Asumiendo que se dispone de un conjunto m -dimensional de un total de n muestras $X = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$, de K clases $C = \{C_1, C_2, \dots, C_K\}$, en el que el número de muestras por cada clase, N , puede ser diferente en las distintas clases, se desea obtener una proyección, $y^{(i)}$, l -dimensional, a partir de una muestra $x^{(i)}$ sobre el vector unitario w (Figura 3.16):

$$y = w^T x \tag{3.9}$$

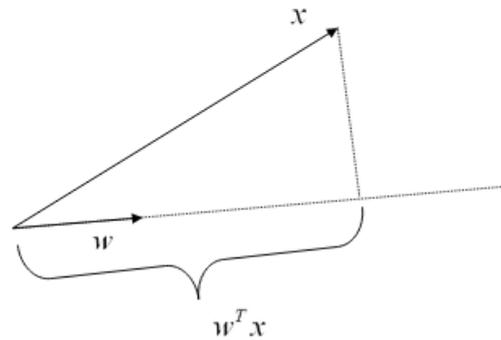


Figura 3.16: Interpretación geométrica de la proyección del vector x .

De todos los posibles hiperplanos de proyección se busca el que maximiza la separabilidad de las proyecciones de diferentes clases. En la Figura 3.17 se ilustra un ejemplo biclase de una distribución de muestras bidimensionales. Como se observa, según el hiperplano de proyección que se escoja (eje de proyección en este caso por tratarse de un problema bidimensional), las muestras proyectadas pueden presentar o no, una separabilidad de las clases.

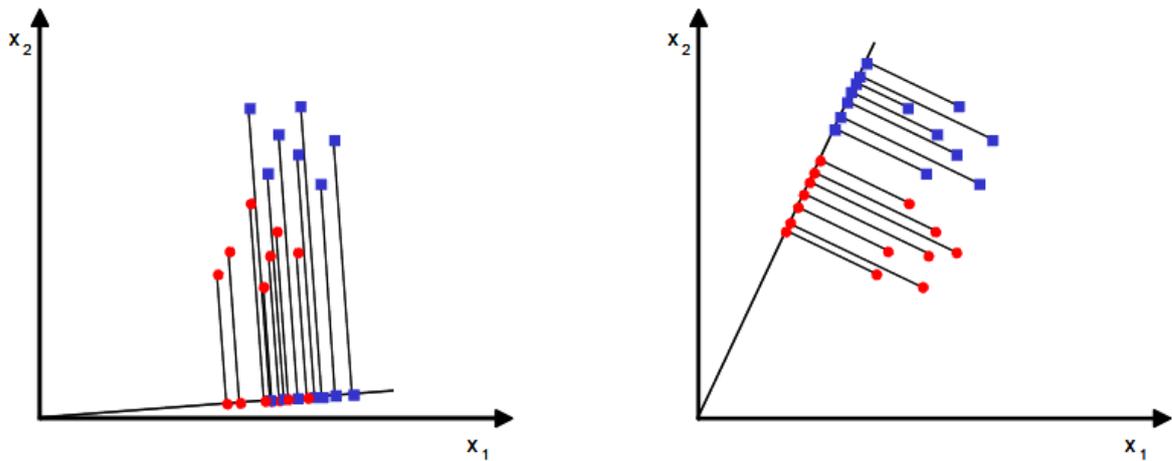


Figura 3.17: Proyección de muestras bidimensionales en diferentes ejes de proyección.

Para encontrar un hiperplano que maximice la separabilidad de las proyecciones, requiere establecerse la definición de dicha separabilidad. Por tanto, se parte de las medias de cada clase, μ_k en el espacio x , y en el espacio y , η_k :

$$\mu_k = \frac{1}{N_k} \sum_{x \in C_k} x \quad (3.10)$$

$$\eta_k = \frac{1}{N_k} \sum_{y \in C_k} y = \frac{1}{N_k} \sum_{x \in C_k} w^T x = w^T \mu_k \quad (3.11)$$

donde N_k es el número de muestras de la clase k .

Con lo que, para el problema biclase, se puede considerar la separabilidad como la distancia entre sus medias:

$$J(w) = |\eta_1 - \eta_2| = |w^T(\mu_1 - \mu_2)| \quad (3.12)$$

Sin embargo, basarse solamente en la distancia entre las medias proyectadas para establecer la separabilidad no es una buena medida puesto que podrían darse casos en los que la dispersión de las muestras de cada clase, presentaran situaciones como la que se ilustra en la Figura 3.18 [74].

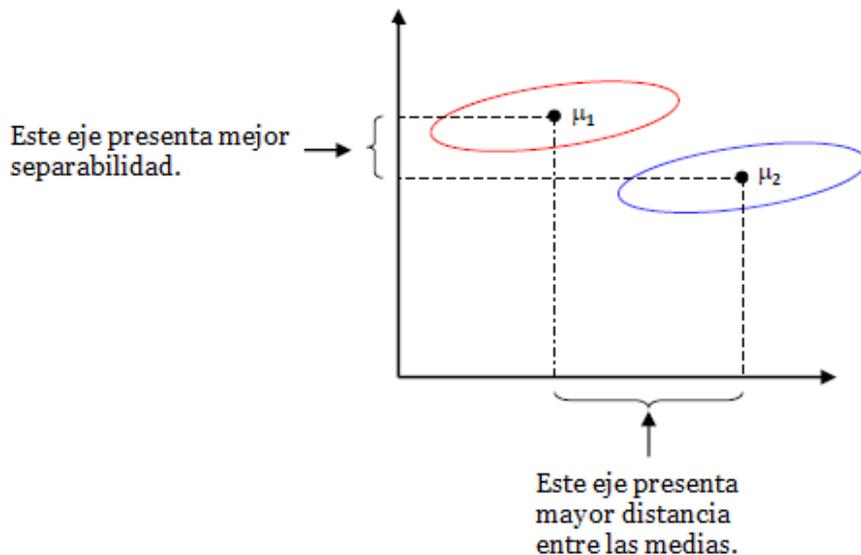


Figura 3.18: Ejemplo biclase y 2-dimensional de la proyección de las medias en ejes diferentes.

Para resolver este tipo de situaciones, Fisher plantea maximizar la distancia entre las medias, normalizadas por una medida de dispersión intraclassa que equivale a la varianza. Para cada clase se define la dispersión como:

$$\sigma_k^2 = \frac{1}{N_k} \sum_{y \in C_k} (y - \eta_k)^2 \quad (3.13)$$

El criterio de Fisher, basado en maximizar la distancia entre las medias de las clases y , al mismo tiempo, minimizar la varianza intraclassa, para el caso biclase; ampliamente usado, es el siguiente:

$$J(w) = \frac{(\eta_2 - \eta_1)^2}{\sigma_1^2 + \sigma_2^2} \quad (3.14)$$

Por tanto se busca un eje de proyección en el que las muestras de una misma clase están muy cerca entre ellas, y al mismo tiempo, las proyecciones de las medias están lo más alejadas posible (Figura 3.19).

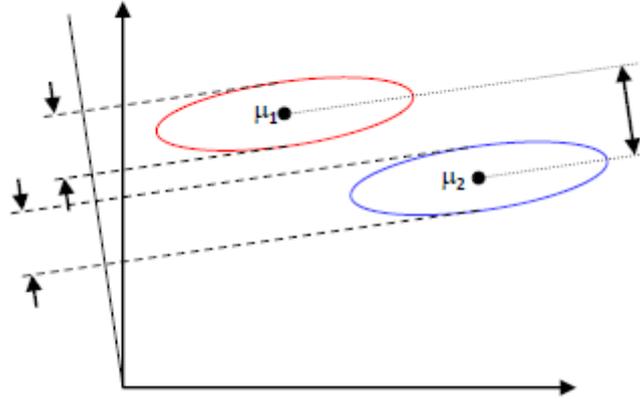


Figura 3.19: Nuevo eje de proyección.

Puesto que la función de decisión es: $y = w^T x$, y las medias y varianzas pueden ser definidas trivialmente, se puede demostrar fácilmente que maximizar el criterio de Fisher es equivalente a maximizar el coeficiente de *Rayleigh*, con respecto al vector de pesos de la función de decisión w en la clasificación multiclase:

$$w^* = \arg \max_w \{J(w)\} = \arg \max_w \left\{ \frac{w^T S_b w}{w^T S_w w} \right\} \text{ (criterio de maximización)} \quad (3.15)$$

$$S_w = \sum_{k=1}^K S_k ; S_k = \sum_{i \in C_k} (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^T \quad (3.16)$$

$$S_b = \sum_{k=1}^K N_k (\mu_k - \mu)(\mu_k - \mu)^T \text{ donde } \mu = \frac{1}{n} \sum_{i=1}^n x^{(i)} \quad (3.17)$$

$$S_T = S_b + S_w \quad (3.18)$$

Donde: S_b es la matriz de dispersión *entre clases* (*between-class*), S_w es la matriz de dispersión *intraclase* (*within-class*) y S_T es la matriz de dispersión total.

Limitaciones

En el caso donde existen más de dos clases (Figura 3.20), se generaliza el discriminante lineal de Fisher para problemas multiclase. En este caso se debe tener en cuenta que se necesita encontrar $(K - 1)$ proyecciones $[y_1, y_2, \dots, y_{K-1}]$ en vez de una sola proyección y . Esto supone medias de $(K - 1)$ vectores de proyección w_i dispuestos en columnas de una matriz de proyección $W = [w_1|w_2| \dots |w_{K-1}]$, es decir, se pasa de: $y_i = w_i^T x \rightarrow y = W^T x$.

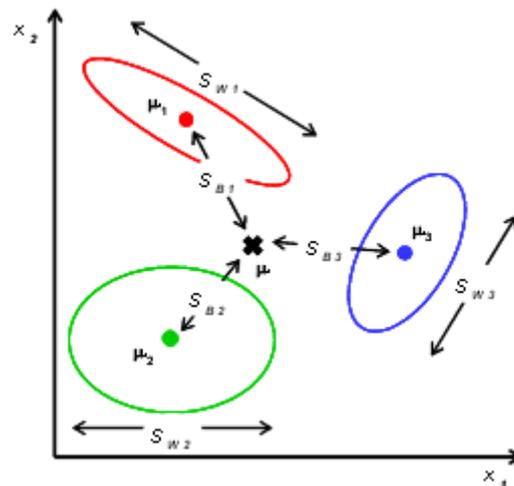


Figura 3.20: Discriminante de Fisher para un problema de 3 clases.

Entonces, si la estimación del error de clasificación establece que se necesitan más características, deben emplearse otros métodos que proporcionen características adicionales. Además LDA es un método paramétrico (que implican probabilidades gaussianas unimodales), con lo que si las distribuciones son significativamente no gaussianas, las proyecciones LDA podrían no soportar estructuras complejas de datos a la hora de clasificarlos (Figura 3.21).

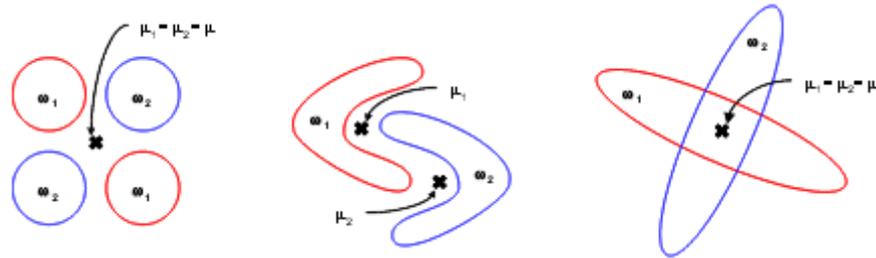


Figura 3.21: Ejemplos de distribuciones que presentan problemas en clasificadores LDA.

Por otro lado, LDA también falla si la información de discriminación se encuentra en la varianza en vez de en la media (Figura 3.22).

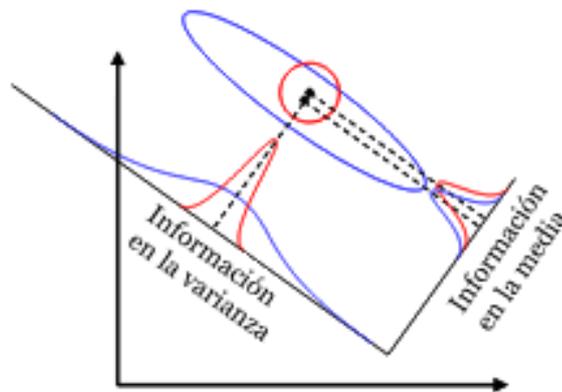


Figura 3.22: Ejemplo de distribución de muestras de dos clases en el que la información de discriminación puede estar en la media o la varianza.

Pese a su buen funcionamiento en muchas aplicaciones, los métodos clásicos basados en LDA no pueden funcionar en problemas mal planteados, cuando el número de características es mayor al número de muestras de entrenamiento. Esto se debe al algoritmo estándar LDA en la computación de las tres matrices de dispersión: matriz de dispersión intraclass, matriz de dispersión entre clases y matriz de dispersión común, definida a través de la matriz de covarianza. En esta situación, para obtener la solución se requiere que la matriz de dispersión común sea invertida, y además, esta debe ser no singular¹⁴. Se pueden seguir diferentes estrategias para que cumpla esta condición.

Ejemplos de aplicación

LDA es una técnica efectiva de subespacios puesto que optimiza el resultado de Fisher. Estas buenas capacidades han motivado su empleabilidad en aplicaciones de teledetección,

¹⁴ Una matriz no singular es una matriz cuadrada cuyo determinante es diferente de cero.

principalmente centradas en la clasificación de imágenes y la selección de bandas. Adicionalmente, ha sido posible realizar una investigación geobotánica basada en el discriminante lineal, en la que se aprovechaba las capacidades de LDA, concretamente en el análisis de perfiles. Los métodos clásicos de LDA han sido explotados, haciendo uso de datos hiperespectrales, en el reconocimiento de especies coníferas [75] así como en la clasificación de especies de árboles de bosques tropicales húmedos [76]. Otros ejemplos de uso pueden encontrarse en segmentación de imágenes y análisis discriminante para la identificación de unidades cartográficas.

3.5.2.4. Markov Random Fields (MRFs)

Las técnicas MRF también conocidas como *Redes de Markov (Markov Networks)* o modelos gráficos no dirigidos, tratan de modelar, de una manera general, la probabilidad conjunta de un grupo de variables aleatorias, $X = \{X_1, \dots, X_n\}$, proporcionando una representación compacta de la distribución en términos de *potenciales o factores locales*, que son definidos sobre subconjuntos de variables. Estas funciones potenciales son definidas como, $\pi_c: x_c \rightarrow \mathbb{R}$, y pueden ser vistas como representación de preferencias sobre configuraciones locales (no confundir con probabilidades marginales). Tal representación compacta se consigue factorizando la distribución conjunta en el producto de los factores locales. Por supuesto, no todas las distribuciones pueden ser factorizadas de esta manera, sin embargo, esta familia es aún muy expresiva. Un campo aleatorio de Markov está fuertemente relacionado a una gráfica no dirigida $G = (V, E)$ donde cada vértice $v \in V$ se asocia a una variable aleatoria $X_i \in X$ mientras que cada factor, π , con el *clique*¹⁵ mayor, c , en G [77].

Fundamentos teóricos de MRF

- Un campo aleatorio de Markov se define formalmente como:
- Un conjunto de emplazamientos $S = \{1, \dots, N\}$. Se corresponden a N localizaciones de píxeles.
- Un conjunto de variables aleatorias $\{w_n\}_{n=1}^N$ asociadas a cada uno de los emplazamientos.
- Un conjunto de vecinos $\{N_n\}_{n=1}^N$ de cada uno de los N emplazamientos.

¹⁵ Un clique es un subgrafo en que cada vértice está conectado a cada otro vértice del subgrafo, es decir, todos los vértices del subgrafo son adyacentes.

Para que sea un MRF, el modelo debe obedecer a la propiedad de Markov:

$$Pr(w_n | w_{S \setminus n}) = Pr(w_n | w_{Nn}) \quad (3.19)$$

En otras palabras, el modelo debe ser condicionalmente independiente de todas las demás variables dadas, sus vecinos. Es como funciona, exactamente, la independencia condicional en un modelo gráfico no dirigido. Consecuentemente, un MRF puede considerarse como un modelo no dirigido descrito por:

$$P(X = x) = \frac{1}{Z} \prod_c \pi_c(x_c) \quad (3.20)$$

Donde $\pi_c(x_c)$ es la c^a función potencial y devuelve siempre un valor no negativo. Este valor depende del estado del subconjunto de variables $c_j \subset \{1, \dots, N\}$. En este contexto, este subconjunto se conoce como clique.

$$Z = \sum_{x \in \Omega} \prod_c \pi_c(x_c) \quad (3.21)$$

Donde Ω es el conjunto de asignaciones legales a \mathcal{X} . Esta constante asegura que el modelo describa una distribución de probabilidad legal, es decir, que la suma de todas sea 1. Una distribución de esta forma se denomina *distribución de Gibbs*.

Se puede usar información *a priori* aplicando un modelo *HMMRF (Hidden Markov Random Field)*. Haciendo uso de este modelo se introducen restricciones espaciales basadas en *vóxeles*¹⁶ vecinos de un cubo de 3 x 3 x 3. El vóxel central tiene 26 vecinos y se puede calcular la energía MRF contando el número de vecinos. Se espera que los vóxeles vecinos tengan las mismas etiquetas de clase. La probabilidad *a priori* de la clase y la probabilidad de la observación se combina para estimar el *Máximo A Posteriori (MAP)*. La probabilidad *a priori* puede ser entre 0, no HMMRF; y 1, máximo HMMRF para datos con mucho ruido; para cubrir diferentes niveles de ruido. Partes de esta versión están basadas en una versión de *GHMMRF (Gaussian Hidden Markov Random Field)* [78].

¹⁶ El vóxel (del inglés *volumetric pixel*) es la unidad cúbica que compone un objeto tridimensional. Constituye la unidad mínima procesable de una matriz tridimensional y es, por tanto, el equivalente del píxel en un objeto 2D.

La energía, U_x , se calcula contando el número de vóxeles vecinos de una clase de tejido en un cubo de $3 \times 3 \times 3$. En la Figura 3.23 se ilustran algunas configuraciones que van desde el $U_x=0$ (sin vecinos) hasta $U_x=-26$ (número máximo de vecinos, 26).

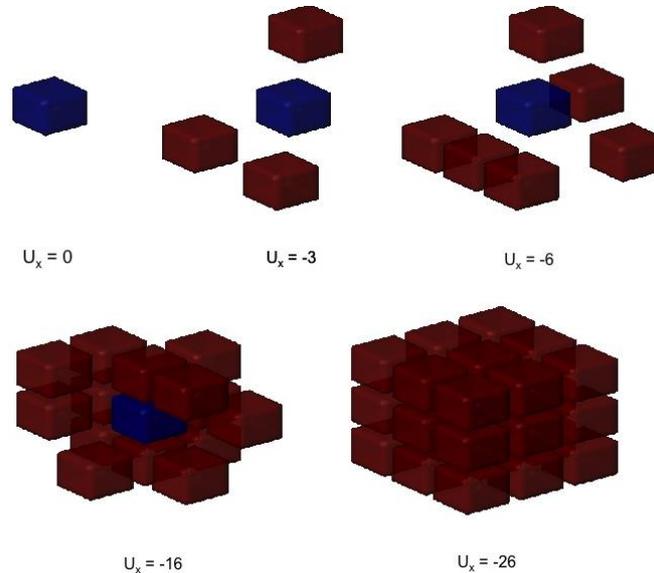


Figura 3.23: Cálculo de la energía de un vóxel

La idea es ir eliminando vóxeles aislados de una clase de tejido, de los cuales es improbable pertenecer al mismo tipo de tejido. Además esto deja huecos en un cluster de vóxeles conectados de un tipo de tejido. En la segmentación resultante el nivel de ruido será minimizado.

Ejemplos de aplicación

Estos modelos fueron introducidos primeramente en el campo de las mecánicas estadísticas para modelar determinados fenómenos físicos. Sin embargo, estos algoritmos se aplican hoy en día a un gran número disciplinas del conocimiento, en especial en medicina, biomedicina y robótica; proporcionan una manera conveniente y constante de modelar entidades dependientes del contexto, tales como píxeles de la imagen y otras características espaciales correlacionadas. Pueden encontrarse en visión por computadora, lenguaje natural, biología computacional y comunicaciones digitales.

La mayoría de los problemas en visión artificial se consideran de etiquetado, usando restricciones, debido al conocimiento *a priori* y a las observaciones. Los estudios basados en MRFs han tenido éxito al modelar problemas de visión de bajo nivel, como restauración

de imágenes, segmentación, reconstrucción de superficies, análisis de texturas, flujo óptico, integración visual y detección de bordes [79].

3.5.2.5. Random Forest (RF)

Random Forest o Random Forests (*bosques aleatorios*) es un conjunto de métodos de aprendizaje para clasificación, regresión y otras tareas, que opera construyendo una multitud de árboles de decisión en tiempo de entrenamiento y produciendo la clase, *moda*¹⁷ de las clases (clasificación) o la predicción media (regresión) de los árboles individuales. Random Forest corrige el hábito de los árboles de decisión, de *overfitting*¹⁸ en su conjunto de entrenamiento [80].

El algoritmo que dio lugar a RF fue desarrollado por Leo Breiman y Adele Cutler, y RF es su marca comercial. El método combina la idea “*bagging*” de Breiman y la selección aleatoria de características, al principio introducida por Ho, y más tarde, independientemente, por Amit y German para construir una colección de árboles de decisión con varianza controlada.

La selección de un conjunto aleatorio de características es un ejemplo del método de subespacio aleatorio, el cual, en la formulación de Ho, es una manera de implementar la clasificación propuesta por Eugene Kleinberg.

Características y ventajas de Random Forests:

- Es uno de los algoritmos de aprendizaje más certeros que hay disponible. Para un set de datos lo suficientemente grande produce un clasificador muy certero.
- Se ejecuta eficientemente en bases de datos grandes.
- Puede manejar cientos de variables de entrada sin excluir ninguna.
- Aporta estimaciones sobre cuáles son las variables importantes en la clasificación.
- Genera una estimación interna imparcial del error general a medida que la construcción del bosque progresa.
- Tiene un método eficaz para estimar datos perdidos y mantener la precisión cuando una gran proporción de los datos está perdida.
- Tiene métodos para balancear el error en conjuntos de datos no balanceados en la población de clase.

¹⁷ En estadística, la moda es el valor con mayor frecuencia en una distribución de datos.

¹⁸ Overfitting es el efecto de sobreentrenar un algoritmo de aprendizaje.

- Los bosques generados pueden ser guardados para usos futuros sobre otros datos.
- Computa los prototipos que dan información sobre la relación entre las variables y la clasificación.
- Computa las proximidades entre los pares de muestras que pueden usarse en *clustering*, localizando *outliers*¹⁹ (valores atípicos), o (escalando) aportando vistas interesantes de los datos.
- Las capacidades anteriores pueden ser extendidas a conjuntos de datos no etiquetados, destacando *clustering* no supervisado, vistas de datos y detección de *outliers*.
- Ofrece un método experimental para detectar las interacciones de las variables.

Por otro lado, RF no produce *overfitting*, se pueden emplear tantos árboles como se quiera de manera rápida. Ejecutar un conjunto de datos de 50000 muestras y 100 variables, empleando 100 árboles en una máquina de 800 Mhz, conlleva un tiempo de computación de 11 minutos. Para conjuntos de datos grandes, el mayor requisito de memoria lo imponen los mismos datos y tres *arrays* de enteros de la misma dimensión de los datos.

Fundamentos teóricos de RF

En Random Forests se construyen muchos árboles de decisión, luego, para clasificar cada nuevo objeto desde un vector de entrada, se introduce el vector a cada uno de los árboles, seguidamente se obtiene una clasificación de cada árbol, se dice entonces que cada árbol a “*votado*” por una clase. Por último, el bosque proporciona la clasificación según el más votado (de los resultados ofrecidos por todos los árboles de dicha bosque) [81].

Cada árbol se construye usando el siguiente algoritmo:

1. Si el número de muestras del conjunto de entrenamiento es N , se escogerán N muestras aleatoriamente (seleccionadas con reemplazo) del conjunto de datos original. Estas muestras conformarán el conjunto de entrenamiento para la construcción del árbol.
2. Sea M el número de variables totales, se especifica un número m , tal que $m \ll M$, de variables para cada nodo. Las m variables son escogidas aleatoriamente de las M variables, y la mejor partición de estas m son empleadas para dividir el nodo. El valor m se mantiene en la construcción del bosque.

¹⁹ En estadística, son observaciones que numéricamente distan del resto de los datos.

3. Cada árbol se construye lo más grande posible sin que haya ninguna *poda*.

La tasa de error del bosque depende de dos cosas:

- La correlación entre dos árboles, de todos los árboles del bosque. A mayor correlación mayor tasa de error del bosque.
- La *fuerza* de cada árbol individual en el bosque. Un árbol con una tasa de error baja es un clasificador *fuerte*. A mayor fuerza de los árboles individuales, menor es la tasa de error del bosque.

Al reducir m se reduce tanto la correlación como la fuerza, mientras que al aumentar m se incrementan ambas. Existe un valor óptimo de m del posible rango de valores, que suele ser bastante amplio. Se puede obtener un valor de m rápidamente haciendo uso de la tasa de error oob^{20} (*out-of-data*). Este es el único parámetro ajustable al que, de alguna manera, el bosque es sensible.

Para entender y usar las diferentes opciones, es útil conocer de qué manera se computa la información. La mayoría de las opciones depende de dos objetos generados por RF.

Al seleccionarse aleatoriamente con reemplazo, el conjunto de entrenamiento, aproximadamente una tercera parte de los datos quedan fuera del conjunto. Estos datos *oob* son usados para obtener la estimación parcial del error de clasificación a medida que se añaden árboles al bosque. También suele obtenerse una estimación de la importancia de la variable.

Después de que se construye cada árbol, se le introducen todos los datos y se computan las proximidades para cada par de muestras. Si dos muestras ocupan el mismo nodo terminal, su proximidad se incrementa en 1. Al final de la ejecución, las proximidades son normalizadas dividiéndolas por el número de árboles. Las proximidades son usadas en el reemplazamiento de datos perdidos, para localizar *outliers* y para producir vistas de baja dimensionalidad de los datos. En la Figura 3.24 se muestra un ejemplo de una clasificación realizada por un árbol de decisión de RF en una aplicación de detección remota.

²⁰ Out-of-bag hace referencia a los datos que se quedan fuera del conjunto de entrenamiento por el hecho de seleccionar las muestras aleatoriamente y con reemplazo.

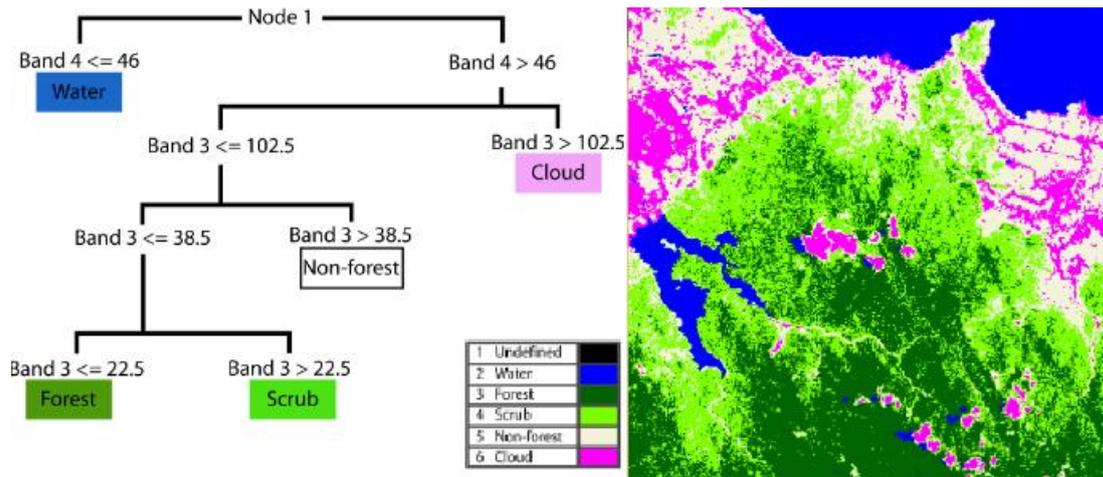


Figura 3.24: Ejemplo de clasificación realizada por un árbol de decisión en teledetección.

Limitaciones

Se han encontrado problemas de *overfitting* empleando determinados grupos de datos, en tareas de regresión y clasificación ruidosas.

Para los datos que incluyen variables categóricas con diferente número de niveles, RF se parcializa a favor de esos atributos con más niveles. Por consiguiente, la posición que marca la variable no es fiable para este tipo de datos. Métodos como las permutaciones parciales se han usado para resolver este tipo de problemas.

Si los datos contienen grupos de atributos correlacionados con similar relevancia para el rendimiento, entonces los grupos más pequeños están favorecidos sobre los grupos más grandes [82].

Ejemplos de aplicación

Random Forest se ha aplicado a exitosamente a varios problemas en biología computacional. La popularidad de RF en este campo surge porque RF puede ser aplicado a un amplio rango de tipos de datos, incluso si los problemas son no lineales o involucran interacciones de elevado orden de complejidad. RF y sus variantes se han empleado en variedad de problemas de bioinformática, tales como: clasificación de expresión genética, análisis de datos de proteómica, basado en espectrometría de masas, descubrimiento de biomarcadores, secuencia de proteínas, predicción de interacciones proteína-proteína, o genética estadística [83].

Adicionalmente, ha sido probado en muchos otros dominios biomédicos. Por ejemplo, RF ha demostrado ser un clasificador estadístico potente en ecología computacional. Cutler comparó las precisiones de RF y otros cuatro clasificadores estadísticos habitualmente usados, en tres conjuntos diferentes de datos de tipo ecológico para describir: la presencia de especies invasivas de plantas en US California; la presencia de especies extrañas de líquen en el Pacífico noroccidental de los EEUU; y los lugares de anidación de aves en Utah. En estas tres aplicaciones RF presentó altas precisiones de clasificación.

Otra aplicación interesante es la prueba computacional de fármacos, donde los paneles de cultivos celulares son usados para testear los fármacos candidatos, por su capacidad de inhibir la proliferación. Riddick construyó modelos de regresión empleando RF para predecir la respuesta al fármaco de 19 cultivos celulares de cáncer de mama y 7 de gliomas. Se usó de tres formas: (1) selección de característica de las firmas de expresión genética del fármaco basado en la importancia de permutación, (2) eliminación de *outliers* de cultivos celulares basado en proximidad, y (3) modelo de regresión multivariado para predicción de respuesta continua del fármaco.

Se pueden encontrar más aplicaciones de RF en otros campos diferentes como el modelado relacional de estructura-actividad cuantitativo, espectroscopia en resonancia magnética nuclear o soporte de decisión clínica en medicina general.

3.5.2.6. Artificial Neural Networks (ANNs)

Una red neuronal es un sistema computacional “*conexionista*”. Los sistemas computacionales que escribimos son procedimentales; un programa comienza con una primera línea de código, la ejecuta, y va a la siguiente, siguiendo instrucciones de una manera lineal. Sin embargo, una red neuronal verdadera no sigue una ruta lineal, sino que la información se procesa colectivamente, en paralelo a través de los nodos de la red (en este caso los nodos son las neuronas) [84].

En la Figura 3.25 se muestra un ejemplo de un sistema complejo. Los elementos individuales de la red, las neuronas, son simples. Cada una de ellas lee una entrada, la procesa y genera una salida. Por el contrario, una red de muchas neuronas puede presentar comportamientos increíblemente ricos e inteligentes.

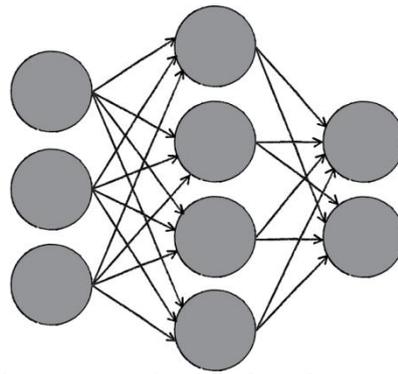


Figura 3.25: Ejemplo de red neuronal.

Uno de los elementos claves en una red neuronal es su capacidad de “aprender”. Una red neuronal no es solamente un sistema complejo sino un sistema complejo *adaptativo*, es decir, que puede cambiar su estructura interna basándose en la información que fluye a través de él. Normalmente esto se consigue haciendo un ajuste de *pesos*. En la Figura 3.25, cada línea representa una conexión entre dos neuronas e indica el camino por el que fluye la información. Cada conexión tiene un peso, *weight* (un número que controla la señal entre las dos neuronas). Si la red genera una “buena” salida, no hay necesidad de ajustar los pesos. Sin embargo, si la red genera una salida “pobre”, un error por así decirlo, entonces el sistema el sistema se adapta alterando los pesos para mejorar los resultados subsecuentes.

En *machine learning* y ciencia cognitiva²¹, las redes neuronales artificiales son una familia de algoritmos estadísticos de aprendizaje inspirados en redes neuronales biológicas (el SNC de los animales, en particular el cerebro) y se usan para estimar o aproximar funciones que pueden depender de un número alto de entradas y que son desconocidas generalmente. Las redes neuronales artificiales son generalmente presentadas como sistemas de “neuronas” interconectadas las cuales pueden computar valores desde las entradas, y son capaces de aprendizaje automático así como de reconocimiento de patrones gracias a su naturaleza adaptativa [85].

Fundamentos teóricos de ANN

Una neurona simple

Una neurona artificial es un dispositivo con varias entradas y una salida. La neurona tiene dos modos de operación; el modo de entrenamiento y el modo de uso. En el modo de

²¹ La ciencia cognitiva es el estudio interdisciplinario de cómo la información es representada y transformada en la mente/cerebro.

entrenamiento, la neurona puede entrenarse para “activarse” (*to fire*) o no, para determinados patrones de entrada. En el modo de uso, cuando un patrón de aprendizaje es detectado a la entrada, su salida asociada se obtiene a la salida de la neurona. Si el patrón de entrada no pertenece a la lista de aprendizaje de patrones de entrada, la “regla de activación” (*firing rule*) se usa para determinar si la neurona se “activa” o no [86].

En la Figura 3.26 puede observarse un modelo de una neurona artificial de n entradas.

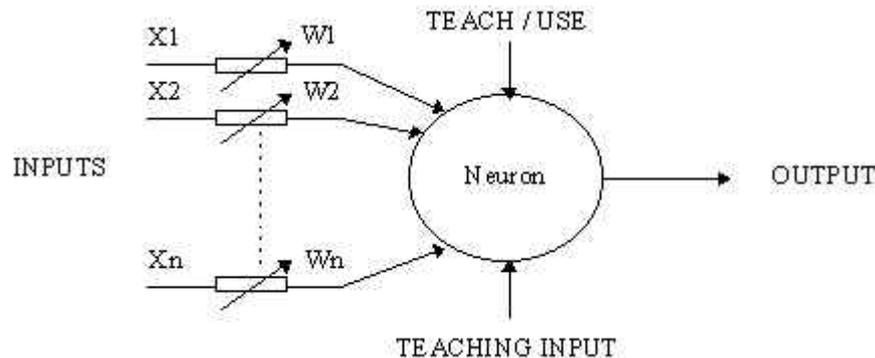


Figura 3.26: Modelo de neurona de McCulloch y Pitts (MCP).

Firing rules

Una *firing rule* determina como se calcula si una neurona debería “activarse” (*fire*) para cualquier patrón de entrada. Esto se aplica a todos los patrones de entrada, no solamente a aquellos que se utilizaron en el entrenamiento del nodo.

Una *firing rule* sencilla puede ser implementada usando la técnica de distancia de Hamming. La regla se explica a continuación:

Se toma una colección de patrones de entrenamiento para un nodo; algunos de los cuales “activa” (*fire*) el nodo (se genera “1” para este conjunto de patrones), y los otros patrones de entrenamiento evitan que se “active” (se genera “0” para este otro conjunto de patrones). Entonces, los patrones que no pertenecen a esta colección producirán que se “active” si, en comparación, tienen más elementos de entrada en común con el patrón “más cercano” del conjunto que genera “1” que con el patrón “más cercano” del conjunto que genera “0”. Si se produce empate, el patrón pertenece en estado indefinido.

En el modelo de neurona MCP (Figura 3.26), las entradas tienen un “peso” asignado, por lo que el efecto que cada entrada tiene en la decisión que se genera es dependiente del peso de esa entrada en particular. El peso de una entrada es un número que se multiplica a una entrada dada proporcionando que dicha entrada esté dotada de dicho peso. Estas

entradas con “mayor peso” luego son sumadas, y si exceden un valor umbral establecido, T , la neurona se “activa”. En cualquier otro caso la neurona no se “activa”. En términos matemáticos, la neurona se “activa”, si y solo si: $X_1W_1 + X_2W_2 + X_3W_3 + \dots > T$.

La suma de estas entradas con pesos asociados y el valor umbral, hacen que esta neurona sea muy flexible y potente. La neurona MCP tiene la capacidad de adaptarse a una situación particular cambiando sus pesos y/o su umbral. Existen varios algoritmos que provocan que la neurona se “adapte”; los más usados son *delta rule* y *back error propagation*. El primero se usa en redes *feed-forward* mientras que el último en redes *feedback*.

Redes feed-forward

Las redes neuronales artificiales *feed-forward* (Figura 3.27) permiten que las señales viajen en un único sentido; desde las entradas a las salidas. No hay bucles de realimentación (*feedback*), es decir, la salida de cualquiera de las capas no afecta a esa capa. Las ANNs *feed-forward* tienden a ser redes directas que asocian entradas a salidas. Son ampliamente usadas en reconocimiento de patrones. Este tipo de organización es también referida como *bottom-up* o *top-down*.

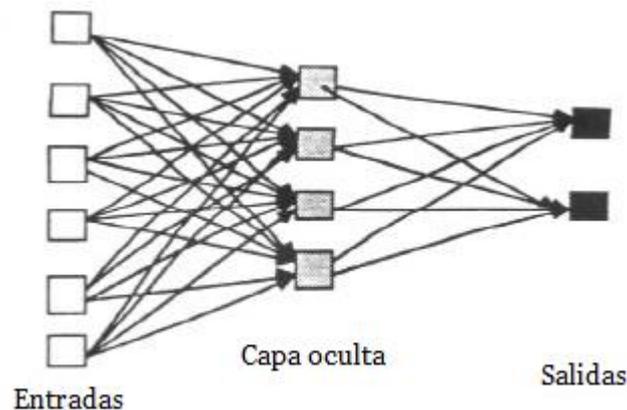


Figura 3.27: Ejemplo de una red sencilla *feed-forward*.

Redes feed-back

Las redes *feed-back* (Figura 3.28) pueden tener señales viajando en ambas direcciones produciendo bucles en la red, son muy potentes y pueden ser extremadamente complicadas. Estas redes son dinámicas; su “estado” está cambiando continuamente hasta que alcanzan un punto de equilibrio, permanecen en el punto de equilibrio hasta que las entradas cambian y cuando esto ocurre, necesitan volver a un nuevo punto de equilibrio.

Las arquitecturas *feedback* son también denominadas como interactivas o recurrentes, aunque el último término se usa a menudo para denotar conexiones *feedback* en organizaciones de una sola capa (*single-layer*).

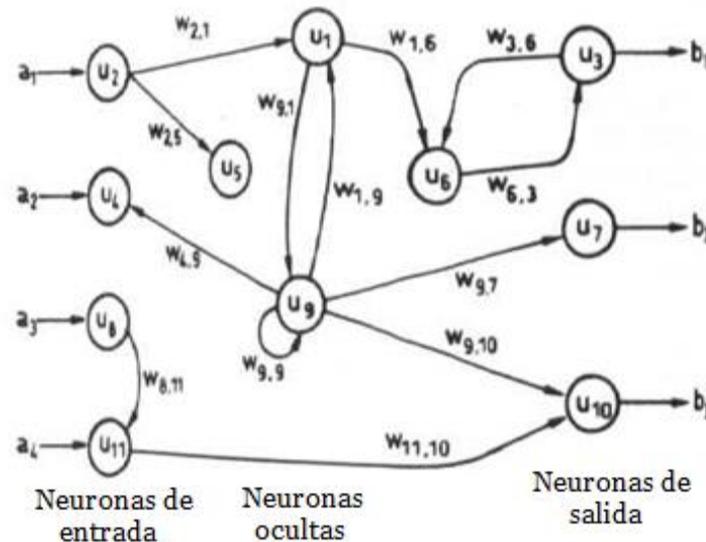


Figura 3.28: Ejemplo de una red *feed-back*.

Capas de las redes neuronales

El tipo más común de una red neuronal artificial consta de tres grupos, capas o niveles, de unidades: una capa de unidades de entrada (*input*) se conecta a una capa de unidades ocultas (*hidden*), las cuales se conectan a una capa de unidades de salida (Figura 3.27).

- La actividad de las unidades de entrada representa la información en bruto que es suministrada al interior de la red.
- La actividad de cada unidad oculta es determinada por la actividad de las unidades de entrada y los pesos en las conexiones entre las unidades de entrada y las unidades ocultas.
- El comportamiento de las unidades de salida dependen de la actividad de las unidades ocultas y los pesos entre las unidades ocultas y las unidades de salida.

Este tipo simple de redes es interesante porque las unidades ocultas son libres de construir sus propias representaciones de la entrada. Los pesos entre las unidades de entradas y las unidades ocultas determinan cuando se activa cada unidad oculta, y por tanto, una unidad oculta puede elegir que representa al modificar estos pesos.

Se puede distinguir entre arquitecturas de capa única y multicapa; la organización de una sola capa, en la cual todas las unidades se conectan a otra unidad, constituye el caso más general y presenta mayor potencial computacional que las jerárquicamente estructuradas en organizaciones multicapa. En las redes multicapa, las unidades son a menudo numeradas por capa, en vez de seguir una numeración global.

Perceptrón

El trabajo más influyente sobre redes neuronales en los años sesenta tenía el título ‘perceptrons’, un término acuñado por Frank Rosenblatt. El *perceptrón* (Figura 3.29) se trata de un modelo MCP con un ajuste de preprocesado adicional. Las unidades etiquetadas como: $A_1, A_2, \dots, A_j, \dots, A_p$ son llamadas unidades de asociación y su tarea es extraer características específicas de las imágenes de entrada. El *perceptrón* imita la idea básica que hay tras el sistema visual mamífero. Fue utilizado principalmente en reconocimiento de patrones pero hoy en día su capacidad se extiende a muchas otras áreas.

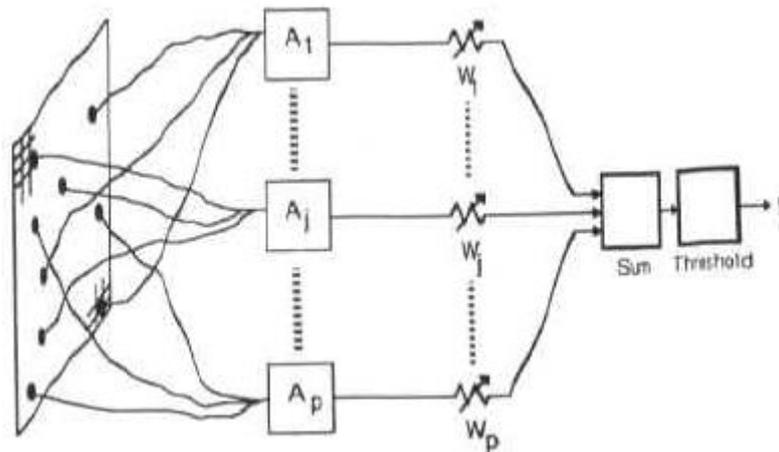


Figura 3.29: *Perceptrón*.

El proceso de aprendizaje

La memorización de patrones y la respuesta subsiguiente de la red neuronal puede ser categorizada en dos paradigmas generales:

- **Associative mapping** (*memoria asociativa*) en el que las redes neuronales aprenden a producir un patrón determinado en el conjunto de unidades de salida cuando se aplica otro patrón particular en el conjunto de unidades de

entrada. La memoria asociativa puede desglosarse generalmente en dos mecanismos:

- *Autoasociación*: un patrón de entrada se asocia con sí mismo y los estados de las unidades de entrada y salida coinciden. Se usa para proporcionar terminación de patrones, es decir, para producir un patrón cuando se presenta una porción de este o un patrón distorcionado. En el segundo caso, la red realmente almacena pares de patrones construyendo una asociación entre dos conjuntos de patrones.
- *Heteroasociación*: está relacionado con dos mecanismos de memorización.
 - *Memorización del vecino más cercano (nearest-neighbour recall)*, donde el patrón de salida producido se corresponde al patrón de entrada almacenado, el que es el más cercano al patrón presentado.
 - *Memorización interpolativa (interpolative recall)*, donde el patrón de salida es una interpolación dependiente de la similaridad, de los patrones almacenados correspondientes al patrón presentado. Otro paradigma, variante de *associative mapping*, es la clasificación, es decir, cuando se dispone de un conjunto fijado de categorías en el que los patrones de entrada deben ser clasificados.
- **Regularity detection** (*detección de regularidad*) en la cual las unidades aprenden a responder a propiedades particulares de los patrones de entrada. Mientras que en *associative mapping* las redes neuronales almacenan las relaciones entre patrones, en *regularity detection*, la respuesta tienen un particular “significado”. Este tipo de mecanismo de aprendizaje es esencial para el descubrimiento de características y la representación del conocimiento.

Toda red neuronal posee un conocimiento contenido en los valores de los pesos de las conexiones. Modificar el conocimiento almacenado en la red como una función de experiencia, implica una regla de aprendizaje para cambiar los valores de los pesos.

La información se almacena en la matriz de pesos, W , de una red neuronal. El aprendizaje es la determinación de los pesos. Continuando con la manera en que se produce el aprendizaje, se puede distinguir entre dos categorías generales de redes neuronales:

- **Redes neuronales fijas** en las que los pesos no pueden ser cambiados, es decir: $\frac{dW}{dt} = 0$. En estas redes neuronales los pesos son fijados *a priori* de acuerdo al problema a resolver.
- **Redes neuronales adaptativas** las cuales son capaces de cambiar sus pesos, en este caso: $\frac{dW}{dt} \neq 0$.

Función de transferencia

El comportamiento de una ANN depende de los pesos y de la función entrada-salida (función de transferencia), que se especifica para las unidades. Normalmente esta función entra en una de estas tres categorías:

- **Lineal (o rampa).** Para las unidades lineales, la actividad de salida es proporcional a la salida total con peso.
- **Umbral.** Para estas unidades la salida se establece en uno de dos niveles, dependiendo de si la entrada total es mayor o menor que un determinado valor umbral.
- **Sigmoidal.** En estas unidades la salida varía continuamente pero no linealmente según cambia la entrada. Las unidades *sigmoidales* presentan mayor parecido a las neuronas reales que las unidades umbrales y lineales, no obstante, los tres tipos deben ser considerados aproximaciones a grandes rasgos.

Para hacer que una red neuronal realice algunas tareas específicas, debe decidirse como se conectan las unidades unas a otras, y deben establecerse los pesos apropiadamente en las conexiones. Las conexiones determinan si es posible que una unidad influya a otra. Los pesos especifican la fuerza de tal influencia.

Se puede “enseñar” a una red neuronal de tres capas a que realice una determinada tarea, empleando el procedimiento siguiente:

1. Se presenta una red con ejemplos de entrenamiento, que consisten de un patrón de actividades para las unidades de entrada, junto al patrón deseado de actividades para las unidades de salida.
2. Se determina como de cerca queda la salida real de la red neuronal en relación a la deseada.

3. Se cambia el peso de cada conexión de manera que la red produzca una mejor aproximación de la salida deseada.

Limitaciones

Existen muchas ventajas y limitaciones con respecto al análisis en redes neuronales, con lo que para discutir este tema correctamente se debería tratar individualmente los diferentes tipos de redes neuronales. Sin embargo, en referencia a redes *backpropagation* (propagación hacia atrás de errores) hay algunos asuntos específicos sobre los que deberían ser conscientes los usuarios potenciales.

Las redes neuronales *backpropagation*, así como muchos otros tipos de redes, son, en cierto sentido, “cajas negras”. Aparte de definir la arquitectura de una red y tal vez inicializarla con números aleatorios, el usuario no tiene rol que el de asignar la entrada, verla entrenar y esperar la salida. De hecho, se suele decir que con *backpropagation*, “casi no sabes lo que estás haciendo”. Algunos paquetes de software libre disponibles, tales como: NevProp, BP, Mactivation, permiten al usuario tomar muestras en intervalos de tiempo regulares del “progreso” de la red, pero el progreso de aprendizaje en sí mismo se realiza sin ayuda. El producto final de esta actividad es una red entrenada que proporciona una relación, como en regresión, sin ecuaciones o coeficientes definidos, al margen de sus propias matemáticas internas. La red es en sí misma la “ecuación final” de dicha relación.

Las redes neuronales *backpropagation* también tienden a presentar un entrenamiento más lento que otros tipos de redes, y algunas veces, requieren cientos de épocas²². Si se ejecuta sobre un sistema computador verdaderamente paralelo, este asunto no presenta un problema importante, sin embargo, si la red neuronal *backpropagation* está siendo simulada sobre un sistema serie estándar, el entrenamiento puede llevarse a cabo en un tiempo considerable. Esto ocurre porque las máquinas *CPU* deben computar la función de cada nodo y conexión separadamente, lo cual puede ser problemático en redes muy grandes con gran cantidad de datos [87].

Ejemplos de aplicación

Las ANNs cubren un área amplia de aplicaciones en el mundo real. Esta tecnología se aplica satisfactoriamente en multitud de industrias debido a que ostentan un gran potencial

²² Una época se define como el ajuste de los pesos para todos los pares de entrenamientos. Normalmente la red requiere muchas épocas para su entrenamiento.

en el reconocimiento de patrones. Los buenos resultados que manifiestan en predicción y pronóstico han hecho que se utilicen en aplicaciones tales como: pronóstico de ventas, control de procesos industriales, análisis de clientes, validación de datos, gestión de riesgos o target marketing.

Por último, las redes neuronales artificiales también son usadas en modelos específicos: reconocimiento de voz en comunicaciones, diagnóstico de hepatitis, recuperación de telecomunicaciones de software dañinos, interpretación de palabras chinas multi-significado, detección de minas submarinas, análisis de texturas, reconocimiento de objetos tridimensionales, reconocimiento de palabras escritas a mano y reconocimiento facial.

3.5.2.7. Support Vectors Machines (SVMs)

A comienzos del siglo XXI se comprobó la gran eficacia de los métodos basados en la teoría del aprendizaje estadístico al trabajar con ambos problemas: alta dimensionalidad y escaso conjunto de entrenamiento. El entrenamiento de los clasificadores, tanto estadístico como de redes neuronales, hacen uso del principio de *Minimización de la Riqueza Empírica (Empirical Risk Minimization, ERM)*, que consiste en permitir la minimización de la tasa de error para un conjunto de entrenamiento dado. El problema viene cuando hay que extender o generalizar esa clasificación al resto de objetos, entonces no se consigue una buena actuación, es decir, el resultado es una tasa de error más alta que para el conjunto de entrenamiento [26].

La máquina de vectores soporte (*Support Vector Machines, SVMs*), es un método supervisado de reconocimiento de patrones introducido recientemente en el marco de la Teoría del Aprendizaje de Vladimir Vapnik y su equipo en los laboratorios AT&T [88]. Combina las siguientes ideas: la técnica de búsqueda de hiperplanos óptimos como solución, la idea de convolución de producto escalar, la extensión de las funciones lineales a no lineales y la noción de *margen ligero* o *soft margin* para permitir errores en los patrones de entrenamiento. Una ventaja importante de este algoritmo es que trabaja con el principio de *Minimización del Riesgo Estructural (Structural Risk Minimization, SRM)*, el cual es mejor que el ERM como hacen muchas otras técnicas. SVM permite entonces una mejor generalización, antes que una mejor clasificación del conjunto de entrenamiento, a nivel de errores.

Existen otras dos razones más que han hecho crecer el interés por este novedoso clasificador. SVM puede ser reducido a un problema de *programación cuadrática convexa*

(*Convex Quadratic Programming, CQP*), más fácil de resolver con respecto a los métodos clásicos y que parecen tener un mejor comportamiento (más robusto) con altas cantidades de datos.

Fundamentos teóricos de SVM

Este clasificador pertenece a la familia de los clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad, a pesar de que se pueden adaptar de forma sencilla para actuar como clasificadores no lineales mediante la aplicación de una función o *kernel* no lineal sobre los datos de entrada. Su principal objetivo es obtener una superficie, o hiperplano, capaz de separar las diferentes clases en las que se puede agrupar una distribución de datos en un espacio N-dimensional, utilizando para ello un proceso de optimización basado en la obtención de vectores que definen los límites de las clases. Estos vectores se denominan normalmente *vectores soporte (support vectors)* [89]. Si vemos los datos de entrada como dos conjuntos de vectores en un espacio N-dimensional, el objetivo del algoritmo SVM simplemente es construir un hiperplano de separación en ese espacio, el cual maximice el margen de distancia a los dos conjuntos de datos [90].

En la Figura 3.30 se puede apreciar cómo se calcula este hiperplano de separación construyendo otros dos hiperplanos paralelos, uno a cada lado del primero.

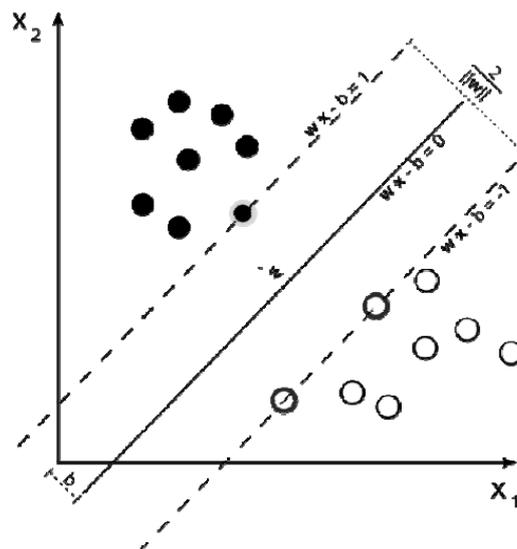


Figura 3.30: Esquema de funcionamiento del clasificador SVM.

Los dos hiperplanos paralelos son empujados o ensanchados, para aproximarse lo más posible a los conjuntos de datos. Intuitivamente se alcanza una buena separación

cuando el hiperplano de separación se encuentra a la mayor distancia de ambas clases. Cuanto mayor sea la distancia menor será en general el error del clasificador. En términos matemáticos, dado un conjunto de entrenamiento:

$$D = \{(x_i, c_i) \mid x_i \in \mathbb{R}^p, \quad c_i \in \{-1, 1\}\}_{i=1}^n \quad (3.22)$$

Donde c_i es 1 ó -1, indicando la clase a la que el x_i pertenece (cada x_i es un vector real p -dimensional). Se busca obtener un hiperplano de distancia máxima a los conjuntos de entrenamiento, y que además divida a aquellas muestras pertenecientes a $c_i = 1$ de aquellas que pertenezcan a $c_i = -1$. Cualquier hiperplano puede ser escrito como un conjunto de puntos x que satisfaga la ecuación (3.23):

$$w \cdot x - b = 0 \quad (3.23)$$

Donde el vector w es un vector normal perpendicular al hiperplano y el parámetro $\frac{b}{\|w\|}$ determina el desplazamiento del hiperplano sobre el origen. Se puede elegir la w y la b que maximice la distancia entre los dos hiperplanos paralelos, que estarán tan apartados como sea posible en función de los datos. Estos hiperplanos pueden ser descritos con las siguientes fórmulas:

$$w \cdot x - b = 1 \quad (3.24)$$

$$w \cdot x - b = -1 \quad (3.25)$$

Nótese que si el conjunto de entrenamiento es linealmente separable, se pueden elegir dos hiperplanos en el borde de los conjuntos de modo que no haya puntos entre ellos, y entonces tratar de maximizar la distancia entre estos dos hiperplanos. Haciendo uso de la geometría, se puede encontrar que la distancia entre ellos es $\frac{2}{\|w\|}$, por lo que se pretende minimizar $\|w\|$. Como tenemos que evitar que los puntos se sitúen en la zona límite, añadimos la restricción de la ecuación (3.26) a los x_i pertenecientes a la primera clase y la restricción de la ecuación (3.27) a los x_i pertenecientes a la segunda:

$$w \cdot x_i - b \geq 1 \quad (3.26)$$

$$w \cdot x_i - b \leq -1 \quad (3.27)$$

Esto puede ser escrito como:

$$c_i(w \cdot x_i - b) \geq 1, \forall 1 \leq i \leq n \quad (3.28)$$

Se puede compactar la expresión para llegar al problema de optimización, esto es: elegir w y b para minimizar $\|w\|$ sujeto a la restricción de la ecuación (3.28). Este problema de optimización manifiesta una cierta complejidad debido a que solamente depende de un valor $|w|$, la razón de su complejidad es que se trata de un problema de optimización no convexo²³. Afortunadamente, es posible sustituir $\|w\|$ por $\frac{1}{2}\|w\|^2$ sin cambiar la solución. Esto es un problema de optimización de programación cuadrática. Más claramente, el problema de optimización puede reformularse de la siguiente manera:

$$\text{minimizar } \frac{1}{2}\|w\|^2, \text{ sujeto a } c_i(w \cdot x_i - b) \geq 1, \forall 1 \leq i \leq n \quad (3.29)$$

El factor $\frac{1}{2}$ se usa como una conveniencia matemática. Ahora el problema que se nos presenta se puede resolver mediante programas y técnicas de programación cuadrática estándar. Escribiendo la regla de clasificación en su forma dual extendida revela que la distancia máxima al hiperplano y, por tanto, la tarea de clasificación, es solo una función de los vectores soporte, es decir, las muestras que están al límite. La segunda forma de SVM se puede derivar como la siguiente expresión:

$$\max \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j c_i c_j x_i^T x_j \quad (3.30)$$

$$\text{Sujeto a } \alpha_i \geq 0 \text{ y } \sum_{i=1}^n \alpha_i c_i = 0 \quad (3.31)$$

Donde los términos α constituyen otra representación del vector de pesos en términos del conjunto de entrenamiento:

$$w = \sum_i \alpha_i c_i x_i \quad (3.32)$$

²³ Un problema de optimización convexo reúne una serie de criterios, condiciones y restricciones matemáticas, estas pueden consultarse en la referencia [72]. Los problemas de optimización no convexos presentan cierta complejidad [119] a la hora de ser resueltos en programación, por ello existen diferentes trabajos orientados a la resolución de este tipo de problemas [120].

El algoritmo original especifica un clasificador lineal, sin embargo, puede modificarse para resolver problemas de clasificación no lineal reemplazando el producto escalar por una función *kernel* no lineal. Esto permite al algoritmo fijar la máxima distancia al hiperplano en un espacio de características transformado. La transformación podría ser no lineal y el espacio transformado de alta dimensionalidad; de este modo, aunque el clasificador es un hiperplano en un espacio de características de alta dimensionalidad, podría ser no lineal en el espacio de entrada original [91]. En la Figura 3.31 se ilustra un ejemplo en el que los datos están definidos en el cuadrado $[-1, 1] \times [-1, 1] \in \mathbb{R}^2$ (situación habitual cuando los datos se corresponden a imágenes en escala de grises) y una posible superficie de decisión tridimensional, \mathbb{R}^3 , no lineal [92].

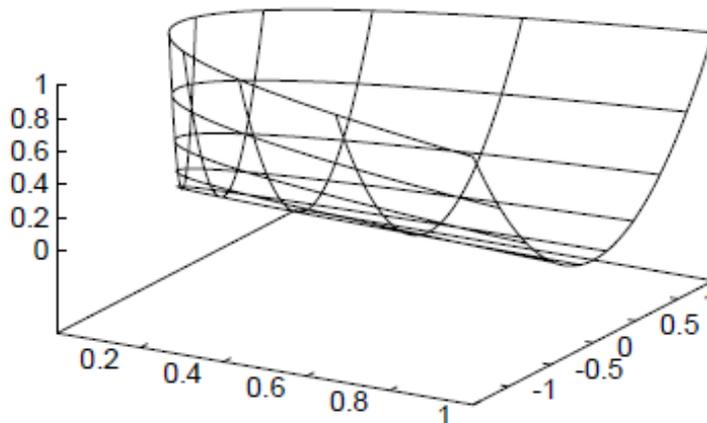


Figura 3.31: Ejemplo de datos definidos en $[-1, 1] \times [-1, 1] \in \mathbb{R}^2$ en el que se ilustra una posible superficie de decisión no lineal.

Si el *kernel* utilizado es de tipo *Gaussian Radial Basis Function*, el espacio de características correspondiente es un espacio de Hilbert de dimensión infinita. La máxima distancia está regulada, por lo que la dimensión infinita no supondrá un problema para obtener los resultados [93]. Algunos de los *kernels* habitualmente utilizados en clasificadores de tipo SVM se enumeran a continuación:

- Polynomial (homogéneo):

$$k(x, x') = (x \cdot x')^d \quad (3.33)$$

- Polynomial (heterogéneo):

$$k(x, x') = (x \cdot x' + 1)^d \quad (3.34)$$

- Radial Basis Function:

$$k(x, x') = e^{-\gamma \|x-x'\|^2}, \text{ para } \gamma > 0 \quad (3.35)$$

- Gaussian Radial Basis Function:

$$k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \quad (3.36)$$

- Sigmoide:

$$k(x, x') = \tan(x \cdot x' + c), \text{ para algunos } k > 0 \text{ y } c < 0 \quad (3.37)$$

En la literatura, podemos encontrar también ejemplos de *kernels* basados en métricas espectrales comúnmente utilizadas en análisis espectral [94]. Como se dijo al inicio, el procesamiento empleado por las SVM no requiere de un gran número de patrones de entrenamiento, siempre y cuando los patrones escogidos sean realmente representativos [95].

Limitaciones

Tal vez, la mayor limitación de SVM reside en la elección del *kernel* que debe utilizarse. No obstante, una vez se ha seleccionado el *kernel* que se va a emplear en la clasificación, se presenta el problema sobre valores que deben asignarse a sus parámetros. Se han realizado algunos trabajos sobre las limitaciones que presentan los *kernels* basados en conocimientos previos, pero la mejor elección del *kernel* para un problema dado, es aún objeto de estudio [92].

Una segunda limitación es la velocidad y el almacenamiento, tanto en el entrenamiento como en la clasificación en fase de test. Mientras que el problema de la velocidad en fase de test se soluciona en [96], aún requiere dos pases del entrenamiento. La fase de entrenamiento para conjuntos de datos muy grandes (millones de vectores soporte), es un problema sin resolver.

Los datos discretos presentan otro problema, aunque con el apropiado reescalado, se consiguen obtener excelentes resultados (Joachims, 1997). Por otra parte, aunque se han realizado muchos trabajos en el entrenamiento para sistemas SVM multiclase en un paso,

el diseño óptimo para sistemas clasificadores multiclase basados en SVM, continúa como área de investigación.

Ejemplos de aplicación

SVM es ampliamente usado en aplicaciones para clasificaciones espectro-espaciales de imágenes hiperespectrales. La atracción al uso de este método recae en que ubica el hiperplano óptimo entre las clases de interés y el resto de clases en un nuevo espacio de características de alta dimensionalidad, tomando en consideración solamente las muestras de entrenamiento situadas en el borde de las distribuciones de las clases. Por tanto, es fácil encontrar muchos estudios de investigación que emplean estos algoritmos en adquisición de imágenes en teledetección por medio de sensores hiperespectrales, tales como los sensores utilizados en *Airborne Visible Infrared Imaging Spectrometer (AVIRIS)* and *Hyperspectral Digital Imagery Collection Experiment (HYDICE)*, los cuales han demostrado su utilidad en numerosas aplicaciones de observación de la Tierra [70].

Aunque la utilización de SVMs en el ámbito de la teledetección parece haber tenido mayor rendimiento respecto a otras técnicas clásicas, se puede encontrar en otro tipo de aplicaciones. SVM se ha empleado en reconocimiento de dígitos aislados escritos a mano, reconocimiento de objetos, identificación de hablantes, detección de *charmed quark*, detección de caras en imágenes, categorización de textos, bioinformática o incluso en neuroimagen [26, 97].

En la Figura 3.32 se ilustra un ejemplo en el que se ha empleado SVM en una aplicación basada en neuroimagen. En la figura se aprecia una imagen obtenida del lóbulo occipital, en la que se representan distintos tipos de información usando diferentes colores. En la fase de procesado de la imagen, esta se divide en áreas de 3 x 3 milímetros y de cada subdivisión se obtiene un histograma de las regiones identificadas.

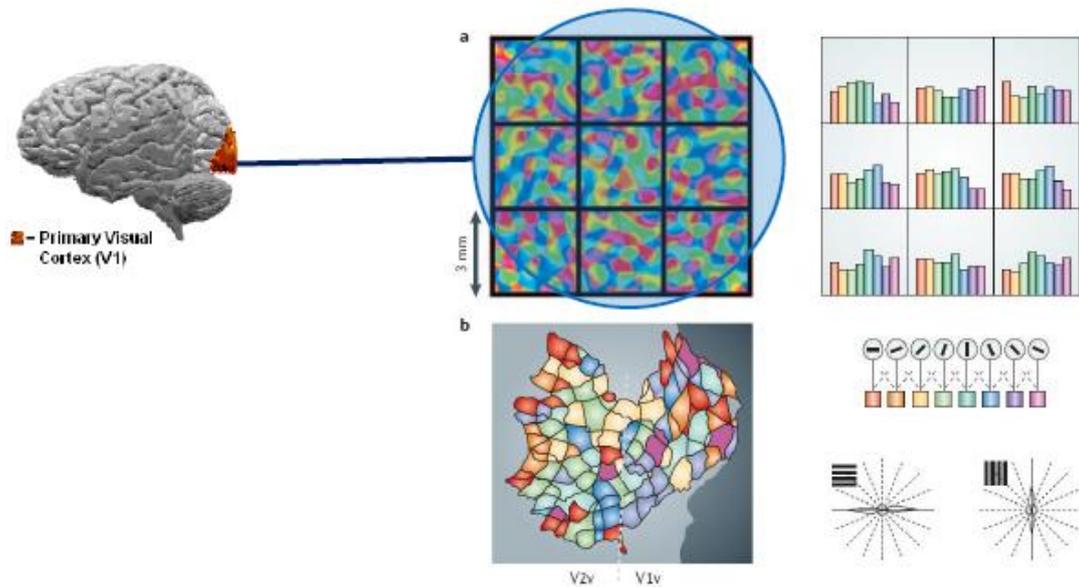


Figura 3.32: Ejemplo de aplicación de SVM en neuroimagen.

3.6 Métodos de validación de algoritmos de clasificación

Ante la gran variedad de técnicas de clasificación existentes, así como la continua proliferación de nuevas tecnologías, se hace necesario el empleo de métricas o esquemas comparativos que permitan analizar, de forma cualitativa, el rendimiento de las nuevas tecnologías planteadas, contrastando sus resultados con los proporcionados por las ya existentes. La mayor parte de las técnicas de evaluación de los algoritmos de análisis de imágenes, se basan en el concepto de *ground-truth* o verdad terrena [98]. Idealmente, es posible definir el concepto de verdad terrena como el resultado de clasificación o interpretación óptimo al que debe llegar un algoritmo [99, 100].

La verdad terrena suele venir caracterizada por información relevante acerca de las propiedades en el mundo real de un conjunto de objetos que se desean identificar o caracterizar. Esta información suele obtenerse mediante mediciones realizadas directamente en la zona de estudio cubierta por la imagen [101], en medicina por ejemplo, puede obtenerse mediante el estudio de la muestra por un médico patólogo, aunque también es posible obtener información de verdad terrena mediante la aplicación de técnicas algorítmicas [27]. No obstante, la primera alternativa es la más fiable, aunque pueda resultar costosa debido a que es necesario etiquetar todas las muestras obtenidas.

Pese a que verdad terrena es un término utilizado principalmente en teledetección, puesto que es el primer campo en el que se ha empleado las imágenes hiperespectrales, se

ha decidido usar además en este estudio para facilitar la comprensión de los resultados mostrados.

A continuación, se presenta una descripción de las diferentes métricas que se pueden aplicar para evaluar la habilidad de un algoritmo en cuanto a la clasificación e identificación de objetos de interés en una imagen. En concreto, se muestra una de las aproximaciones más ampliamente usadas y que servirá para el posterior estudio comparativo de este proyecto, la matriz de confusión, a partir de la cual se derivan otras métricas como el porcentaje de acierto en la clasificación, la especificidad o la sensibilidad.

La matriz de confusión es una técnica que permite evaluar la precisión de algoritmos de clasificación de imágenes [102]. Esta técnica presupone que la información verdadera viene expresada en forma de un mapa temático [103, 104], caracterizado por las siguientes propiedades:

- Cada píxel se encuentra etiquetado como perteneciente a una determinada clase, de forma que se tienen N clases o regiones de referencia, $\{R_i\}_{i=1}^N$
- Las regiones de referencia son mutuamente excluyentes entre sí, es decir, dos regiones diferentes no tienen ningún píxel en común: $R_i \cap R_j = \emptyset, \forall i \neq j$

Suponiendo que cada píxel i de la imagen a evaluar, I , es asignado por el algoritmo como perteneciente a una determinada clase C_i , teniendo N clases, los conjuntos C_i determinan una partición de la imagen a evaluar, es decir, la unión de todos ellos da como resultado la imagen y dos conjuntos distintos que no tienen ningún elemento en común.

$$\bigcup_{i=1}^N C_i = I \text{ y } C_i \cap C_j = \emptyset, \forall i \neq j \quad (3.38)$$

Dadas las consideraciones anteriores, en la Figura 3.33 se ilustra a modo de ejemplo el mapa temático de clasificación de la verdad terrena, asociada a la imagen a clasificar, y el resultado de clasificación proporcionado por un determinado algoritmo para dicha imagen. En la Tabla 3.2 se muestra el proceso de construcción de la matriz de confusión que cuantifica la precisión del algoritmo en la tarea de clasificación. Como puede apreciarse, las entradas de la matriz vienen expresadas en la forma a_{jk} , donde $a_{jk} = \text{cardinal} \{C_j \cap R_k\}$, que se corresponde al número de píxeles de la región resultante al efectuar la intersección entre una clase C_j obtenida por el algoritmo y una clase verdad terrena R_k [102].

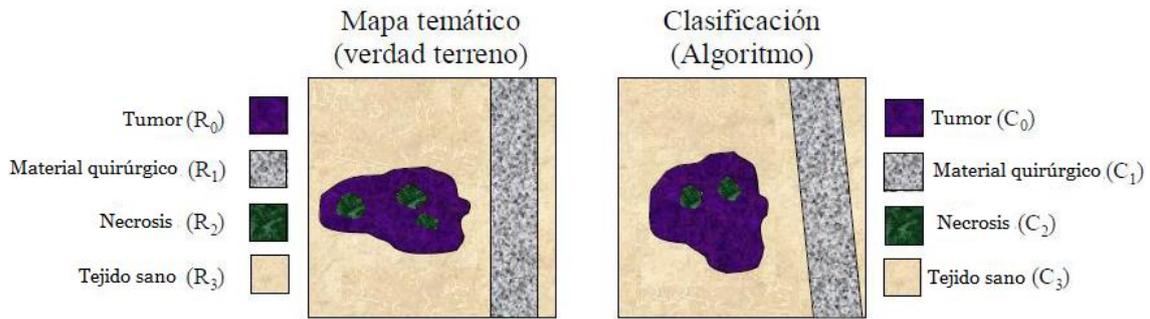


Figura 3.33: Ejemplo de un mapa temático de clasificación a partir de la verdad terrena asociada a la imagen a clasificar.

Tabla 3.2: Matriz de confusión del ejemplo ilustrado en la Figura 3.33.

	R_0	R_1	R_2	R_3
C_0	$a_{00} = C_0 \cap R_0 $	$a_{01} = C_0 \cap R_1 $	$a_{02} = C_0 \cap R_2 $	$a_{03} = C_0 \cap R_3 $
C_1	$a_{10} = C_1 \cap R_0 $	$a_{11} = C_1 \cap R_1 $	$a_{12} = C_1 \cap R_2 $	$a_{13} = C_1 \cap R_3 $
C_2	$a_{20} = C_2 \cap R_0 $	$a_{21} = C_2 \cap R_1 $	$a_{22} = C_2 \cap R_2 $	$a_{23} = C_2 \cap R_3 $
C_3	$a_{30} = C_3 \cap R_0 $	$a_{31} = C_3 \cap R_1 $	$a_{32} = C_3 \cap R_2 $	$a_{33} = C_3 \cap R_3 $

A partir de la matriz de confusión pueden derivarse algunas medidas de confusión genéricas [105] como:

- **Porcentaje de acierto global (Overall Accuracy, OA).** Porcentaje de píxeles de todas las clases correctamente clasificados:

$$OA = \frac{\sum_j a_{jj}}{\sum_{jk} a_{jk}} \times 100 \quad (3.39)$$

En caso de disponer de un clasificador ideal que clasifica todas las muestras correctamente, el porcentaje de acierto global debe ser 100%. Cuando el conjunto de referencia no está bien definido, OA no es representativo con respecto a la verdadera actuación del clasificador. Por ejemplo, si una de las clases tiene muy pocos píxeles de referencia, su influencia sobre la computación de OA será muy baja.

Por otro lado, la medicina es una ciencia de probabilidades en la que se maneja la incertidumbre y la causalidad. Esta incertidumbre se extiende a las actividades diagnósticas donde es importante la validez, la fiabilidad y el riesgo del diagnóstico. Una prueba diagnóstica es buena cuando ofrece resultados positivos en enfermos y negativos en sanos [106].

En esta situación se puede establecer un modelo de clasificador binario en el que los resultados se etiquetan positivos (p) o negativos (n). En este marco teórico, el problema de predicción ofrece cuatro posibles resultados a partir de la clasificación realizada. Si el resultado de una exploración es p y el valor dado es también p , entonces se conoce como un *Verdadero Positivo* (V_p); sin embargo, si el valor real es n entonces se conoce como un *Falso Positivo* (F_p). De igual modo, tenemos un *Verdadero Negativo* (V_n) cuando tanto la exploración como el valor dado son n , y un *Falso Negativo* (F_n) cuando el resultado de la predicción es n pero el valor real es p [107].

De lo anterior se puede definir un experimento a partir de P instancias positivas y N negativas. Los cuatro posibles resultados se pueden formular en una matriz de confusión de 2×2 (Tabla 3.3).

Tabla 3.3: Matriz de confusión para la clasificación biclase de ámbito médico.

		CONDICIÓN VERDAD TERRENA	
		Condición Positiva	Condición Negativa
CONDICIÓN PREDICHA	Población Total		
	Predicción Positiva	Verdaderos Positivos (V_p)	Falsos Positivos (F_p)
	Predicción Negativa	Falsos Negativos (F_n)	Verdaderos Negativos (V_n)

A partir de esta matriz de confusión se derivan dos métricas adicionales empleadas habitualmente en el ámbito médico:

- **Sensibilidad.** Nos indica la capacidad del clasificador de dar como casos positivos, los casos realmente enfermos; proporción de enfermos correctamente identificados. Es decir, la sensibilidad caracteriza la capacidad del clasificador para detectar la enfermedad. Se define como:

$$\text{Sensibilidad} = \frac{V_p}{V_p + F_n} = \frac{\text{Verdaderos positivos}}{\text{Total de positivos en verdad terrena (población de } p)} \quad (3.40)$$

- **Especificidad.** Nos indica la capacidad del clasificador de dar como casos negativos, los casos realmente sanos; proporción de sanos correctamente identificados. Es decir, la especificidad caracteriza la capacidad del clasificador para detectar la ausencia de enfermedad en sujetos sanos. Se define como:

$$\text{Especificidad} = \frac{V_n}{V_n + F_p} = \frac{\text{Verdaderos negativos}}{\text{Total de negativos en verdad terrena (población de } n)} \quad (3.41)$$

Hay que destacar que, si bien OA nos da la proporción de muestras clasificadas correctamente a nivel global, la sensibilidad y la especificidad nos indican la proporción de muestras, enfermas y sanas respectivamente, correctamente clasificadas. Es decir, en un clasificador biclase en el que se toma en consideración una clase concreta, la sensibilidad y la especificidad respectivamente, nos dan una idea de las muestras correctamente seleccionadas y de las muestras correctamente rechazadas.

En el campo médico, en un clasificador perfecto se obtendría un porcentaje de sensibilidad del 100% (todas las muestras enfermas diagnosticadas como enfermas), y un porcentaje de especificidad del 100% (todas las muestras sanas diagnosticadas como sanas).

3.7 Resumen

Las firmas espectrales de los datos espectrográficos y de los píxeles que conforman las imágenes hiperespectrales, contienen información unívoca que caracteriza a diferentes tipos de sustancias. Por ello, este tipo de datos hacen posible la detección, diferenciación o clasificación de los diferentes tipos de tejidos cerebrales, tumorales o sanos, debido a la diversidad que se manifiesta en su composición química.

La espectroscopia, aunque se trata de una técnica de imagen hiperespectral limitada espacialmente a un único píxel, hace posible reducir el coste computacional que presentan las imágenes hiperespectrales convencionales, incluso cuando presenta un número de bandas espectrales elevado. Si los datos espectrográficos contienen mayor número de bandas espectrales que las imágenes hiperespectrales, es posible que además estén dotados de mayor información relevante en la clasificación de tejidos cerebrales. Es por esta razón por la que en este proyecto se ha prescindido de emplear técnicas de preprocesado que reduzcan el número de bandas de los datos espectrográficos. Por otra parte, al tratarse de datos de un único píxel, no tiene sentido aplicar técnicas de procesado espacial.

Por otro lado, la información recogida en el estudio del estado del arte sobre los algoritmos clasificadores empleados en el campo de las imágenes hiperespectrales, ha sido necesaria en la toma de decisión sobre los sistemas de clasificación implementados en este proyecto. Tras contrastar las diferentes opciones y tomando en consideración la base de datos de muestras de las que se dispone, se ha escogido el método supervisado basado en las Máquinas de Vectores Soporte (SVMs).

Esta técnica ha sido elegida por varias razones. En primer lugar, destaca entre otros métodos por aparecer en gran parte de la literatura relacionada con la clasificación de imágenes hiperespectrales. Esto se debe a que manifiesta ser un clasificador muy robusto y a que ostenta buenos resultados trabajando con este tipo de datos. En segundo lugar, demuestra gran eficacia al trabajar con datos de muchas bandas espectrales en conjuntos de datos no balanceados. A nivel de errores, permite una mejor generalización hacia muestras de test, antes que una mejor clasificación del conjunto de entrenamiento. Esta razón es relevante puesto que para la elaboración de este proyecto se ha contado con una base de datos muestral muy poco balanceada, ya que para algunas clases se dispone de gran cantidad de muestras, mientras que para otras clases, se dispone de cantidades muy reducidas.

Capítulo 4

Base de datos de las muestras espectrográficas

4.1 Introducción

Como se comentó en el apartado 1.2 de esta memoria, la base de datos con la que se ha contado en la realización de este Proyecto Fin de Carrera, y que ha servido como soporte muestral para el desarrollo de los sistemas de clasificación que se han implementado en el mismo, consiste de un conjunto de datos espectrográficos suministrados por el Hospital Universitario de Southampton de Reino Unido. Los datos se obtuvieron a partir de muestras de tejido cerebral cuya resección tuvo lugar en intervenciones quirúrgicas, craneotomías, realizadas a pacientes con tumores cerebrales. Los datos espectrográficos fueron recogidos en dos campañas, empleando dos espectrómetros diferentes que se encargaron de capturar las muestras de tejido cerebral, para proporcionar la respuesta o firma espectral de un único píxel perteneciente a cada una de las muestras.

En este capítulo se exponen las características del proceso de adquisición de las muestras, incluyendo el proceso de preparación de las mismas que realizó el personal médico, así como las características de los espectrómetros empleados.

Adicionalmente, se incluye un apartado en el que se explica brevemente la tarea de eliminación de determinadas muestras calificadas como *outliers*, realizada mediante la técnica de Mapeo del Ángulo Espectral en el PFC “Clasificación de tumores cerebrales mediante espectroscopia utilizando máquinas de vectores soporte (SVM)” [4].

4.2 Adquisición de los datos espectrográficos

Los datos espectrográficos fueron proporcionados por personal médico del Hospital Universitario de Southampton (UHS). Estas muestras fueron recabadas en dos campañas diferentes, *campana 1* y *campana 2*, de manera que para cada una de las campañas se hizo uso de un espectrómetro diferente.

Las muestras de tejido cerebral de ambas campañas se han obtenido en diferentes procedimientos quirúrgicos, en los que, los médicos tenían la sospecha sobre el padecimiento de tumores cerebrales por parte de los respectivos pacientes. Esta toma de muestras ha sido posible gracias a que los pacientes sometidos a las craneotomías han expresado su consentimiento para participar en el proyecto de investigación.

Hay que destacar que el conjunto de muestras de ambas campañas es no balanceado, tomando en consideración la baja proporción de muestras no tumorales frente a la gran cantidad de muestras tumorales. Esto se debe a que la resección de un tumor cerebral supone que la mayor parte del tejido resecado sea tumoral.

Como se explicó en el apartado 2.3, los neurocirujanos no resecan únicamente tejido tumoral en las intervenciones quirúrgicas. Esto se debe a ese margen de seguridad que reduce el riesgo de que el tumor se vuelva a reproducir. Por tanto, el hecho de que se cuente con una baja proporción de muestras de tejido no tumoral en la base de datos, es un problema intrínseco e inevitable.

Durante la cirugía, el tejido extraído era considerado tejido tumoral si era identificado como tal por al menos uno de dos cirujanos. Posteriormente, las muestras fueron etiquetadas por un patólogo que clasificó cada ejemplar con una diagnosis detallada, el tipo de tejido del que provenía y la fecha.

Una vez finalizada la intervención y la clasificación patológica, las muestras fueron sometidas a un proceso de preparación antes de ser introducidas en los espectrómetros. En dicho proceso algunas de las muestras fueron lavadas con una solución salina del 0,9% para

eliminar los rastros de sangre visibles; a continuación, se pesaron y se secaron; una vez secas se molieron en un mortero con *bromuro de potasio (KBr)* en polvo; y finalmente, la mezcla resultante fue dividida en porciones de aproximadamente 0,5 gramos, a las que se le aplicó una presión de aproximadamente 10 toneladas con una prensa en vacío para crear una muestra sólida adecuada para introducirla en el espectrómetro. Las muestras se almacenaron individualmente a -20°C , con una bolsita de gel de sílice con el fin de reducir la absorción de agua por el bromuro potásico.

4.2.1. Muestras de la campaña 1

Las muestras de la campaña 1 fueron adquiridas empleando el espectrómetro *Varian FT-IR 600* (Varian Inc. Palo Alto, US) (Figura 4.1), conjuntamente con el software *Agilent Resolutions Pro V.5* (Agilent Technologies UK Ltd. Stockport, UK / U.S. Santa Clara). Este espectrómetro obtiene un único píxel, pero con una alta resolución espectral y ha sido uno de los instrumentos de esta clase pioneros en áreas como la biotecnología, la alimentación, la agricultura y la seguridad. Además, los espectrómetros de la serie 600-IR son especialmente adecuados para aplicaciones de investigación y desarrollo de métodos de investigación [4].



Figura 4.1: Espectrómetro Varian FT-IR 600 [108].

Este espectrómetro emplea el método de adquisición de muestras hiperespectrales denominado *Espectroscopia Infrarroja* mediante transformada de Fourier (*Fourier Transform Infrared spectroscopy, FT-IR*). Dicho método registra la imagen con un interferómetro y luego emplea la transformada rápida de Fourier para obtener la imagen en

el dominio de la frecuencia (espectro). Esta técnica de recolección de imágenes hiperespectrales es la más empleada en el ámbito médico [32].

Estas muestras fueron tomadas por el espectrómetro en modo transmitancia y transformadas por el software Agilent al valor correspondiente de absorción. Esta relación entre transmitancia y absorbancia se ha mostrado previamente en la ecuación (3.3).

Los valores de absorción fueron obtenidos en el rango comprendido entre 400 nm y 6000 nm, con una resolución espectral de 1,93 nm, disponiendo de esta forma de 2906 bandas por cada una de las muestras de esta campaña.

Las muestras de tejido cerebral de la campaña 1 proceden de 26 adultos de ambos sexos (hombres y mujeres). El total de muestras de esta campaña es de 220, donde 202 de ellas se corresponden a muestras tumorales mientras que las 18 restantes provienen de tejidos sanos.

En concreto para esta campaña, el patólogo especificó en las muestras tumorales, el tipo de tumor; y en las muestras sanas, si se trataba de sustancia gris o blanca.

En la Figura 4.2, se muestra un ejemplo de algunas firmas espectrales asociadas a uno de los pacientes de esta campaña.

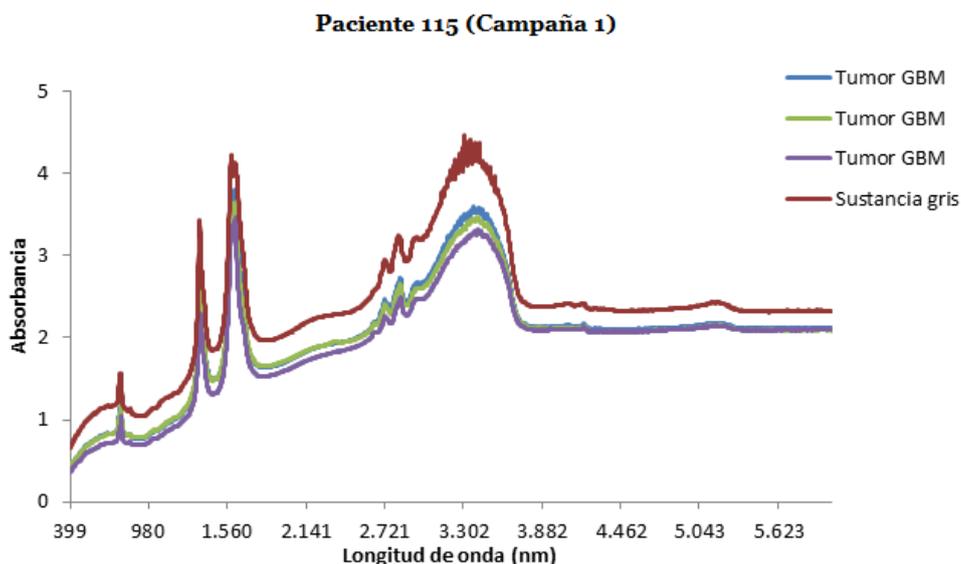


Figura 4.2: Firmas espectrales de muestras asociadas al paciente 115 [4].

Por otro lado, en la Tabla 4.1 se presenta la descripción de las muestras de la campaña 1. Se han ordenado siguiendo el etiquetado proporcionado por el Hospital de Southampton.

No obstante, las muestras tumorales también se han distinguido según el grado de tumor y si este es considerado de alto o bajo grado. Las muestras de tejido sano han sido diferenciadas según el tipo de sustancia al que pertenecen. Estas agrupaciones fueron empleadas en el PFC anterior [4], para definir las clases utilizadas en los sistemas de clasificación implementados en dicho proyecto. Sin embargo, se debe aclarar que la primera muestra de esta tabla, diagnosticada como Astrocitoma pilocítico de grado I, fue descartada en el proyecto anterior [4] debido a que se trata de la única muestra de grado I en la campaña 1. Al no disponerse de un mayor número de muestras de grado I en la campaña 1, se hacía imposible definir conjuntos de muestras de entrenamiento y test para clasificar este grado. En el presente PFC, esta muestra se añadió al conjunto de muestras de grado I empleado en las clasificaciones de la fusión de ambas campañas.

Como se refleja en la Tabla 4.1, la campaña 1 no solo presenta un conjunto total de muestras no balanceado respecto al número de muestras sanas y muestras tumorales, sino que además, en la subdivisión de muestras tumorales puede observarse que el número de muestras de grado I, II y grado III, es muy escaso frente al número de muestras de grado IV. Este hecho se debe a que los tumores de grado I, II y III pueden ser no diagnosticados hasta haberse transformado en un tumor de grado IV, ser asintomáticos, o simplemente, a que los médicos consideran que una craneotomía es un proceso muy arriesgado en el caso de un tumor que puede no llegar a convertirse en maligno.

Entre las muestras de la campaña 1 que proporcionó el Hospital Universitario de Southampton, también destaca una única muestra del tumor glial astroblastoma. Como se mencionó en el apartado 2.3.4, la organización mundial de la salud aún no ha establecido un grado histológico de modo definitivo para este tipo de tumor [17]. Por no disponerse de más muestras tumorales para este tipo en ninguna de las campañas, haciendo imposible el desarrollo de un sistema basado en SVM que trate de clasificarlas, se ha decidido excluir dicha muestra de la base de datos del presente PFC.

Tabla 4.1: Relación de muestras proporcionadas por el Hospital Universitario de Southampton (campaña 1).

Clases		Tipo	Etiquetas (UHS)	Número de muestras (por tipo)	Número de muestras (por clase)	
Muestras tumorales	Grado bajo	Grado I	Astrocitoma pilocítico	“Pilocytic astro”	1	
			Astrocitoma dfuso	“II astro”	10	
	Grado II		Oligodendroglioma	“II oligo”	12	22
		Grado III	Astrocitoma anaplásico Oligodendroglioma anaplásico	“III astro” “III oligo”	16 14	30
	Grado alto	Grado IV	Glioblastoma multiforme con astrocitos y oligodendrocitos malignos	“GBM mixed astro/oligo”	5	149
			Glioblastoma multiforme con gliosarcoma	“GBM with gliosarcomatous”	12	
			Glioblastoma multiforme	“GBM astro”	16	
			Gliosarcoma	“IV gliosarcoma”	20	
	Glioblastoma multiforme	“GBM”	96			
Muestras sanas	Sustancia blanca	Sustancia blanca	“White matter”	11	18	
	Sustancia gris	Sustancia gris	“Grey matter”	7		
Número total de muestras de la campaña 1					220	

A continuación, se ilustran las firmas espectrales de las muestras de la campaña 1 especificando en cada caso el grado y el diagnóstico, si se trata de un tumor, o el tipo de sustancia, si se trata de tejido sano. En estas ilustraciones (Figura 4.3 a Figura 4.14) se agrupan las firmas espectrales pertenecientes a un mismo diagnóstico.

Al disponerse de más de una muestra por diagnóstico, a excepción de la diagnosis de astrocitoma pilocítico y de astroblastoma (esta última no representada por ser descartada), las firmas espectrales de las muestras que manifiestan la misma diagnosis, se representan en una misma gráfica. Además, se ilustra su media y su desviación típica en gráficas anexas, en la misma figura.

Se asume que la longitud de onda de todas las firmas espectrales ilustradas a lo largo de este PFC, se representarán en nanómetros, puesto que los datos espectrográficos se han obtenido en estas unidades.

- Firma espectral de la muestra de grado I:
 - *Astrocitoma pilocítico*

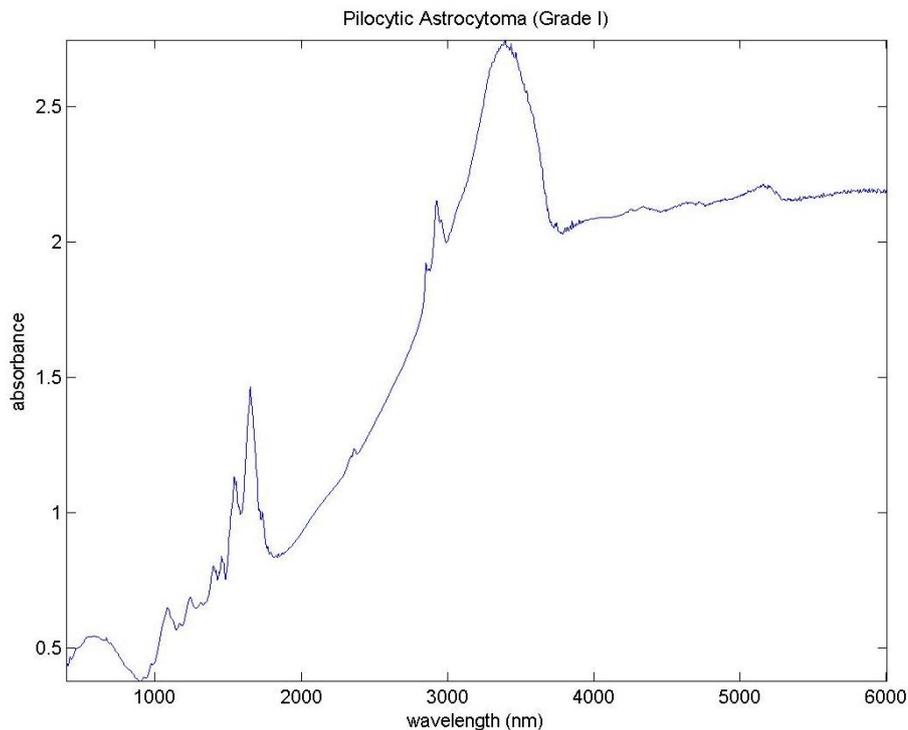


Figura 4.3: Firma espectral de la muestra de astrocitoma pilocítico (Grado I).

- Firmas espectrales, media y desviación típica de las muestras de grado II:
 - *Astrocitoma difuso*

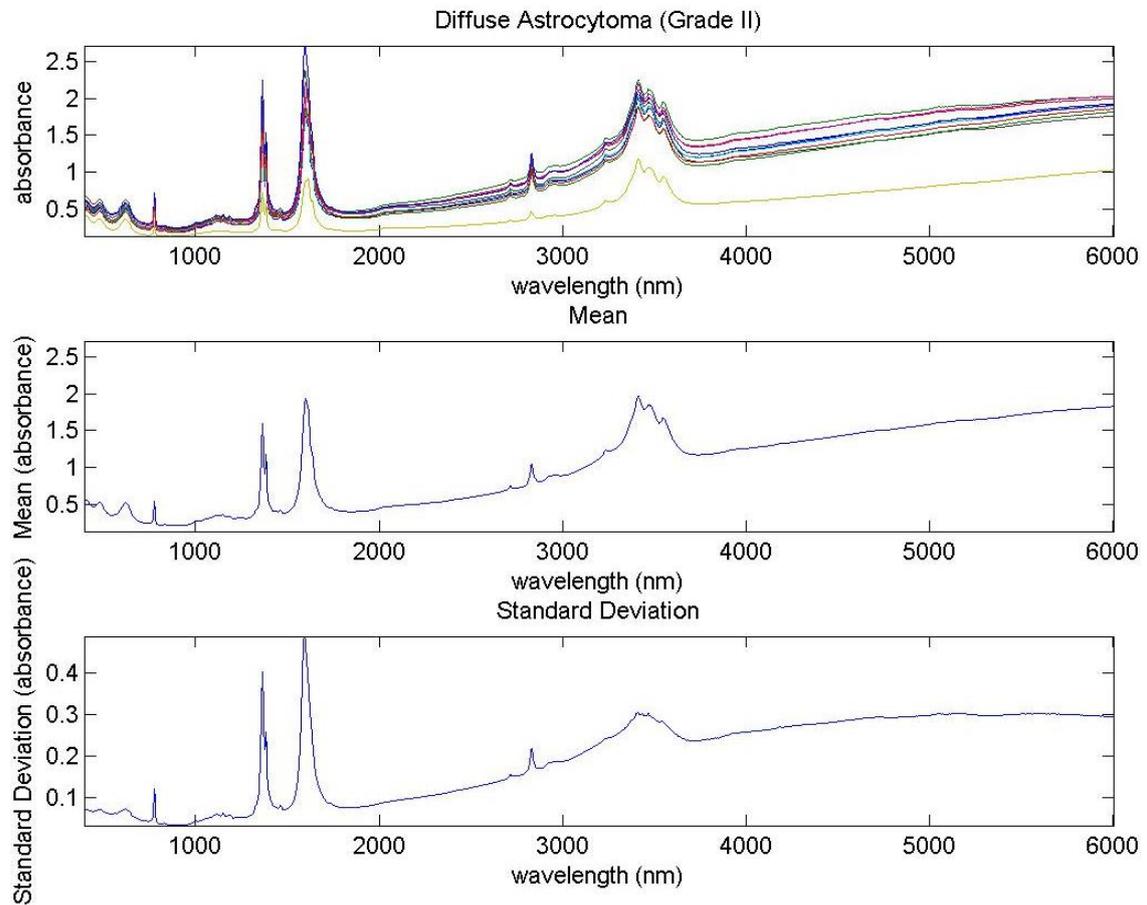


Figura 4.4: Firmas espectrales, media y desviación típica de las muestras de astrocitoma difuso (Grado II).

- *Oligodendroglioma*

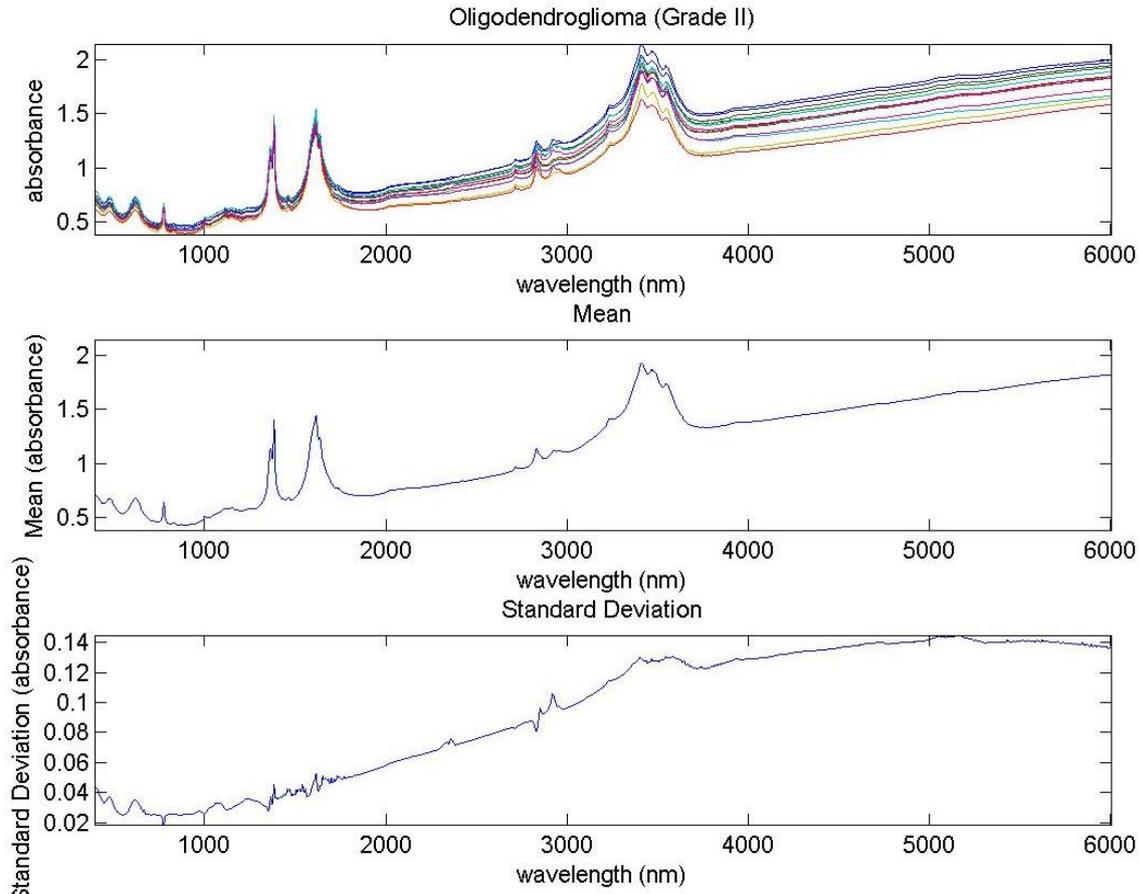


Figura 4.5: Firmas espectrales, media y desviación típica de las muestras de oligodendroglioma (Grado II).

- Firmas espectrales, media y desviación típica de las muestras de grado III:
 - *Astrocitoma anaplásico*

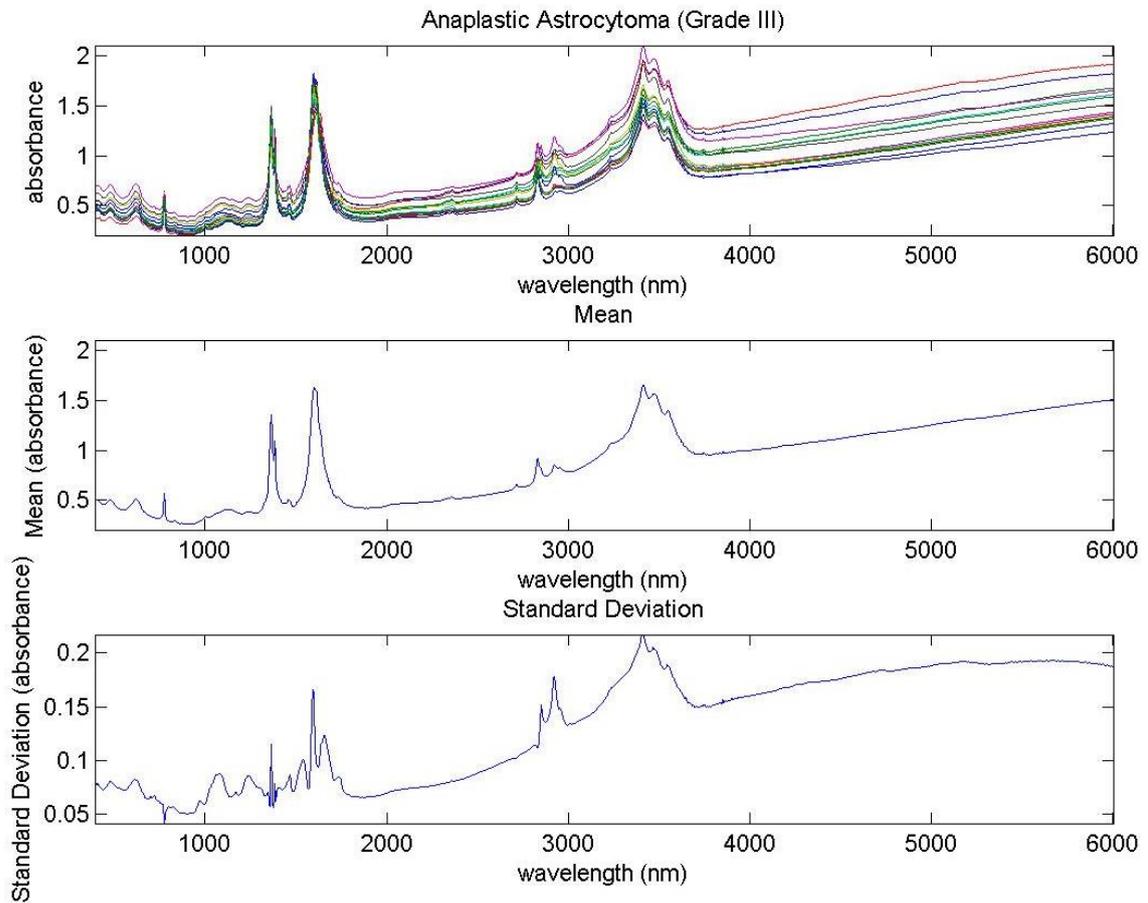


Figura 4.6: Firmas espectrales, media y desviación típica de las muestras de astrocitoma anaplásico (Grado III).

- *Oligodendroglioma anaplásico*

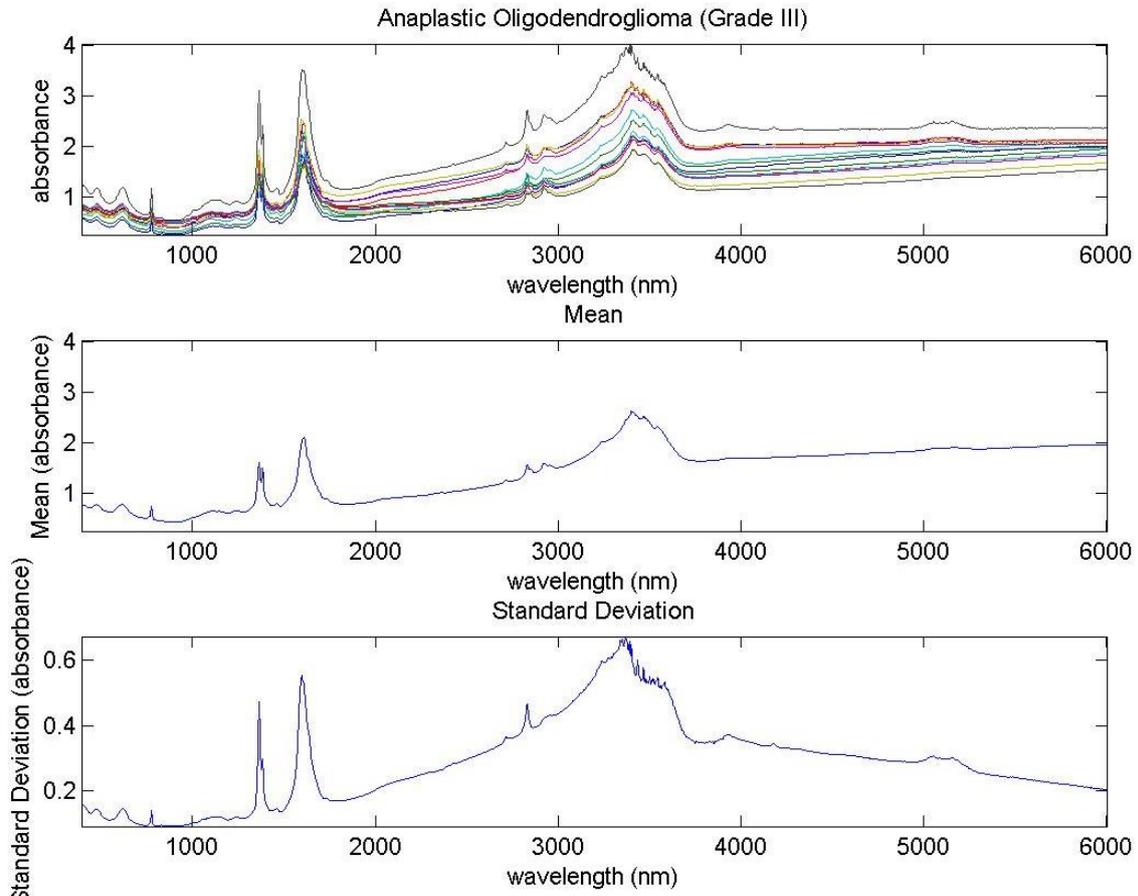


Figura 4.7: Firmas espectrales, media y desviación típica de las muestras de oligodendroglioma anaplásico (Grado III).

- Firmas espectrales, media y desviación típica de las muestras de grado IV:
 - Etiqueta: “GBM mixed astro/oligo”

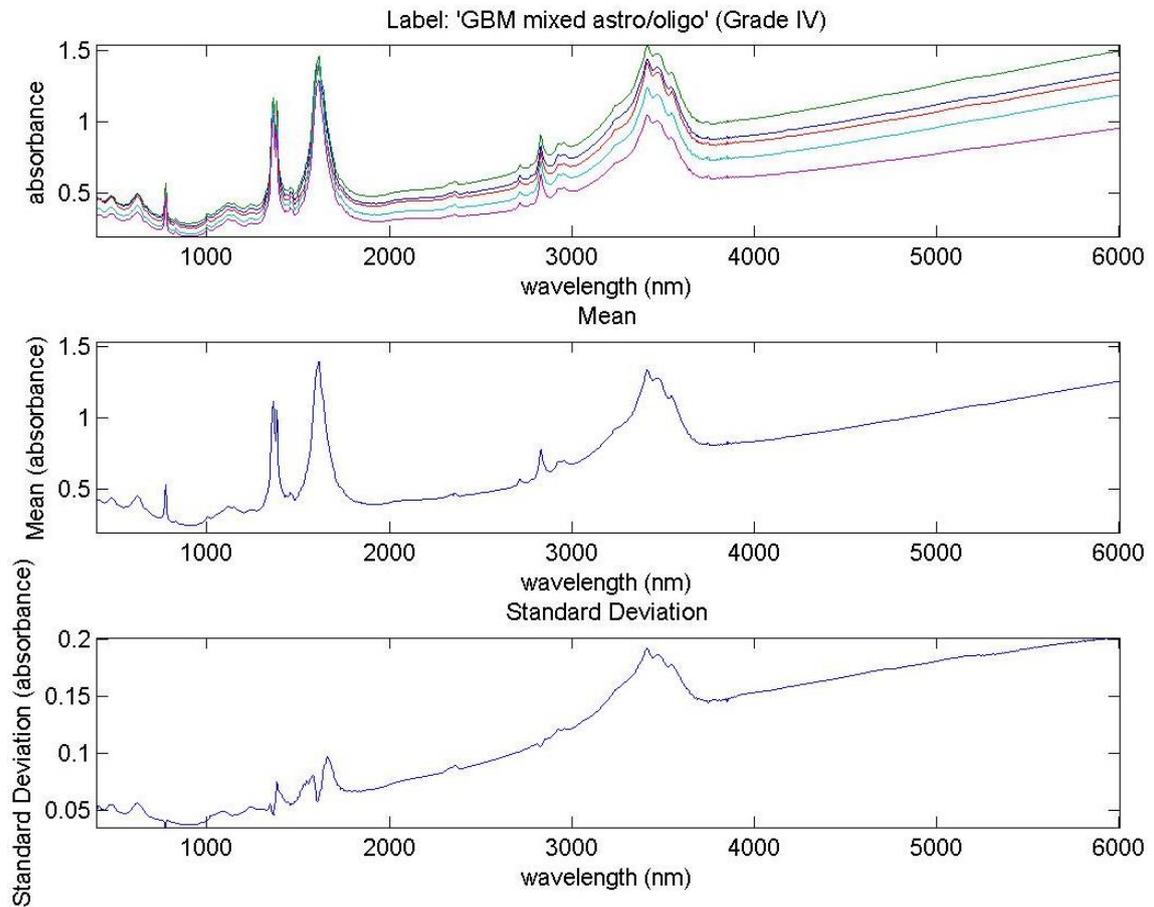


Figura 4.8: Firmas espectrales, media y desviación típica de las muestras de “GBM mixed astro/oligo” (Grado IV).

- *Etiqueta: "GBM with gliosarcomatous"*

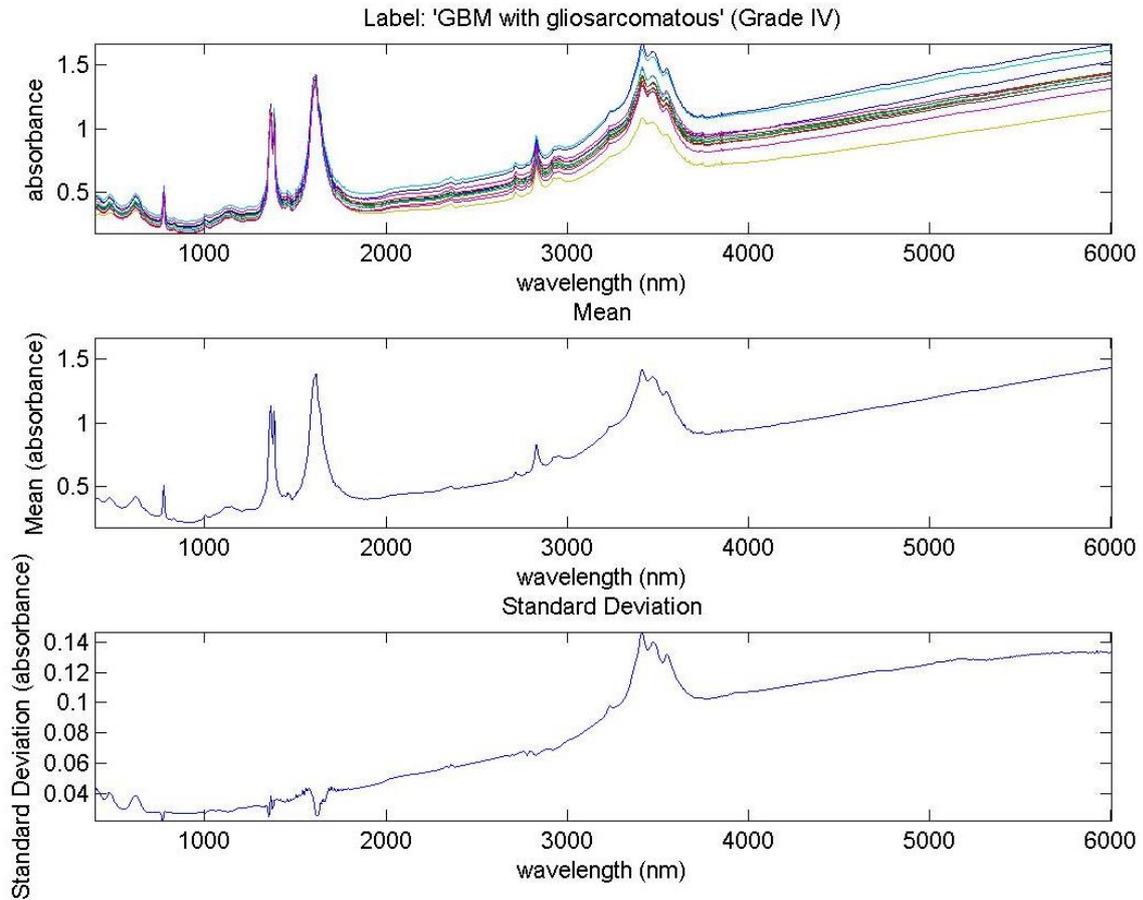


Figura 4.9: Firmas espectrales, media y desviación típica de las muestras de "GBM with gliosarcomatous" (Grado IV).

- Etiqueta: "GBM astro"

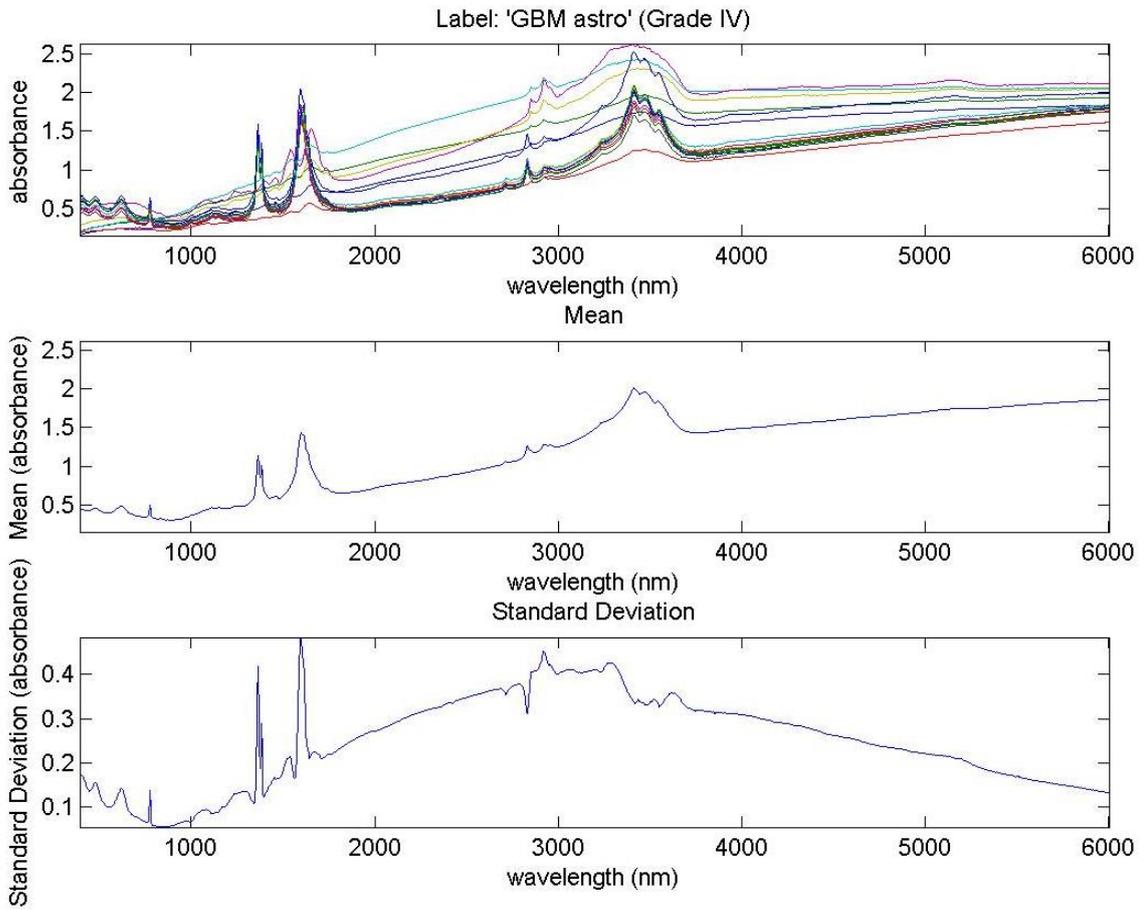


Figura 4.10: Firmas espectrales, media y desviación típica de las muestras de "GBM astro" (Grado IV).

- *Etiqueta: "IV gliosarcoma"*

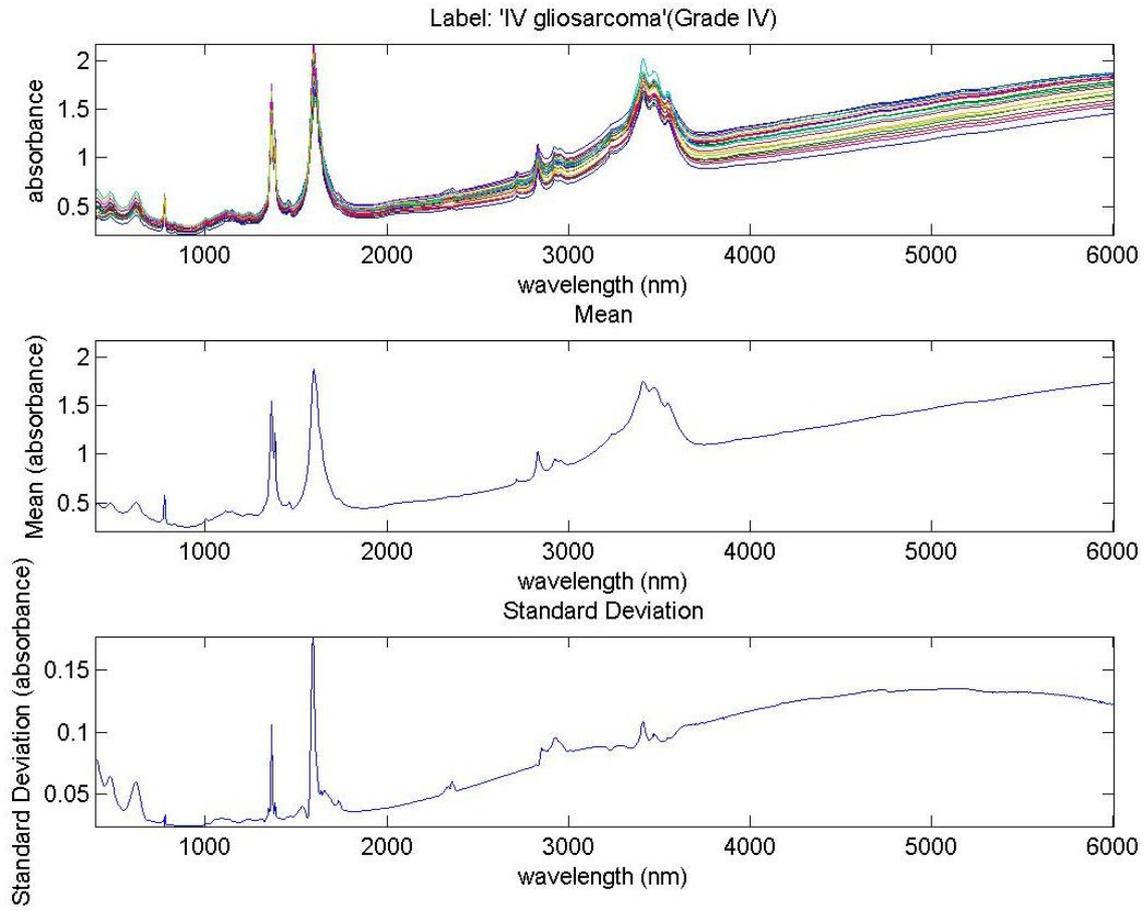


Figura 4.11: Firmas espectrales, media y desviación típica de las muestras de "IV gliosarcoma" (Grado IV).

- *Glioblastoma multiforme*

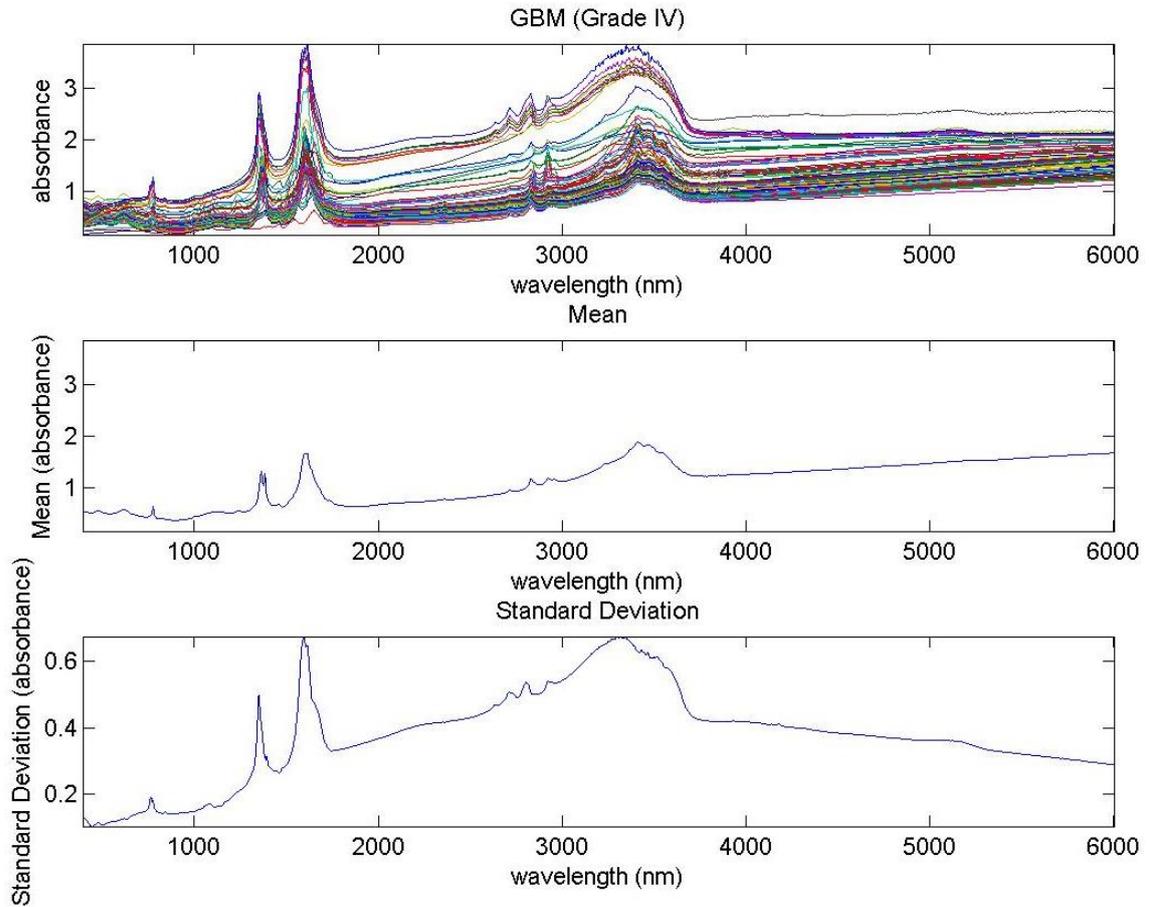


Figura 4.12: Firmas espectrales, media y desviación típica de las muestras de glioblastoma multiforme (Grado IV).

- Firmas espectrales, media y desviación típica de las muestras de tejido sano:
 - *Sustancia blanca*

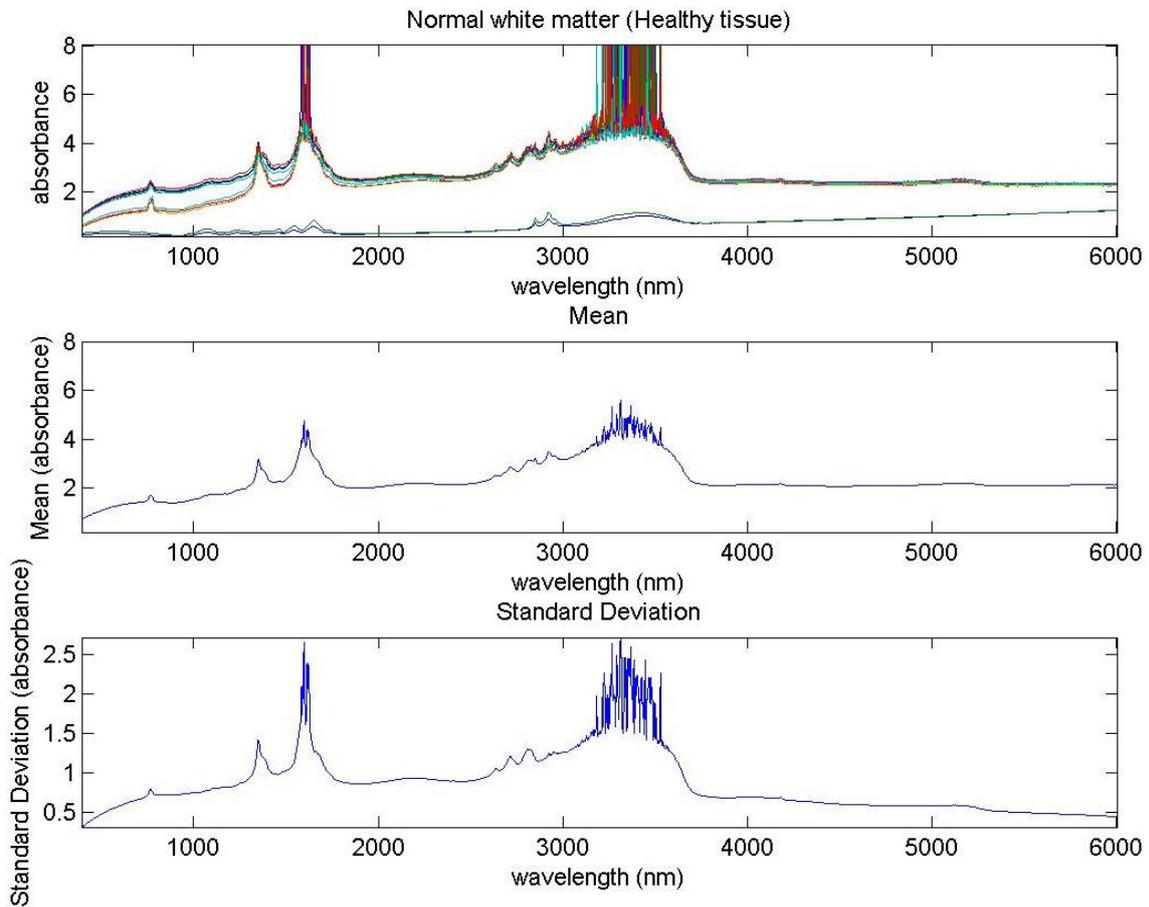


Figura 4.13: Firmas espectrales, media y desviación típica de las muestras de sustancia blanca (Tejido sano).

Como se aprecia en la Figura 4.13, algunas de las firmas espectrales de las muestras presentaron altos niveles de ruido en rangos de su espectro. Este ruido se debe al proceso de obtención de estos datos a través del espectrómetro, no obstante, como se explica en el Capítulo 5, se decidió eliminar el ruido de estas muestras implementando un filtrado de ruido de tipo *smooth*.

- Sustancia gris

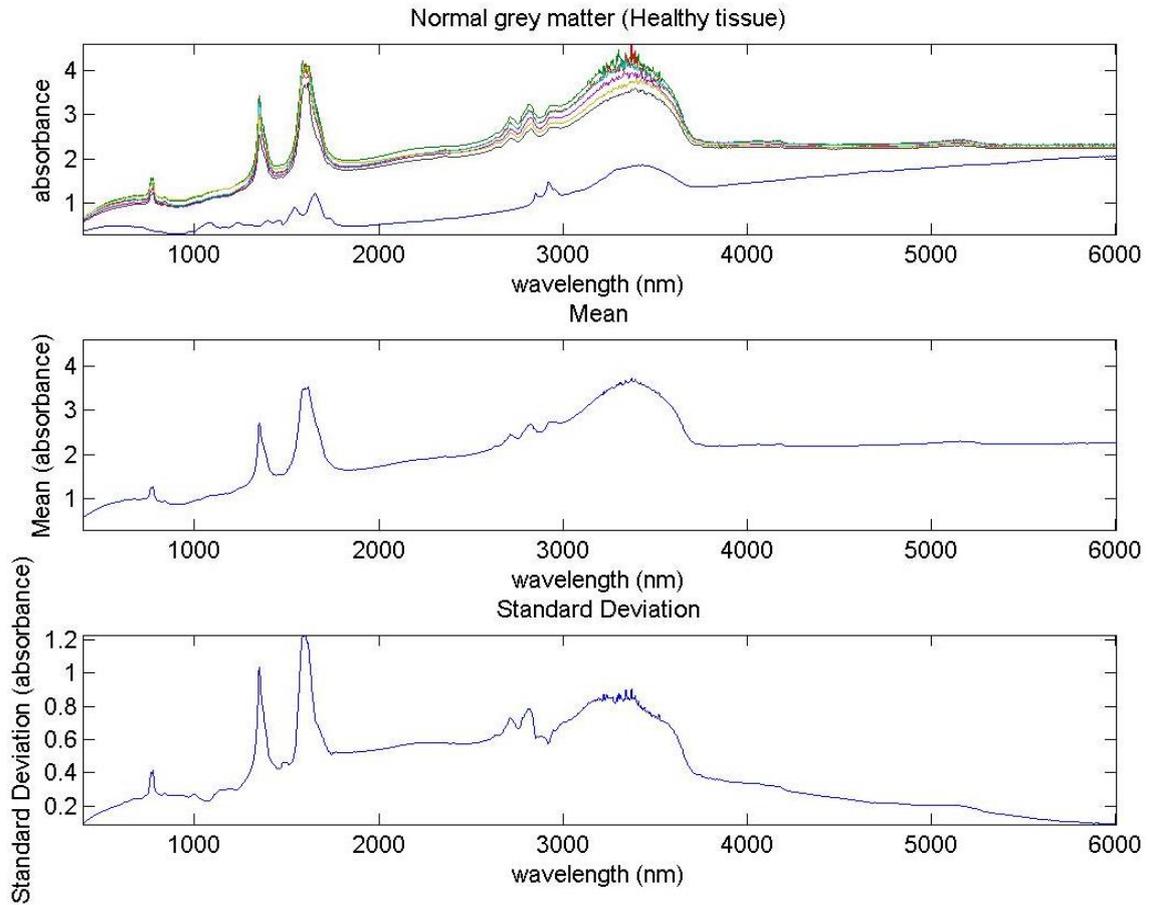


Figura 4.14: Firmas espectrales, media y desviación típica de las muestras de sustancia gris (Tejido sano).

4.2.1.1. Eliminación de muestras de la campaña 1

Debido a que todas las muestras no han recibido exactamente el mismo tratamiento antes de obtenerse los datos espectrográficos, esto es: algunas muestras se lavaron con una solución salina mientras que otras no y la cantidad de bromuro de potasio aplicado no era siempre la misma; en el PFC anterior [4], se decidió realizar una eliminación de muestras consideradas no válidas.

Para el descarte de estas las muestras se empleó la técnica de Mapeo del Ángulo Espectral de forma iterativa a cada una de las siguientes agrupaciones de muestras: grado II, grado III, grado IV y muestras sanas. El umbral de ángulo espectral, para la eliminación de muestras, que se estableció experimentalmente fue de 6 grados.

En la Tabla 4.2 se expone el número de muestras de la campaña 1 que quedaron tras la eliminación de muestras no válidas. Como puede verse en la tabla, se sigue manteniendo la muestra de grado I de esta campaña puesto que, como se ha comentado anteriormente, se ha tenido en consideración a la hora de realizar las clasificaciones del conjunto de muestras que surge de la fusión de las campañas. Evidentemente, la muestra de grado I no se empleó en las clasificaciones realizadas en el presente proyecto en los casos en los que se clasificaba exclusivamente la campaña 1.

En esta campaña se dispone de un total de 182 muestras válidas, de las cuales 167 se corresponden a tumorales mientras que las 15 restantes se corresponden a muestras sanas.

Tabla 4.2: Muestras válidas de la campaña 1 empleadas en este proyecto.

Clases		Tipo	Muestras eliminadas	Muestras válidas	Muestras (por clase)	
Muestras tumorales	Grado bajo	Grado I	Astrocitoma pilocítico	0	1	1
		Grado II	Astrocitoma dfuso	1	9	21
	Oligodendroglioma		0	12		
	Grado alto	Grado III	Astrocitoma anaplásico	0	16	25
			Oligodendroglioma anaplásico	5	9	
		Grado IV	Glioblastoma multiforme con astrocitos y oligodendrocitos malignos	0	5	120
			Glioblastoma multiforme con gliosarcoma	0	12	
			Glioblastoma multiforme	6	10	
			Gliosarcoma	0	20	
	Glioblastoma multiforme	23	73			
Muestras sanas	Sustancia blanca	Sustancia blanca	2	9	15	
	Sustancia gris	Sustancia gris	1	6		
Número de muestras empleadas en este PFC (campaña 1)					182	

4.2.2. Muestras de la campaña 2

Las muestras de la campaña 2 fueron adquiridas empleando el espectrómetro *PerkinElmer FT-IR System, Spectrum BX* (PerkinElmer Inc. Waltham, MA) (Figura 4.15), conjuntamente con el software *PerkinElmer Spectrum V5.3.1* (PerkinElmer Inc. Waltham, MA). Este espectrómetro obtiene un único píxel con alta resolución espectral y también emplea el método de adquisición de muestras hiperespectrales Espectroscopia Infrarroja mediante transformada de Fourier.



Figura 4.15: Espectrómetro PerkinElmer FT-IR System, Spectrum BX.

A diferencia de la campaña 1, las muestras de la campaña 2 fueron obtenidas en transmitancia. Los valores de transmitancia de la campaña 2 se adquirieron en el rango que va desde 1000 nm a 4000 nm, con una resolución espectral de 2 nm, por lo que se dispone de 1501 bandas para cada una de las firmas de esta campaña.

Las muestras de tejido cerebral de la campaña 2 proceden de 17 pacientes, mientras que el número de muestras disponible es de 128. En general, respecto a las muestras de esta campaña no se puede hablar estrictamente de muestras sanas en ningún caso, ya que todas presentan una determinada patología. No obstante, para algunas de estas muestras, en vez de sanas se les ha aplicado el término “no tumorales”, puesto que la patología que presentan no es ningún tipo de tumor cerebral, y no es objeto de estudio de este PFC profundizar en otras patologías que no entren en el ámbito de los tumores cerebrales. La patología con la que son diagnosticadas las muestras no tumorales de esta campaña, es epilepsia con esclerosis del hipocampo.

Por otro lado, hay que destacar también que esta campaña incluye muestras de un tipo de tumor que no se encontraba en la campaña 1. Se trata de la categoría de los tumores secundarios. Concretamente, las metástasis que incluye esta base de datos son del tipo adenocarcinoma pulmonar y carcinoma gástrico.

Por lo tanto, las muestras de esta campaña se pueden dividir en varias categorías. Por un lado, están las muestras tomadas de los tumores primarios, las cuales se pueden subdividir en agrupaciones según su graduación, por otro lado, se dispone de muestras de tumores secundarios (metástasis), y por último, las muestras no tumorales.

En la Tabla 4.3 se detalla la base de datos de la campaña 2. Al igual que la campaña 1, se han ordenado atendiendo al etiquetado proporcionado por el Hospital de Southampton, diferenciando en este caso, los tumores primarios de los secundarios, y en el caso de los primarios, agrupándose por grado. A diferencia de la campaña 1, en esta campaña se ha omitido una subdivisión entre las muestras no tumorales, esto se debe a que algunas de estas muestras presentan una mezcla de sustancia blanca y sustancia gris. Estas agrupaciones fueron aprovechadas para definir las clases utilizadas en los sistemas de clasificación implementados en el presente PFC.

Como se puede comprobar, la campaña 2 consiste también en un conjunto de muestras no balanceado en cualquiera de sus agrupaciones. Por ejemplo, de las 128 muestras que conforman la campaña, solamente 17 son no tumorales frente a las 111 muestras tumorales; de las muestras tumorales, 12 son de metástasis mientras que las 99 restantes provienen de tumores primarios; y de esta última agrupación, 53 son muestras de grado IV mientras que solo se dispone de 6 muestras de grado III.

Por otra parte, salta a la vista que las muestras de grado I llevan la misma etiqueta que las muestras no tumorales, “*Epilepsy*”. No obstante, se especifica literalmente que cada una de estas 16 muestras de grado I, correspondientes al paciente identificado con el número 3, se diagnostican como: “*Epilepsy - Hippocampal sclerosis + Low Grade Glioneuronal tumour WHO Grade I*”. Aunque existe una mezcla de tejidos de diferentes tipos en estas muestras, por presentar tejido tumoral en parte de su composición, se ha decidido asignar estas muestras al conjunto de muestras de grado I.

Tabla 4.3: Relación de muestras proporcionadas por el Hospital Universitario de Southampton (campaña 2). Todas las muestras de esta campaña han sido empleadas en este proyecto.

		Clases	Tipo	Etiquetas (UHS)	Número de muestras (por tipo)	Número de muestras (por clase)		
Muestras tumorales	Tumor primario	Grado bajo	Grado I	Tumor glioneuronal	“Epilepsy”	16	16	
			Grado II	Astrocitoma difuso	“II Astrocytoma”	12	24	
		Grado alto		Oligodendroglioma	“II Oligo”	12		
			Grado III	Astrocitoma anaplásico	“III Astrocytoma”	6	6	
			Grado IV	Glioblastoma multiforme con componentes/características oligodendrogliales		“GBM with oligo”	12	53
				Glioblastoma multiforme		“GBM”	41	
	Tumor Secundario	Adenocarcinoma pulmonar		“Metastasis - lung”	6	12		
		Carcinoma gástrico		“Metastasis - gastric”	6			
		Muestras no tumorales		Epilepsia – Esclerosis del hipocampo	“Epilepsy”	17	17	
	Número total de muestras de la campaña 2						128	

Debe destacarse que no se ha prescindido de ninguna de las muestras que conforman la campaña 2 en el presente PFC, por lo que no ha sido necesario utilizar ningún método de eliminación de muestras.

De la Figura 4.16 a la Figura 4.26, se presentan las firmas espectrales de las muestras de la campaña 2. En ellas se especifica el diagnóstico, el grado, si se trata de tumor primarios, y el tipo de metástasis, si se trata de tumor secundario. En las muestras no tumorales se especifica su patología. En estas gráficas se agrupan las firmas espectrales pertenecientes a un mismo. En la campaña 2 existe más de una muestra por diagnóstico, por ello en todas las figuras también se representa sus respectivas medias y desviaciones típicas.

En esta campaña solamente aparece un grupo de muestras que presentan dudas acerca de los detalles de su diagnóstico. Específicamente, aquellas que se han etiquetado como "*GBM with oligo*". Sin embargo, conocer el grado al que pertenecen nos ha sido suficiente para clasificarlas mediante los sistemas de clasificación que se han implementado en este PFC.

Sin embargo, antes de realizar las clasificaciones correspondientes con los sistemas diseñados en este PFC, se procedió a normalizar todas las muestras entre los valores 0 y 1, tanto en los conjuntos de muestras de entrenamiento como en los de test, para aislar las muestras de las condiciones de iluminación.

- Firmas espectrales, media y desviación típica de las muestras de tumor primario (grado I):
 - *Tumor glioneuronal*

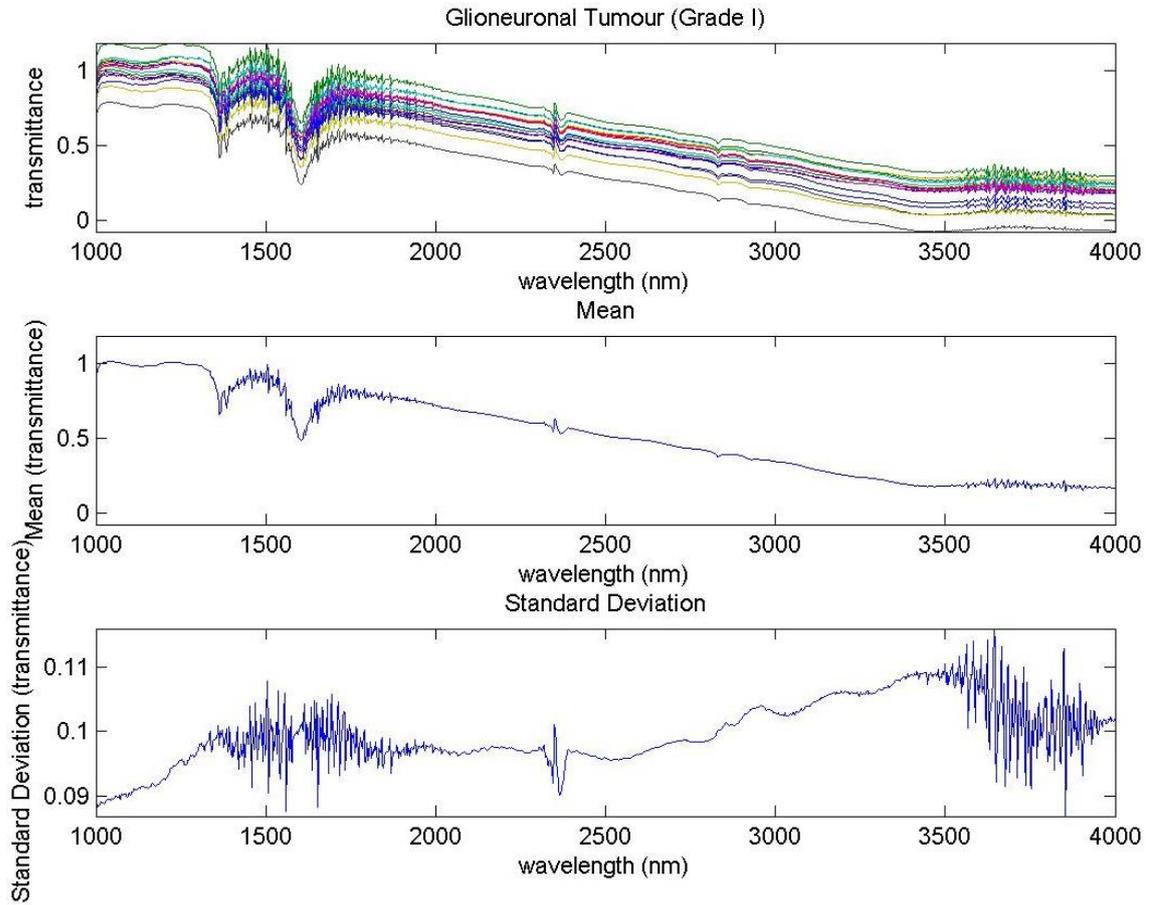


Figura 4.16: Firmas espectrales, media y desviación típica de las muestras de tumor glioneuronal (Grado I).

En las muestras representadas en la Figura 4.16 se refleja claramente la tendencia decreciente que manifiestan las firmas espectrales con el aumento de la longitud de onda. Se trata de las únicas muestras espectrales que se disponen en la campaña 2. Todas provienen del mismo paciente (paciente 3) y hay que recalcar que, aunque no se aprecia gran diferencia en la forma de sus firmas espectrales, algunas provienen de tejido de sustancia blanca mientras que otras de tejido de sustancia gris.

- Firmas espectrales, media y desviación típica de las muestras de tumor primario (grado II):
 - *Astrocitoma difuso*

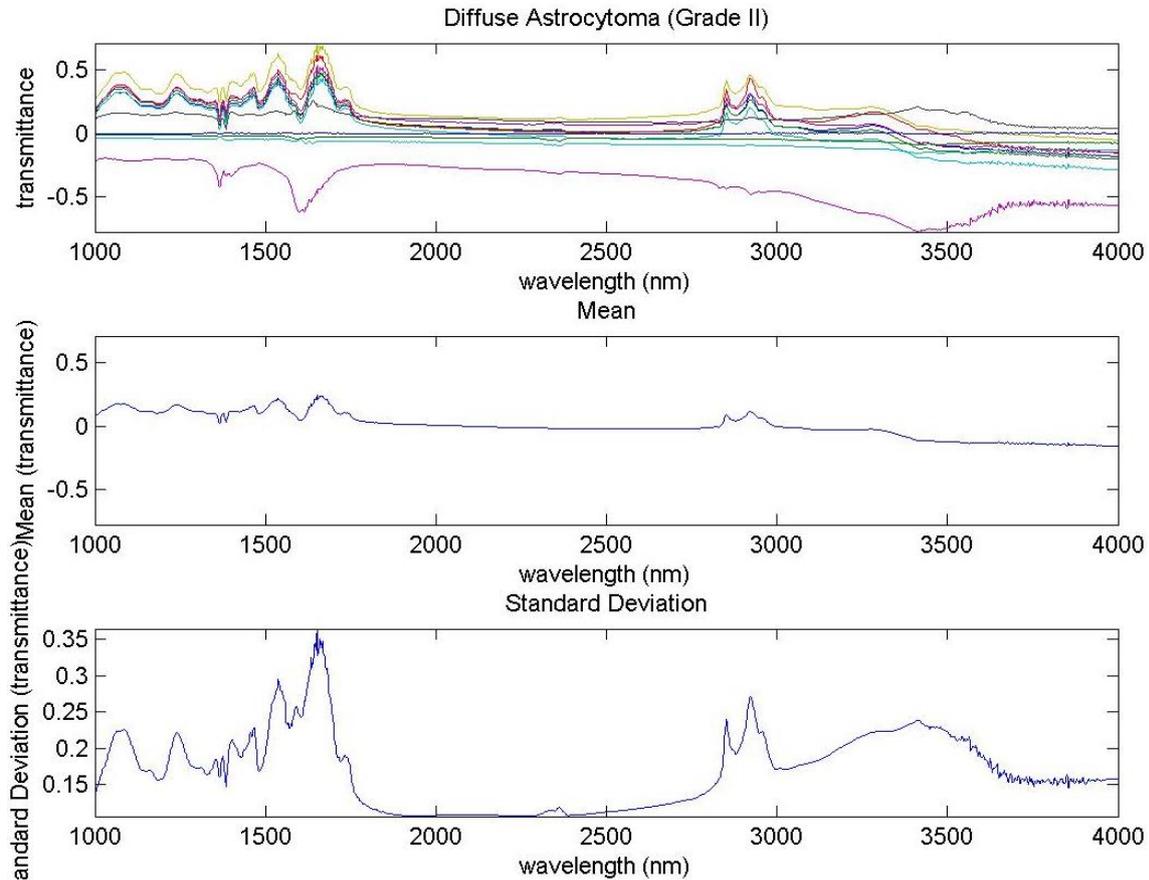


Figura 4.17: Firmas espectrales, media y desviación típica de las muestras de astrocitoma difuso (Grado II).

Aunque todas las muestras de la Figura 4.17 comparten el mismo diagnóstico, puede observarse a simple vista que existe una agrupación de firmas espectrales diferenciable al resto. Concretamente, las firmas de la gráfica que presentan la misma forma y los valores más altos de transmitancia. Se ha comprobado que esa forma común la presentan las 6 muestras obtenidas del centro del tumor (se especifica en el tipo de tejido de estas muestras). No obstante, todas las firmas de esta diagnosis pertenecen al mismo paciente (identificado con el número de paciente 20).

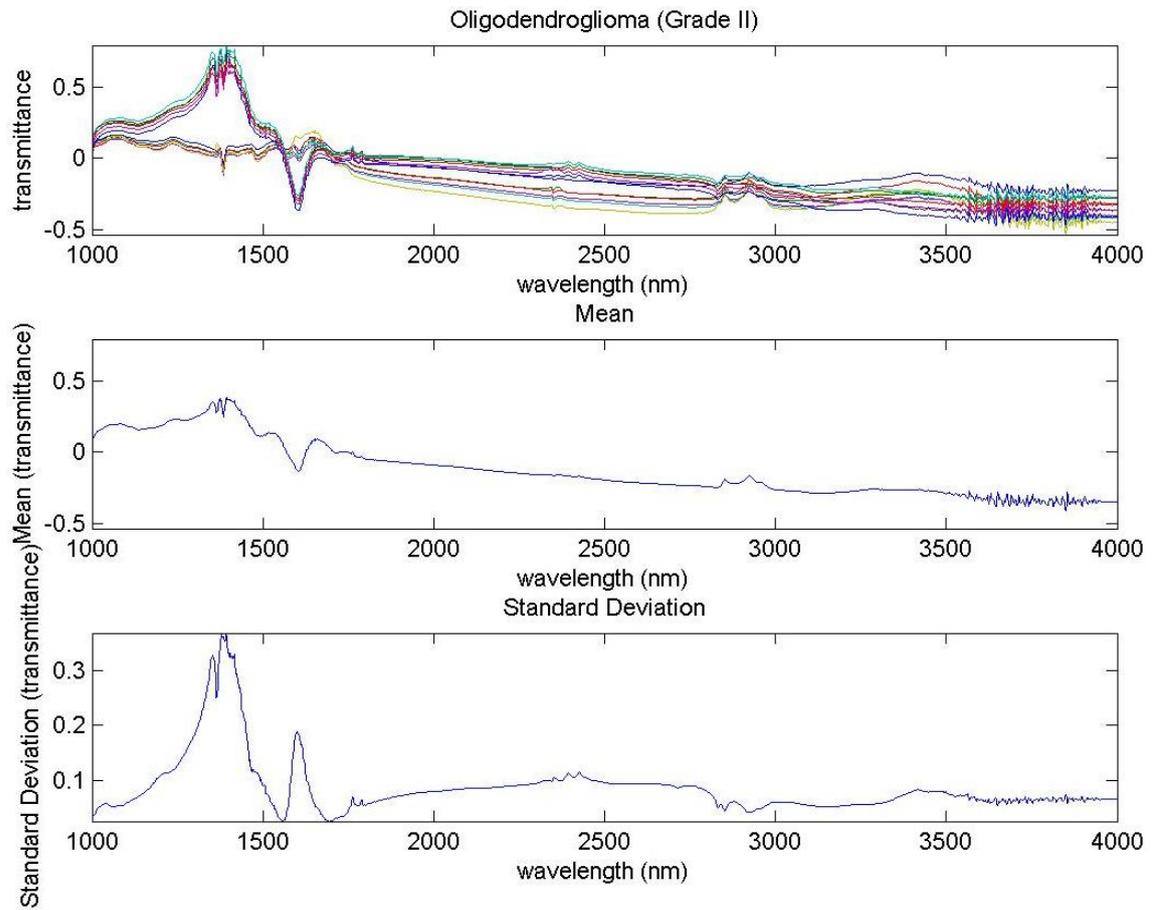
- *Oligodendroglioma*

Figura 4.18: Firmas espectrales, media y desviación típica de las muestras de oligodendroglioma (Grado II).

En la Figura 4.18 también se puede distinguir entre dos agrupaciones de firmas espectrales bien diferenciadas. En este caso, además de que las agrupaciones proceden de dos pacientes diferentes, las muestras de uno de ellos fueron tomadas del centro del tumor (paciente 25), mientras que para el otro paciente (paciente 11) no se especifica la zona (la descripción del tipo de tejido es “*Tumour*”).

- Firmas espectrales, media y desviación típica de las muestras de tumor primario (grado III):
 - *Astrocitoma anaplásico*

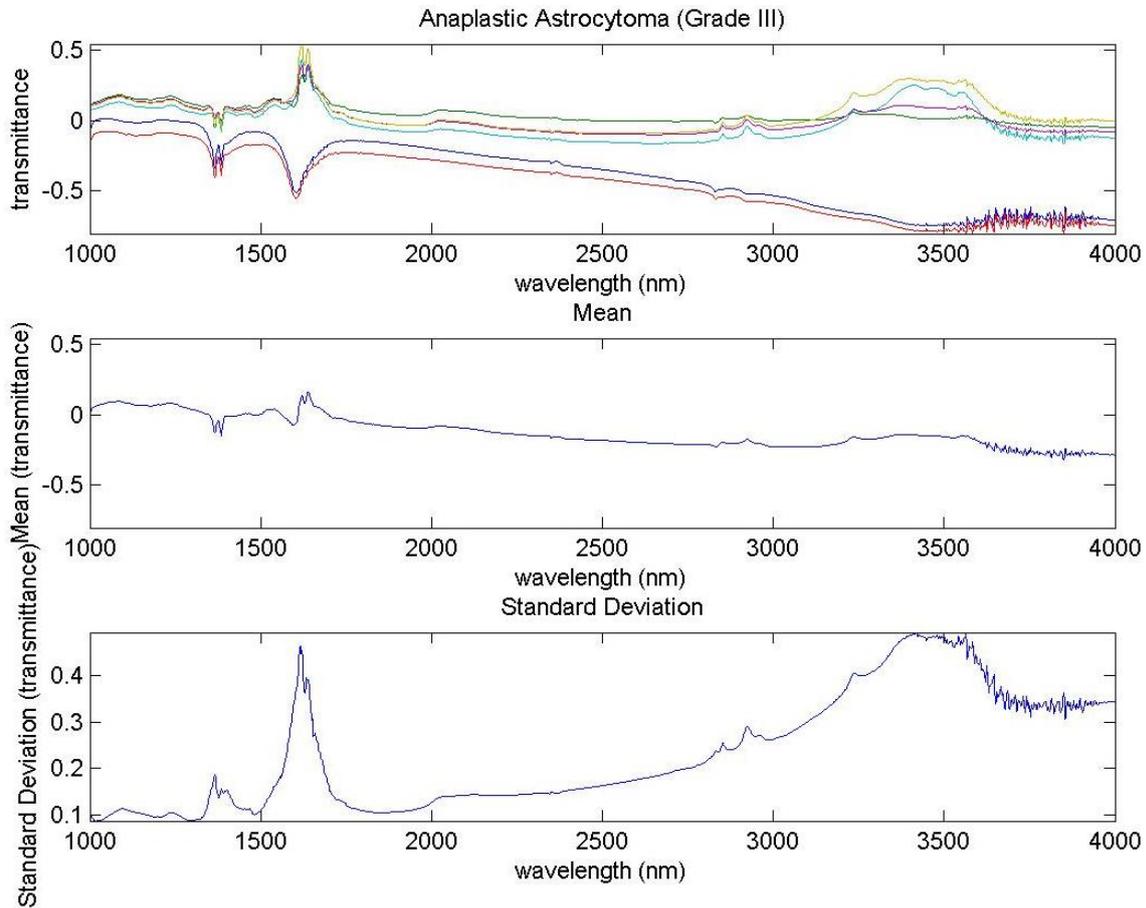


Figura 4.19: Firmas espectrales, media y desviación típica de las muestras de astrocitoma anaplásico (Grado III).

En el caso de las seis únicas muestras que presentan el tumor de grado III (Figura 4.19), también se distinguen claramente dos agrupaciones diferenciadas. Sin embargo, no se tiene conocimiento acerca de dichas diferencias puesto que todas estas muestras proceden del mismo paciente (paciente 9), y todas tienen las mismas especificaciones de tejido (sustancia gris). Tampoco se puede cotejar que la diferencia tenga relación con su datación puesto que en su descripción se indica que todas fueron procesadas el mismo día.

- Firmas espectrales, media y desviación típica de las muestras de tumor primario (grado IV):
 - Etiqueta: “GBM with oligo”

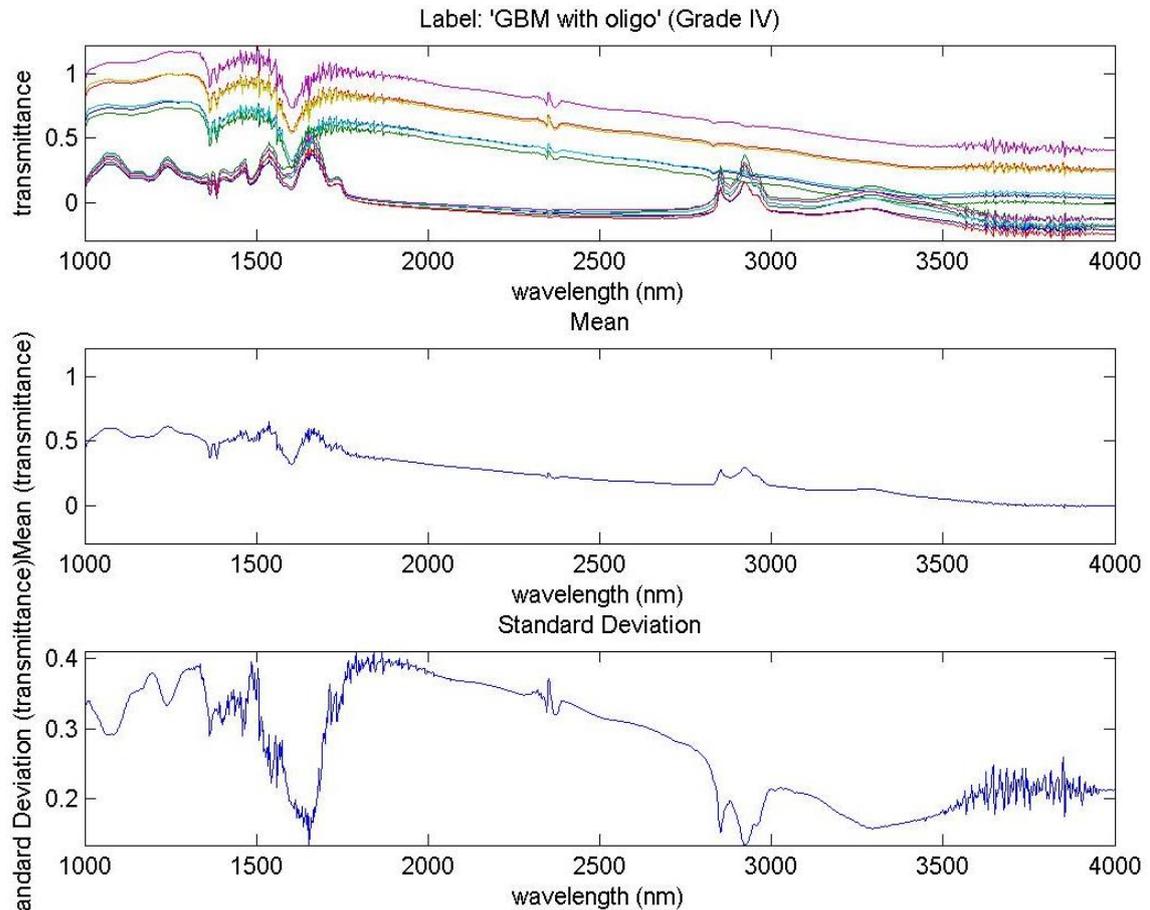


Figura 4.20: Firmas espectrales, media y desviación típica de las muestras de “GBM with oligo” (Grado IV).

Las dos agrupaciones de firmas espectrales que se aprecian en la Figura 4.20, provienen de pacientes diferentes. Sin embargo, en la descripción del tipo de tejido, en ambos pacientes se especifica “*Mixed normal*”.

- *Glioblastoma multiforme*

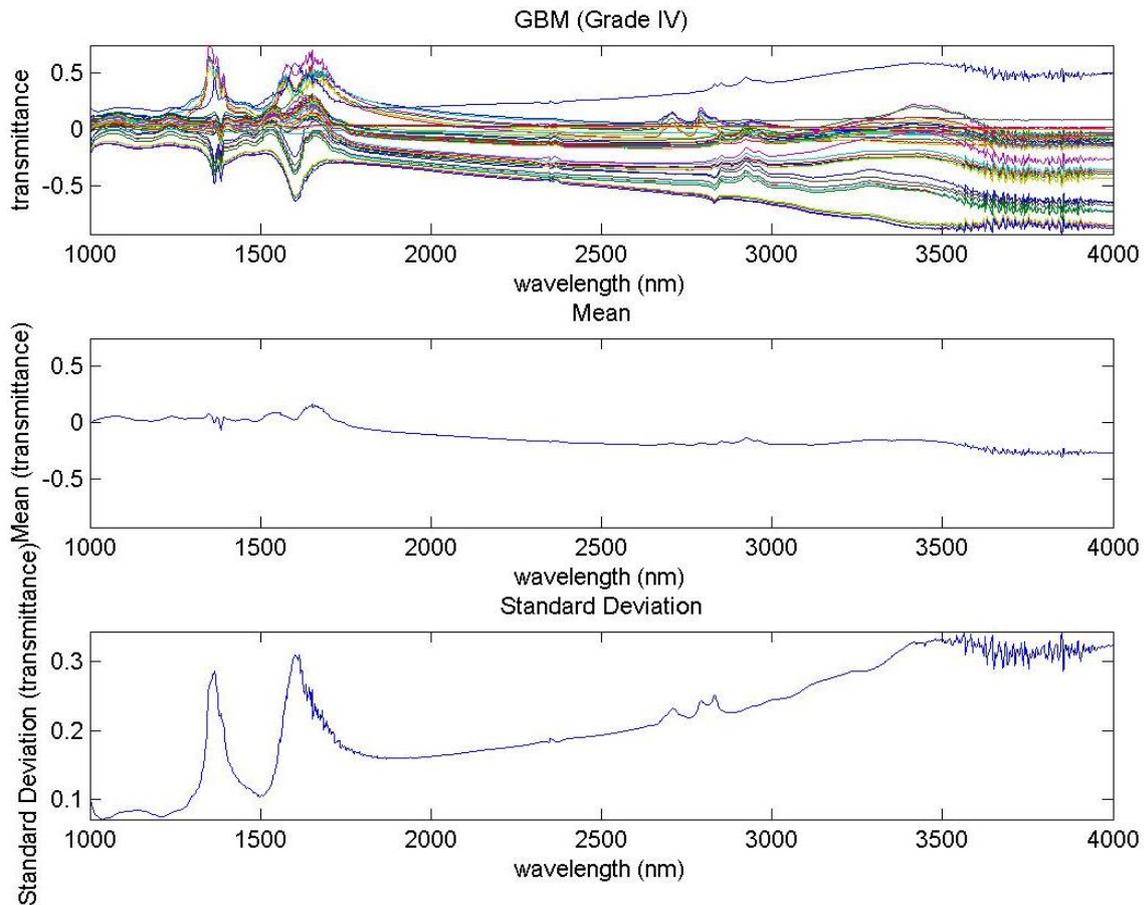


Figura 4.21: Firmas espectrales, media y desviación típica de las muestras de glioblastoma multiforme (Grado IV).

Dado que para el tumor glioblastoma multiforme tenemos muchas muestras, y por tanto es más complicado distinguir sus firmas espectrales en una sola gráfica, se presentan a continuación en sus diferentes agrupaciones.

Las muestras de glioblastoma multiforme se han obtenido de seis pacientes diferentes. Las seis gráficas de la Figura 4.22 se corresponden a cada uno de sus respectivos pacientes, en estas se especifica el número de identificación del paciente, así como la descripción que especificaron los médicos sobre el tipo de tejido. En el último caso, puede observarse que las muestras son tomadas de sustancia gris en los márgenes del tumor.

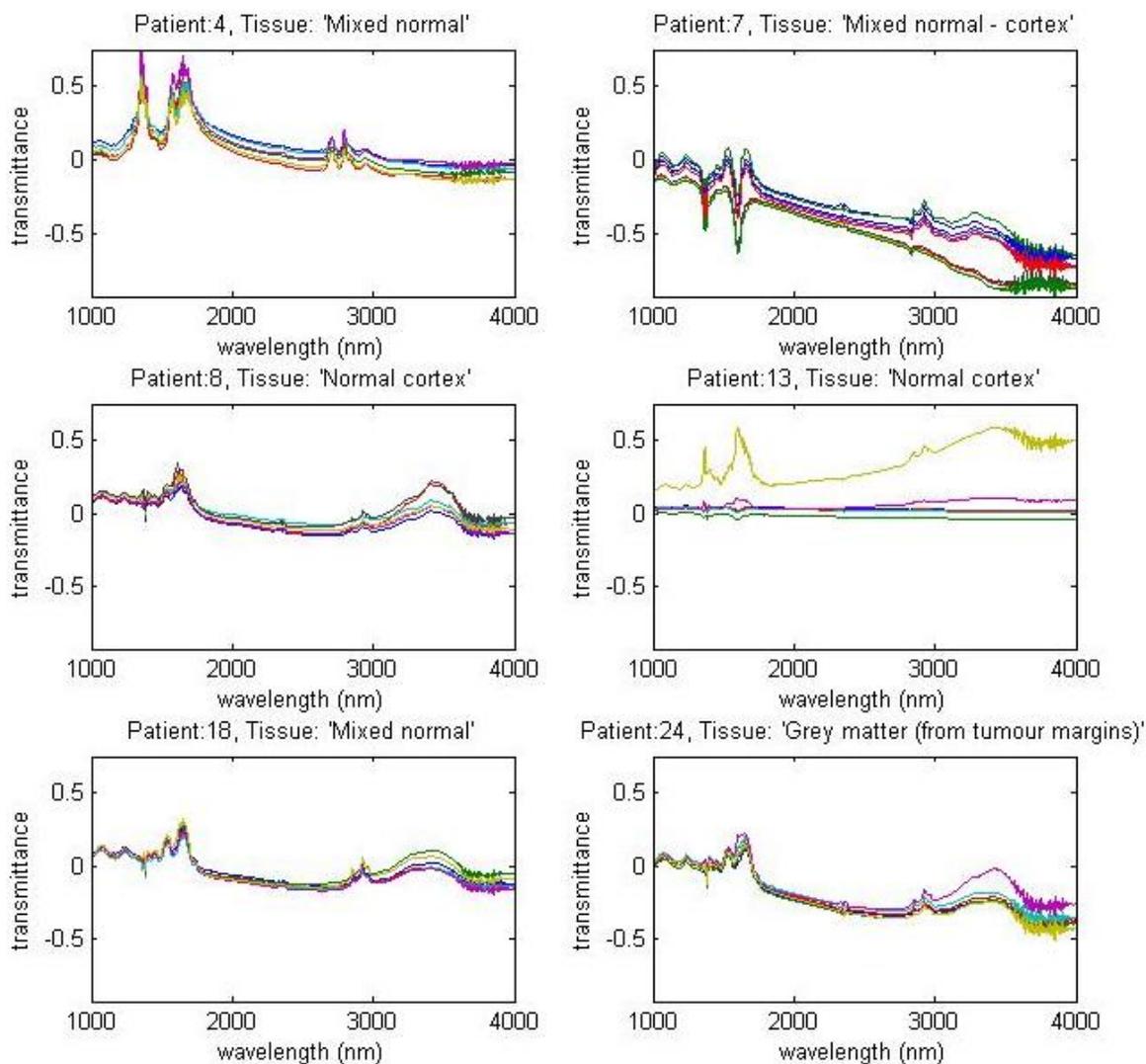


Figura 4.22: Firmas espectrales de las muestras de glioblastoma multiforme (Grado IV) por paciente.

- Firmas espectrales, media y desviación típica de las muestras de tumor secundario:

- *Adenocarcinoma pulmonar*

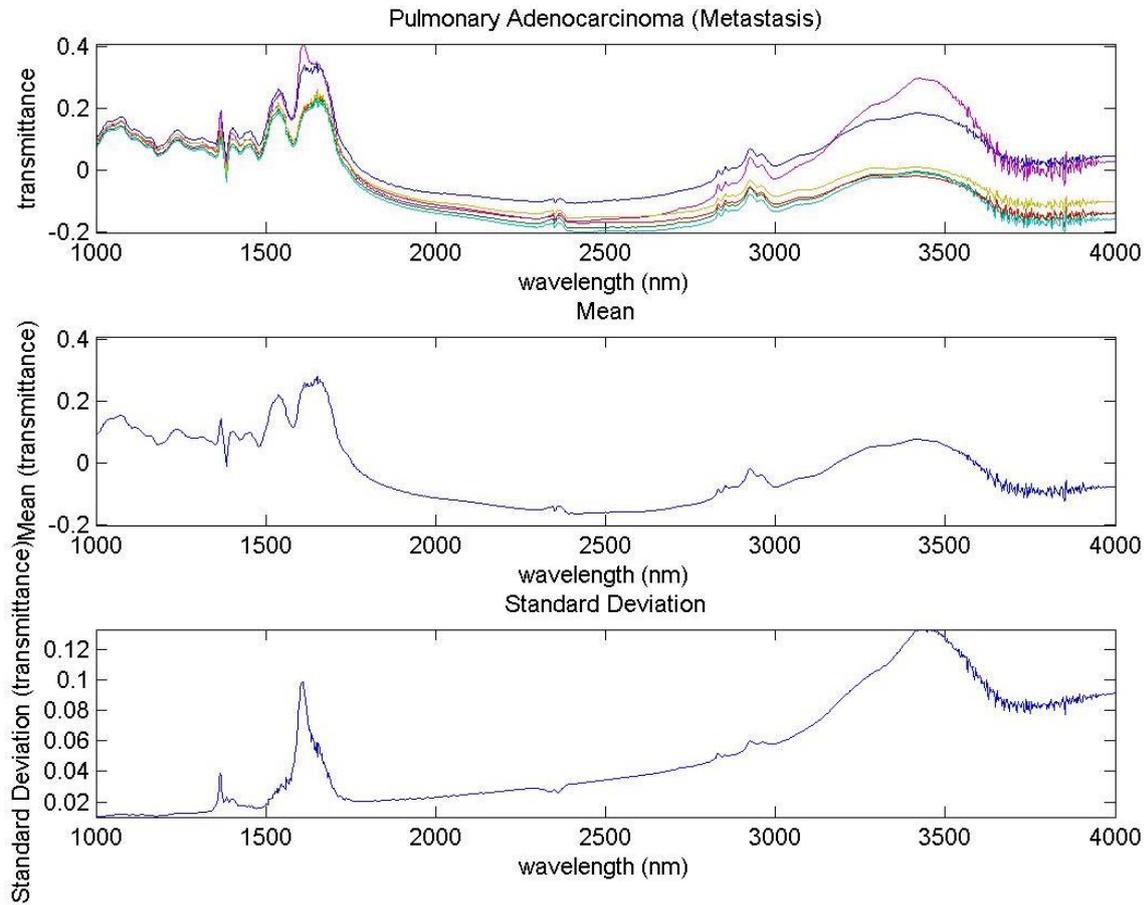


Figura 4.23: Firmas espectrales, media y desviación típica de las muestras de adenocarcinoma pulmonar (Metástasis).

Todas las muestras del adenocarcinoma pulmonar provienen del mismo paciente, y se describe en el tipo de tejido de todas ellas, que las muestras son tomadas de la zona central del tumor.

- *Carcinoma gástrico*

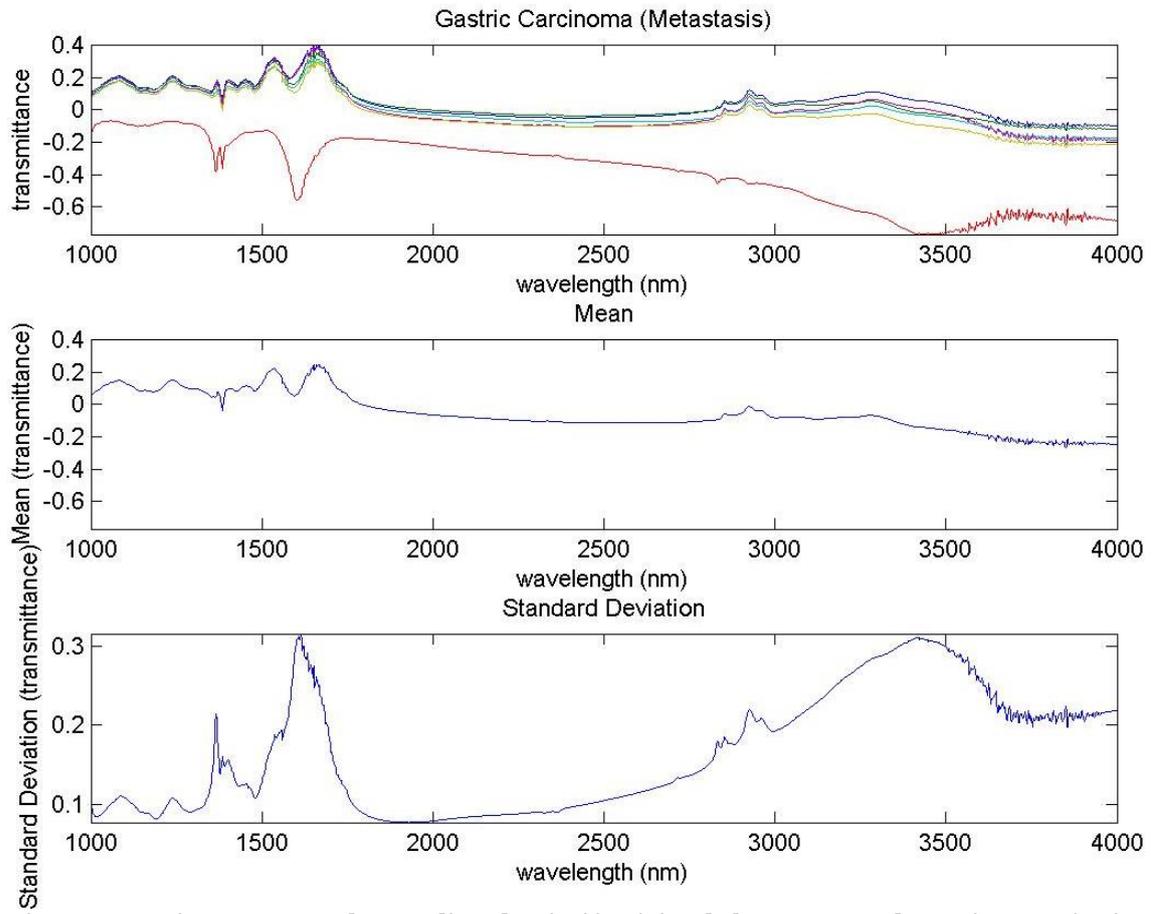


Figura 4.24: Firmas espectrales, media y desviación típica de las muestras de carcinoma gástrico (Metástasis).

Al igual que en las muestras del adenocarcinoma pulmonar, las muestras de carcinoma gástrico pertenecen a un mismo paciente y son tomadas de la zona central del tumor.

- Firmas espectrales, media y desviación típica de las muestras de tejido no tumoral:
 - *Epilepsia con esclerosis del hipocampo.*

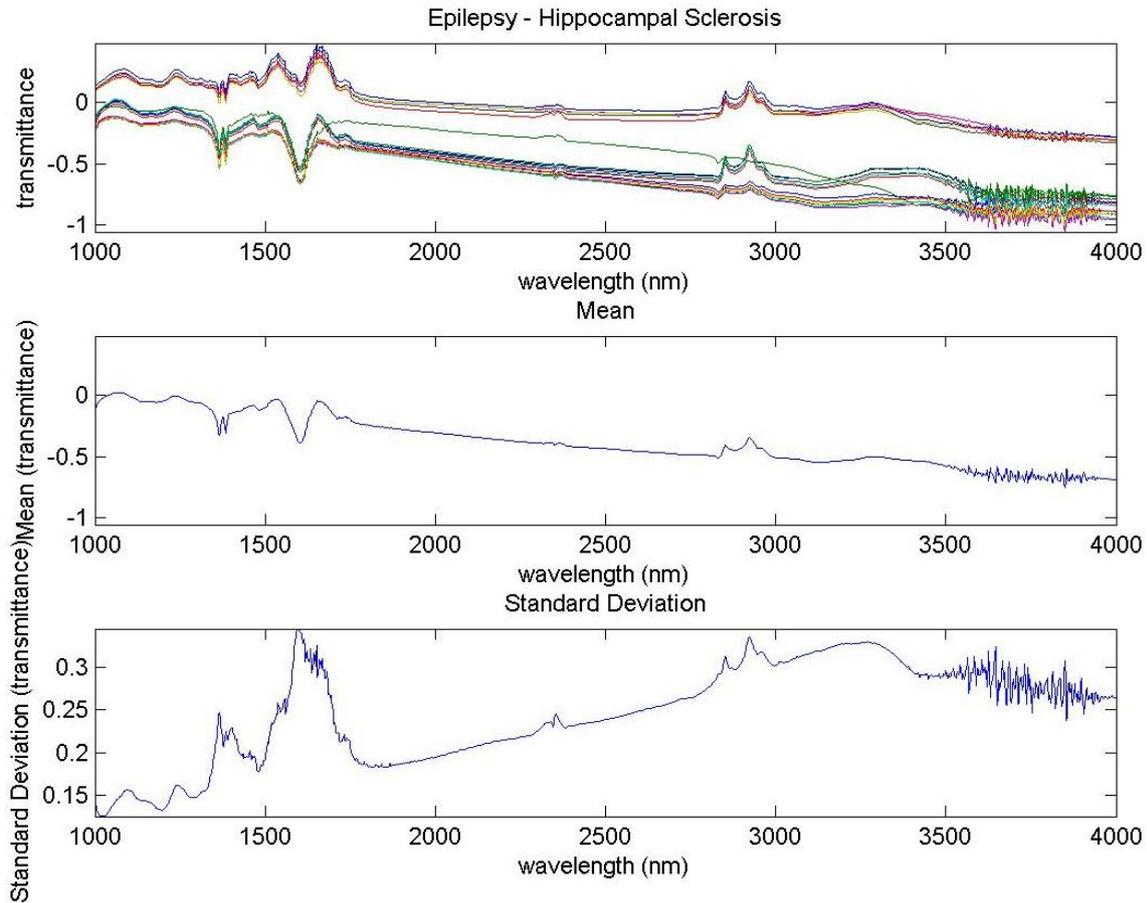


Figura 4.25: Firmas espectrales, media y desviación típica de las muestras de epilepsia con esclerosis del hipocampo (no tumoral).

Las muestras no tumorales se han tomado de dos pacientes con este tipo de epilepsia, los cuales se identifican con los números de paciente 17 y 21 respectivamente. Para el tipo de tejido del paciente 21 se ha especificado en todas las muestras “*Mixed normal*”. Los dos pacientes presentaron la misma patología.

En la Figura 4.26 se ilustran las firmas espectrales de estas muestras separadas en tres grupos. Para el paciente identificado con el número 17 se distinguen seis muestras con tejido de sustancia gris (gráfica central) y cinco de sustancia blanca (gráfica superior). En la gráfica inferior se representan las firmas espectrales tomadas del paciente 21, cuya descripción de tejido es “*Mixed normal*”.

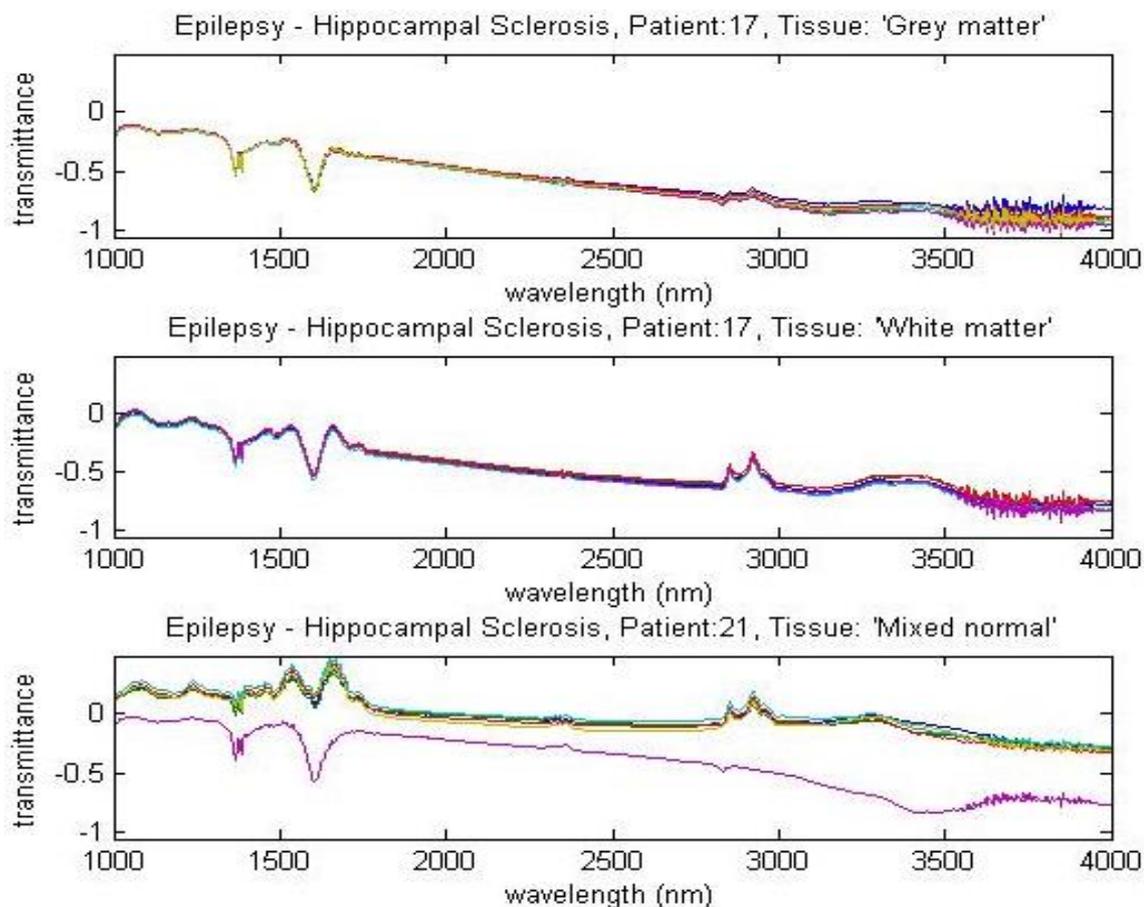


Figura 4.26: Firmas espectrales de las muestras no tumorales (pacientes 17 y 21).

4.3 Resumen

Las muestras proporcionadas por el Hospital de Southampton se dividen en dos campañas: campaña 1 y campaña 2. Las muestras de la campaña 1 sirvieron como soporte muestral a un proyecto anterior [4], y tanto esta campaña como la campaña 2 conforman la base de datos empleada en el desarrollo del presente PFC.

Las muestras de ambas campañas provienen de diferentes pacientes y fueron obtenidas en sus respectivas craneotomías. Se trata de muestras de diferentes tipos de tejido resecadas de tumores primarios, de los cuatro grados establecidos por la WHO, de dos tipos de metástasis, así como de tejido sano o no tumoral. Tras la resección de las muestras de tejido, estas fueron analizadas y clasificadas por un patólogo, etiquetadas según su diagnóstico, tipo de tejido y fecha; y luego, sometidas a un proceso de preparación. En este proceso se lavaron los rastros de sangre de algunas de ellas por medio de una solución salina, se pesaron, se secaron y se molieron en un mortero con bromuro potásico. Esta mezcla resultante fue dividida en porciones de 0,5 gramos aproximadamente a las que se le

aplicó presión en una prensa de vacío para crear una muestra sólida. Estas muestras se almacenaron individualmente en una bolsita de gel de sílice a -20°C . Posteriormente se procedió a captar los datos espectrográficos de las muestras. Para cada campaña se empleó un espectrómetro diferente.

Los datos espectrográficos de la campaña 1 fueron obtenidos en absorbancia y presentan 2906 bandas espectrales (desde 400 nm a 6000 nm), mientras que los de la campaña 2 constan de 1501 bandas (desde 1000 nm a 4000 nm), y la forma de sus firmas espectrales revela que fueron obtenidos en transmitancia.

En este PFC se mantiene el conjunto de muestras de la campaña 1 que se empleó en el proyecto anterior [4] tras la eliminación de muestras no válidas que tuvo lugar en dicho proyecto. No obstante, se ha añadido a este conjunto la única muestra de grado I con la que cuenta la campaña 1 en las clasificaciones de la fusión de ambas campañas. Por otro lado, en el presente PFC se ha decidido no prescindir de ninguna de las muestras de la campaña 2, por lo que la base de datos final empleada en el proyecto cuenta con 182 muestras de la campaña 1 y 128 muestras de la campaña 2.

Capítulo 5

Método desarrollado

5.1 Introducción

El principal objetivo de este PFC ha sido diseñar un sistema de clasificación que proporcione, a partir de un dato espectrográfico de una muestra de tejido cerebral, un diagnóstico médico. Concretamente, se desea que el sistema de clasificación sea capaz de identificar si se trata de un tejido tumoral o no tumoral, y en el caso de que se trate de un tejido tumoral, discernir si se trata de metástasis o de tumor primario, y en este último caso, identificar el grado del tumor. Teóricamente este tipo de distinciones son conocidas como clases en los algoritmos de clasificación.

Se han diseñado varios sistemas de clasificación empleando diferentes métodos basados en Máquinas de Vectores Soporte (*SVMs*). Estos sistemas de clasificación se han desarrollado aprovechando, tanto en sus etapas de entrenamiento como en sus etapas de test, los diferentes conjuntos de muestras de la base de datos presentada en el Capítulo 4.

Los algoritmos implementados se pueden dividir en dos métodos de clasificación: árboles de clasificación jerárquica y clasificadores multiclase. Ambos métodos se han explotado para clasificar las muestras de la campaña 1 y las de la campaña 2, sin embargo,

las clasificaciones de la fusión de las campañas se han realizado solamente utilizando sistemas multiclase. Esto se ha hecho así puesto que los sistemas multiclase ofrecieron mejores resultados que los clasificadores jerárquicos basados en árboles de decisión, en las clasificaciones de las campañas por separado.

Se realizaron transformaciones de datos de absorbancia a transmitancia y viceversa, para luego realizar clasificaciones de los mismos datos en ambas magnitudes y cotejar los resultados.

Adicionalmente, se crearon diferentes modelos de clasificadores haciendo uso de diferentes *kernels* para el clasificador SVM, y empleando métodos de optimización de parámetros.

5.2 Cadena de procesamiento de los datos espectrográficos

Una vez se dispone de la base de datos cedida por el Hospital de Southampton, se realiza un procesamiento de las muestras en el que estas se ven sometidas a diferentes tipos de operaciones matemáticas que afectan a la forma de las firmas espectrales hasta obtener una serie de resultados finales. En este PFC, la cadena de procesamiento aplicada a los datos espectrográficos presentados en el Capítulo 4 puede dividirse en tres etapas: selección de datos de entrada, preprocesado y clasificación. Esta cadena se muestra en la Figura 5.1.

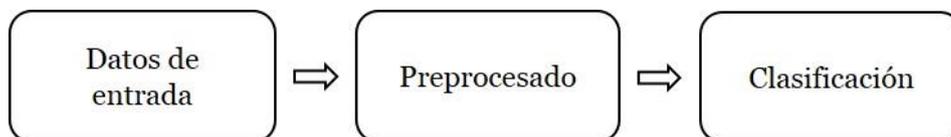


Figura 5.1: Etapas de la cadena de procesamiento empleada en este proyecto.

5.2.1. Etapa de datos de entrada

En la primera etapa se estudia el conjunto inicial de las muestras disponibles en ambas campañas y se hace una selección de las muestras válidas. Además, las muestras válidas se asignan a diferentes conjuntos de datos dependiendo de las clases que se verán involucradas en la etapa de clasificación de esta cadena. Estas agrupaciones o clases se han presentado en la Tabla 4.2 y la Tabla 4.3 del Capítulo 4, y como puede observarse, hay clases que pueden dividirse en otras subclases. Por ejemplo, la clase “muestras tumorales” de la campaña 1 se divide en dos clases, “grado alto” y “grado bajo”, y la clase “grado alto” a su vez se divide en dos clases finales: “grado III” y “grado IV”.

Por otra parte, como se ha comentado a lo largo de esta memoria, también se desarrollaron sistemas multiclase para clasificar la fusión de las campañas 1 y 2. Cuando se especifica “fusión de las campañas”, simplemente se está haciendo referencia a un conjunto de datos que contiene todas las muestras válidas de ambas campañas, y que, además, incluye la única muestra de grado I de la campaña 1.

Por tanto, a la hora de plantearse establecer las clases de este nuevo conjunto de datos, se tuvieron que considerar todas las clases finales de ambas campañas (en el apartado 5.6 se explicará que en los sistemas de clasificación multiclase se emplean únicamente las clases finales). Es evidente que se necesitan las clases comunes a las dos campañas: “grado II”, “grado III” y “grado IV”; como en este nuevo conjunto también se dispone de muestras de grado I y muestras de metástasis, se requieren sus clases respectivas: “grado I” y “tumor secundario”; finalmente las muestras de “sustancia blanca”, “sustancia gris” y “no tumorales” se agrupan en una única clase que también se denominó “no tumorales”. Esto último se hizo de este modo porque, como ya se ha comentado, las muestras no tumorales de la campaña 2 presentaban una mezcla de tejidos. Plantearse dividir en tres clases diferentes las muestras no tumorales, como por ejemplo: “sustancia blanca”, “sustancia gris” y “epilepsia”; no es especialmente relevante en este trabajo ya que, además de salirse del ámbito de los tumores cerebrales, se estaría complicando innecesariamente el modelo SVM para los sistemas de clasificación multiclase, y no se dispondría en este tipo de sistemas, de la robustez en la clasificación de la clase “no tumoral”, al dividirse las muestras de esta clase en subclases.

Tomando en cuenta estas consideraciones, se ha establecido que el conjunto de muestras de la fusión de campañas, se agrupe en las mismas clases que se habían planteado para la campaña 2.

5.2.2. Etapa de preprocesado

En el siguiente paso de la cadena de procesado, se realiza un preprocesado de los datos espectrográficos catalogados como válidos en la primera etapa. En esta fase de preprocesado se realizan diferentes transformaciones de las muestras según la magnitud y forma de las firmas espectrales.

Por un lado, se realiza una normalización de todas las muestras entre los valores 0 y 1 de sus respuestas en absorbancia o transmitancia y, por otro lado, se obtienen todas las firmas espectrales de los diferentes conjuntos en ambas magnitudes. Es decir, las firmas

espectrales de la campaña 1 (obtenidas en absorbancia) se transforman para obtenerse su respuesta en transmitancia; de las firmas espectrales de la campaña 2 (obtenidas en transmitancia), se obtiene su respuesta en absorbancia; y las firmas espectrales de la fusión de las campañas se disponen en absorbancia y transmitancia.

En la etapa de preprocesado también se necesita realizar una adaptación de las muestras de la campaña 1 a la hora de fusionar las campañas.

Esta adaptación consiste en reajustar la dimensión espectral de la campaña 1 puesto que las muestras de esta campaña no comparten ni el rango ni las bandas espectrales de las muestras de la campaña 2. Como se explicó en el apartado 4.2 de esta memoria, las muestras de la campaña 1 tienen un rango de longitudes de onda que va desde 400 nm a 6000 nm (con resolución espectral de 1,93 nm), mientras que el espectro de las muestras de la campaña 2 va desde 1000 nm a 4000 nm (con resolución espectral de 2 nm). Por tanto, se limitó el rango de las muestras de la campaña 1 al rango que comparte con las muestras de la campaña 2, y se realizó una interpolación *spline* cúbica para obtener las firmas espectrales de la campaña 2 en las bandas de la campaña 1. Se decidió emplear este tipo de interpolación porque en la literatura se comenta que ofrece buenos resultados, no obstante, tras aplicarse esta operación a las muestras de la campaña 1 se revisaron sus firmas espectrales, comprobando que los resultados eran satisfactorios para nuestro estudio. Una vez adaptadas las muestras de la campaña 1 se tuvieron que volver a normalizar.

Al final del proceso de adaptación, las muestras de la fusión de las campañas se disponen en el rango 1000-4000 nm con una resolución espectral de 2 nm.

De este modo, al final de esta etapa se dispone de las muestras de las campañas por separado y las muestras de la fusión de las campañas, normalizadas tanto en absorbancia como en transmitancia.

Adicionalmente, en esta etapa se aplica un filtrado tipo *smooth* a algunas firmas espectrales con el fin de eliminar el alto nivel de ruido que presentan. Esto se ha hecho porque, aunque los sistemas SVM presentan una cierta tolerancia al ruido, altos niveles de ruido en las muestras pueden producir problemas en la clasificación.

5.2.3. Etapa de clasificación

Finalmente, una vez se dispone de los datos espectrográficos preprocesados y correctamente asignados a sus respectivas clases, se pasan a la etapa de clasificación. En esta etapa de la cadena se realiza el proceso de clasificación de los datos y se obtienen las métricas que indican el grado de acierto en las clasificaciones. La etapa de clasificación se explicará en detalle en los siguientes apartados.

5.3 Proceso de clasificación

En el proceso de clasificación, la base de datos se divide en dos conjuntos aleatorios: muestras de entrenamiento y muestras de validación o test. A partir del conjunto de muestras de entrenamiento y de las etiquetas que identifican sus respectivas clases, se genera un modelo de clasificador SVM. Posteriormente, este modelo se valida introduciendo las muestras restantes (conjunto de muestras de test). Después de realizar la clasificación de las muestras de test, se cotejan los resultados obtenidos con sus correspondientes etiquetas. Finalmente, se calculan las métricas de validación del clasificador. En la Figura 5.2 se presenta en detalle el proceso de clasificación de los datos espectrográficos.

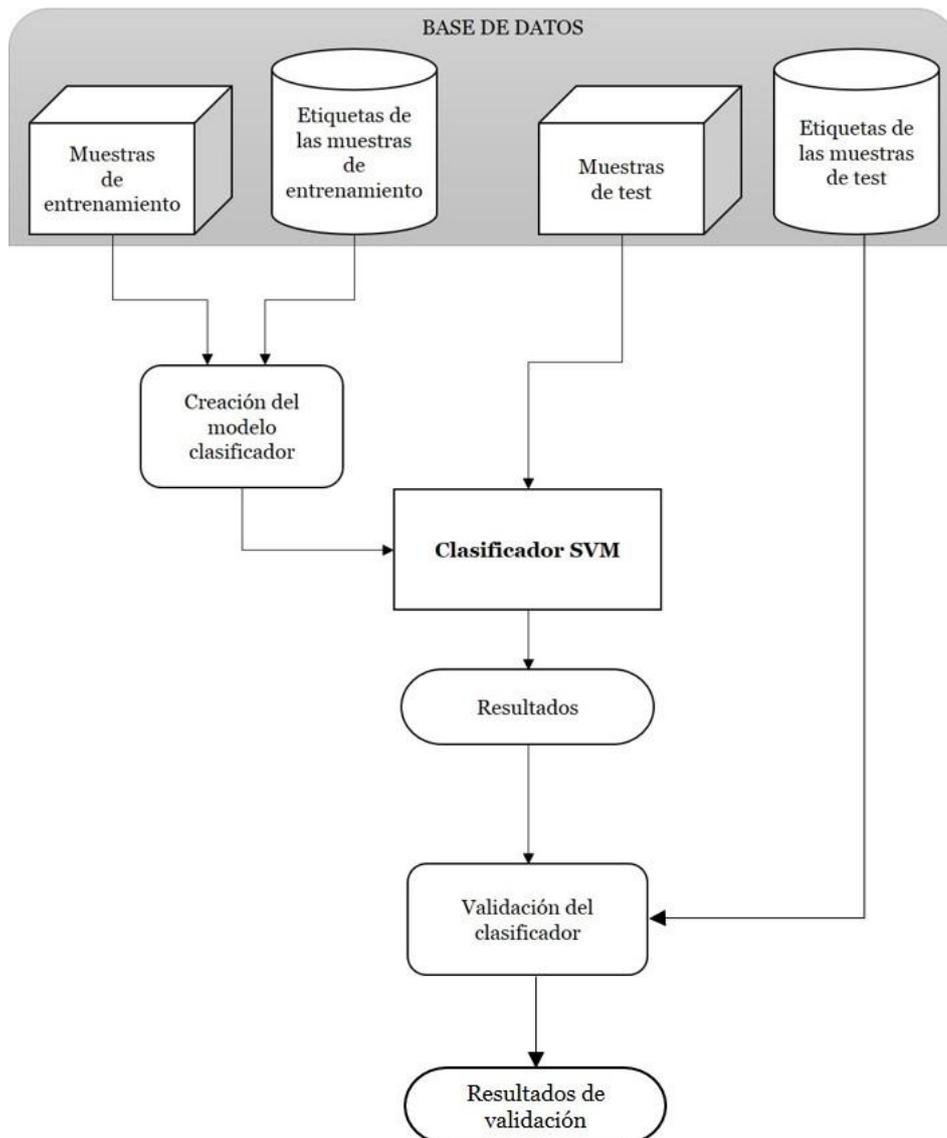


Figura 5.2: Proceso de clasificación.

Debe tenerse en cuenta la necesidad de repetir el proceso de clasificación un número de veces considerable para que el modelo de clasificación sea válido. Esto se debe a que, si se realiza el proceso una sola vez, puede darse el caso de que el conjunto de muestras de entrenamiento escogido aleatoriamente no sea representativo de la diversidad de muestras de cada clase, dando lugar a un modelo SVM que falsea los datos de validación del método. Hay que aclarar que un conjunto de muestras no representativo podría ser, un conjunto de muestras *outliers* que no ofrece vectores soportes apropiados para establecer un buen hiperplano de separación; o en contraste, un conjunto de muestras cercanas a los centroides de los clústeres, que no ofrecen información sobre la dispersión de las muestras de cada clase. Al repetir el proceso de clasificación muchas veces, con diferentes conjuntos de

entrenamiento, la probabilidad de escoger conjuntos de muestras no representativos se reduce, es decir, la variedad de las muestras permite que los resultados de validación sean más fiables. En otras palabras, replicar el proceso de clasificación un número considerable de veces con diferentes conjuntos de muestras, reduce la dependencia del conjunto de muestras que se ha empleado, que presenta el sistema clasificador.

Existen varias técnicas a la hora de establecer los diferentes conjuntos de entrenamiento y test. Entre ellas destaca la *validación cruzada (cross-validation)*. Esta técnica consiste en realizar particiones del conjunto total de muestras, de manera que, de cada una de las particiones se establezca un conjunto de entrenamiento y un conjunto de test, para realizar el proceso de clasificación. Una vez se obtienen los resultados de validación para cada partición, se realiza la media de dichos resultados. Aunque esta técnica es muy utilizada y en general ofrece buenos resultados, se ha descartado en el desarrollo de este PFC debido a que en algunas clases de la base de datos se dispone de una cantidad muy reducida de muestras. Por tanto, si se tuviera que subdividir estos conjuntos de muestras para cada una de las particiones de la validación cruzada, a la hora de realizar cada repetición en el proceso de clasificación se contaría con muestras insuficientes para las clases con poca población, con lo que el modelo SVM generado no sería robusto.

La idea que subyace tras la técnica empleada en el presente PFC, para establecer los diferentes conjuntos de entrenamiento y test, es permitir que se repitan algunas muestras en los conjuntos establecidos para cada repetición del proceso de clasificación. La técnica consiste en fijar un porcentaje de muestras de test y otro porcentaje de muestras de entrenamiento sobre el conjunto total de muestras, de manera que las muestras del conjunto de entrenamiento se seleccionen de manera aleatoria, y el conjunto de muestras restantes conformen el conjunto de test. Este algoritmo se repite el número de veces que se desea realizar el proceso de clasificación, sin tener en cuenta las muestras, de test y entrenamiento, seleccionadas en cada repetición.

El hecho de escoger las muestras aleatoriamente en cada repetición del proceso de clasificación, supone que en las distintas repeticiones existan muestras de entrenamiento o test, que fueron o serán empleadas en otras repeticiones del proceso, lo que contrasta con la técnica de *cross-validation*. Sin embargo, esta técnica permite disponer de un mayor número de muestras de entrenamiento (y también de test), para generar el modelo de clasificación.

En general, en este PFC se estableció el 70% de muestras de entrenamiento frente al 30% de muestras de test por tratarse de valores comúnmente establecidos en este tipo de aplicaciones. No obstante, se realizaron algunas pruebas con otros porcentajes con el fin de contrastar los resultados. Se debe aclarar que esos porcentajes se mantienen cuando se seleccionan las muestras de cada clase, es decir, se selecciona aleatoriamente el 70% de las muestras para cada una de las clases, conformando el conjunto de muestras de entrenamiento; y 30% restante de cada clase conforma el conjunto de muestras de test. Por otro lado, se decidió establecer en cincuenta el número de repeticiones del proceso de clasificación por considerarse un número lo suficientemente alto, y, al igual que en la técnica *cross-validation*, tras las repeticiones se realizan las medias de las métricas de validación.

Todos los sistemas diseñados en este proyecto se han implementado sobre el entorno de desarrollo integrado *MATLAB*[®]. Además, para realizar los experimentos de clasificación en este estudio se ha hecho uso de la librería “*LIBSVM*”. Esta librería ha sido desarrollada por Chih-Chung Chang y Chih-Jen Lin, y está disponible en internet en su versión 3.22 (diciembre de 2016) [109].

5.4 Métricas de validación

Las principales métricas utilizadas en la validación de los sistemas desarrollados en este PFC se han definido en el apartado 3.6 de la memoria. Se trata del porcentaje de acierto global (*Overall Accuracy*), la sensibilidad y la especificidad. Sin embargo, hay que destacar que ha sido necesario, para la elaboración de este proyecto, redefinir las métricas de sensibilidad y especificidad. Esto se debe a que en el ámbito médico se emplean únicamente en la clasificación de pacientes que presentan una enfermedad determinada, y pacientes que no la presentan; y en la línea de investigación en la que se enmarca este proyecto se realizan clasificaciones que van más allá de la situación de padecerse un tumor o no padecerlo.

Por ejemplo, en algunas situaciones se da el caso de tener que clasificar entre diferentes tipos o grados de tumor, o cuando debe distinguirse entre diferentes tipos de tejido sano.

Por lo tanto, sobre los resultados experimentales proporcionados en este PFC, la sensibilidad y la especificidad no deben entenderse, generalmente, como métricas que indican la proporción de muestras, enfermas y sanas respectivamente, correctamente clasificadas; sino como métricas que indican la proporción de muestras correctamente

casificadas como pertenecientes a una clase y la proporción de muestras clasificadas como no pertenecientes a la clase. Esto se explica más detalladamente a continuación.

Mientras que OA nos da una idea global de los datos correctamente clasificados, sin centrarse en el porcentaje de acierto en la clasificación de cada clase en concreto; la sensibilidad y la especificidad ofrecerán, respectivamente, una idea del porcentaje de acierto de la clasificación de una clase concreta (llamada en este proyecto *clase bajo test*), y una idea del porcentaje de acierto al clasificar muestras no pertenecientes a esa clase que se está tratando, es decir, el porcentaje de acierto al rechazar muestras que no pertenecen a la clase bajo test.

Para poder obtener la sensibilidad y la especificidad de la manera que aquí se plantea, se hace imprescindible recalcular la matriz de confusión tomando en consideración la clase bajo test que se está tratando. Esta matriz se muestra en la Tabla 5.1.

Tabla 5.1: Matriz de confusión considerando una clase concreta bajo test.

		CONDICIÓN REAL	
		Condición Positiva	Condición Negativa
CONDICIÓN PREDICHA	Población Total		
	Predicción Positiva	Verdaderos Positivos (V_p)	Falsos Positivos (F_p)
	Predicción Negativa	Falsos Negativos (F_n)	Verdaderos Negativos (V_n)

A diferencia de la matriz de confusión de la Tabla 3.3, esta matriz de confusión no se construye en base a una prueba diagnóstica que ofrece resultados positivos en enfermos y negativos en sanos. En este caso, los resultados se etiquetan como positivos (p) cuando son considerados como pertenecientes a la clase bajo test, y negativos (n) cuando son considerados como pertenecientes a otra clase diferente.

Este problema de predicción también ofrece cuatro posibles resultados a partir de la clasificación realizada. No obstante, en este caso se deben redefinir las entradas de la matriz en función de la clase bajo test, que puede ser, por ejemplo: tumor de grado II, metástasis, tumor de alto grado, sustancia blanca, no tumoral, etc. Si el resultado de la clasificación de una muestra es la clase que se está considerando, y la etiqueta real de dicha muestra

coincide con esa clase bajo test, entonces se considera como un *Verdadero Positivo* (V_p); sin embargo, si la etiqueta se corresponde a otra clase, se considerará como un *Falso Positivo* (F_p). De igual modo, tenemos un *Verdadero Negativo* (V_n) cuando, tanto la clasificación como la etiqueta no pertenecen a la clase bajo test y un *Falso Negativo* (F_n) cuando el resultado de la clasificación no es la clase bajo test pero la etiqueta de esa muestra coincide con la clase bajo test.

Una vez se construye la matriz de confusión para cada clase, con sus entradas definidas en función de las clasificaciones realizadas, considerando a cada una de las clases como clase bajo test, se pueden calcular la sensibilidad y la especificidad para cada clase de igual forma, es decir, se emplean las ecuaciones (3.40) y (3.41).

Por último, se debe aclarar que en los clasificadores biclase (en los sistemas basados en árboles de clasificación) en los que se discrimina entre muestras tumorales y muestras sanas, o muestras tumorales y muestras no tumorales; las definiciones de sensibilidad y especificidad coinciden con las empleadas en el ámbito médico ya que en estos casos se toma la clase tumoral como clase bajo test.

5.5 Árboles de clasificación

Los árboles de clasificación consisten en un método de clasificación jerárquica de varios niveles en los que, en cada nivel, se realizan clasificaciones binarias de diferentes clases o subclases. Es decir, a excepción del primer nivel, en cada nivel se realizan clasificaciones en subclases o subcategorías de las muestras previamente obtenidas en clasificaciones previas. Dependiendo de las clases predichas en esas clasificaciones anteriores, las muestras se pueden enviar a diferentes clasificadores subsiguientes, o se asignan a las clases finales si se ha llegado a las ramas finales del árbol de clasificación. En el desarrollo del presente proyecto se diseñaron árboles de clasificación para clasificar cada una de las campañas de datos.

A continuación, se presentan los sistemas de clasificación jerárquicos diseñados en este PFC. Primero se describirá la estructura diseñada para clasificar la campaña 1, y en el apartado subsiguiente, se tratará la estructura diseñada para clasificar la campaña 2.

5.5.1. Árbol de clasificación de la campaña 1

Para clasificar la campaña 1 se diseñó un sistema de clasificación que mantuviera la estructura del árbol diseñado en el proyecto anterior [4]. Sin embargo, hay que aclarar que se implementaron algoritmos propios para cotejar los resultados obtenidos en ambos proyectos. Debe recordarse además que, como se ha comentado en el apartado de antecedentes de este proyecto (apartado 1.2), el trabajo realizado en el proyecto anterior se centra únicamente en la clasificación de la campaña 1.

En la Figura 5.3 se ilustra el árbol de clasificación de tres niveles desarrollado para clasificar los datos de la campaña 1.

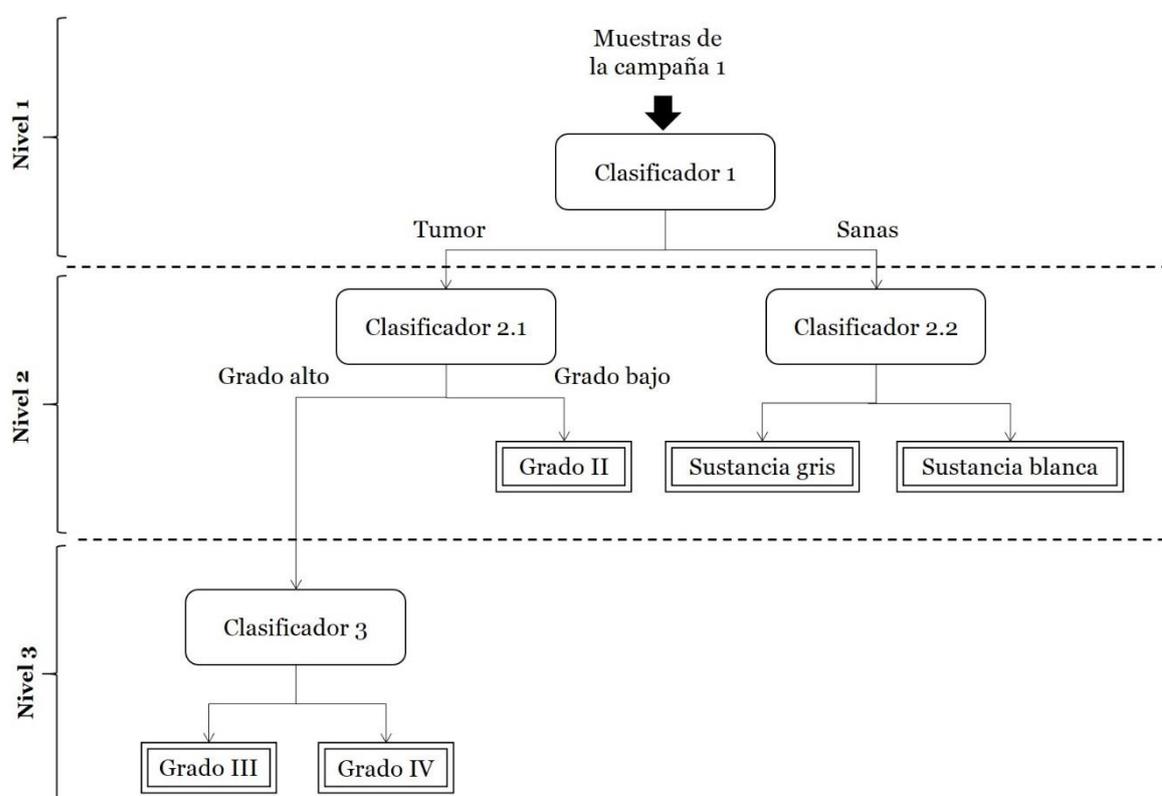


Figura 5.3: Árbol de clasificación de la campaña 1.

Como puede observarse, en el primer nivel de este clasificador jerárquico se realiza una clasificación biclase en un clasificador que se ha denominado *Clasificador 1*. Se trata del nivel más importante en un clasificador de este tipo puesto que, si se produce un error en este nivel, este se propaga hacia el resto de niveles dando lugar a errores seguros en los subsecuentes clasificadores. Además, este nivel de decisión es el más relevante tomando en consideración el ámbito médico, puesto que se distingue entre muestras sanas y muestras tumorales. Es decir, cometer un error al clasificar una muestra sana como muestra tumoral,

o viceversa, resulta más comprometido que, por ejemplo, cometer un error relacionado con la identificación del grado del tumor.

Los resultados de las clasificaciones del *Clasificador 1*, pasan a ser las muestras de entrada de los clasificadores del siguiente nivel del árbol de clasificación. Concretamente, las muestras clasificadas como tumores son consideradas como muestras de entrada al *Clasificador 2.1*, mientras que las muestras clasificadas como sanas, son consideradas como muestras de entrada del *Clasificador 2.2*.

El *Clasificador 2.1* se utiliza para discriminar entre tumores de grado bajo y tumores de grado alto. Teniendo en cuenta que para clasificar la campaña 1 no se empleó la única muestra de grado I que se dispone en esta campaña, las muestras clasificadas por este clasificador como grado bajo, son consideradas finalmente muestras de grado II. Aunque este clasificador no es tan relevante como el *Clasificador 1*, en la mayoría de los casos es de vital importancia conocer si el tumor se encuentra en un grado avanzado. Esto se debe a que dicha distinción proporciona información sobre el tratamiento a seguir por el paciente, y sobre su esperanza de vida. Además, los errores cometidos en este clasificador también se propagan al clasificador del siguiente nivel si son clasificados incorrectamente como grado alto.

En el nivel 2 de este árbol de clasificación también se presenta el *Clasificador 2.2*. En este clasificador se plantea distinguir entre muestras de sustancia blanca y muestras de sustancia gris, las muestras clasificadas en el nivel 1 como muestras sanas. A efectos de diagnóstico, este clasificador no resulta trascendental en este estudio ya que, en este punto, las muestras ya han sido clasificadas como sanas. Además, los errores cometidos por este clasificador no serán propagados puesto que finaliza una de las ramas del árbol, disponiéndose a su salida los resultados finales de las clases: sustancia blanca y sustancia gris. Se trata del clasificador menos relevante en este proyecto.

Finalmente, en el nivel 3 de este árbol de clasificación se dispone únicamente del *Clasificador 3*. En este último clasificador se introducen las muestras clasificadas como muestras de grado alto, las cuales se clasificarán para obtener como resultados finales: la clase de grado III y la clase de grado IV. Hay que tener en cuenta que este clasificador es el que presenta más posibilidades de errores de propagación, esto es porque, todos los errores que se cometan en los clasificadores 1 y 2.1, y que se propaguen por sus respectivas ramas izquierdas (Figura 5.3), generarán errores de propagación en este clasificador.

Para la implementación de este algoritmo se ha realizado un proceso de reetiquetado de las muestras clasificadas en las zonas de cambio de nivel del árbol. Esto es: tras la clasificación realizada por el *Clasificador 1*, las muestras serán etiquetada con un número distintivo de la clase “Tumor” o distintivo de la clase “Sanas” (al ser un árbol de clasificadores biclase en todos los casos se etiquetan como “1” ó “2”), y según el valor de esta etiqueta, las muestras se encauzarán al *Clasificador 2.1* o al *Clasificador 2.2* respectivamente; posteriormente, estas muestras se vuelven a etiquetar en función de cómo se hayan predicho en estos clasificadores; y así sucesivamente hasta que se alcancen las clases indivisibles de la campaña 1.

5.5.2. Árbol de clasificación de la campaña 2

A la hora de afrontar el diseño del sistema de clasificación de la campaña 2, empleando el método de clasificación jerárquica, fue necesario cambiar la estructura del árbol diseñado para la campaña 1. Esto se debe a que las clases de la campaña 2 difieren sustancialmente de las clases de la campaña 1. En este caso contamos con muestras de Grado I, muestras de metástasis, y en el caso de las muestras no tumorales, muestras cuyo tejido es una mezcla entre sustancia blanca y sustancia gris.

Además, a diferencia del árbol diseñado para la campaña 1, el árbol diseñado para clasificar la campaña 2 consta de cuatro niveles. En la Figura 5.4 se ilustra la estructura del sistema desarrollado para clasificar los datos de la campaña 2.

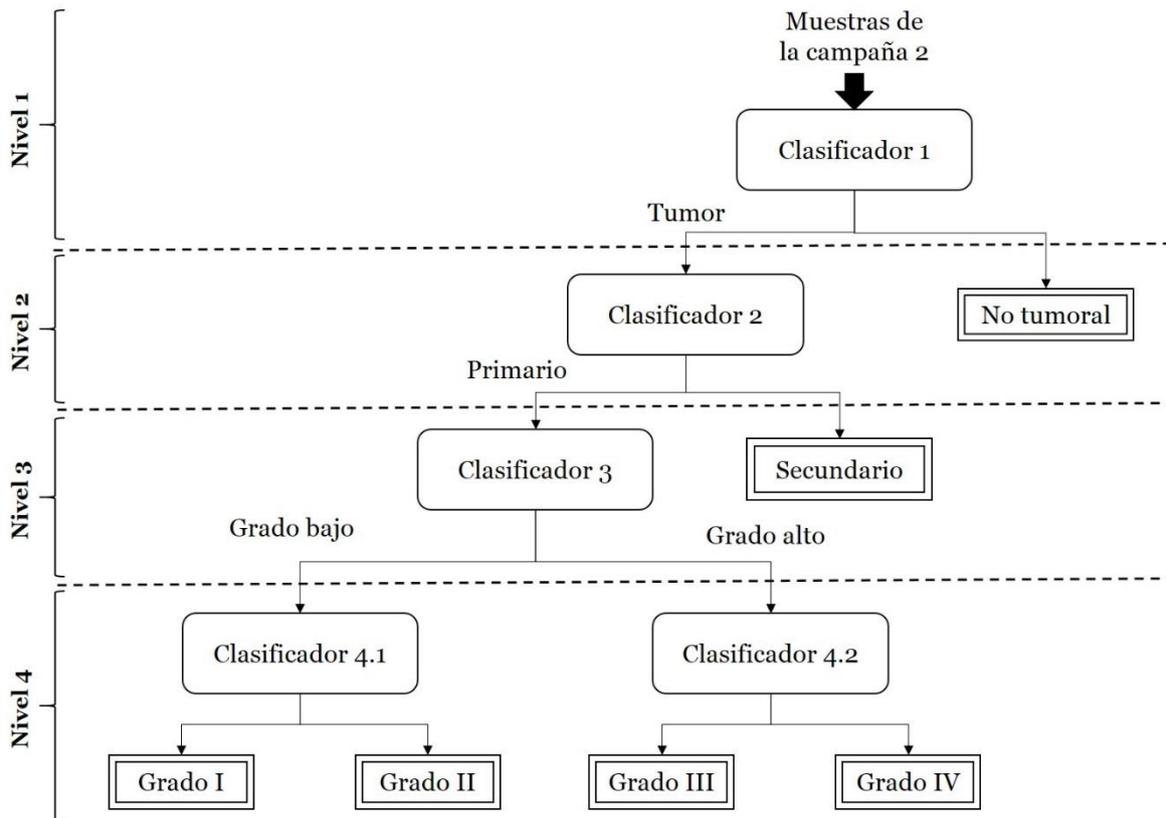


Figura 5.4: Árbol de clasificación de la campaña 2.

Análogamente al árbol de clasificación de la campaña 1, en este árbol, el clasificador llamado *Clasificador 1* se encuentra en el nivel más relevante del árbol por discriminar entre muestras tumorales y no tumorales, y de la misma manera, porque si se comete un error de clasificación al predecir una muestra no tumoral como tumoral, se propagará el error a los siguientes niveles que llegue. Las muestras que se clasifiquen en este nivel como tumorales, pasarán al *Clasificador 2*.

El *Clasificador 2* se encuentra en el nivel 2 de este árbol. Se emplea para discriminar entre tumores primarios y tumores secundarios. Aunque este clasificador no es tan importante como el *Clasificador 1*, la distinción que hace es crucial. Esto es porque aporta información sobre si el tumor se ha diseminado desde otra parte del cuerpo, en tal caso la metástasis podría haberse diseminado también a otras zonas del organismo.

Las muestras clasificadas como tumores primarios pasan a clasificarse por el *Clasificador 3*, en el nivel 3. Este clasificador desempeña exactamente la misma función que el *Clasificador 2.1* del árbol de clasificación de la campaña 1, y presenta la misma relevancia.

Sin embargo, las muestras que sean clasificadas como muestras de grado bajo deberán pasar al *Clasificador 4.1* de nivel 4.

En el *Clasificador 4.1* se introducen las muestras clasificadas como muestras de grado bajo, las cuales se clasificarán para obtener como resultados finales la clase de grado I y la clase de grado II.

Finalmente, en el nivel 4 también tenemos el *Clasificador 4.2*. En este caso se introducen las muestras clasificadas como muestras de grado alto, a las que este clasificador discriminará para obtener los resultados finales de clase de grado III y clase de grado IV. Este clasificador desempeña la misma función y presenta la misma relevancia que el *Clasificador 3* del árbol de clasificación de la campaña 1. En el árbol de clasificación de la campaña 2, los clasificadores del nivel 4 son los más expuestos a errores de propagación.

Para la implementación de este sistema también se ha realizado el proceso de reetiquetado, análogo al realizado en el árbol de clasificación de la campaña 1.

5.6 Sistemas de clasificación multiclase

Como se ha comentado someramente en la introducción de este capítulo, los sistemas de clasificación desarrollados en este proyecto se pueden enmarcar en dos métodos bien diferenciados: árboles de clasificación y clasificadores multiclase. Los dos métodos fueron empleados para clasificar cada una de las campañas por separado, sin embargo, para clasificar la fusión de las campañas, se optó por aplicar los sistemas de clasificación multiclase.

En los sistemas de clasificación multiclase implementados en este estudio, se especifican las clases indivisibles o finales según sea la campaña de datos a clasificar, de manera que, cada muestra a la entrada del sistema se clasifica como una de las clases especificadas en la campaña que se está tratando, sin pasar por diferentes niveles de clasificación. Debe tenerse en cuenta que, a diferencia de los sistemas de clasificación basados en árboles de decisión, los sistemas de clasificación multiclase diseñados no requieren de algoritmos de reetiquetado, esto es porque para cada muestra se realiza una única clasificación.

Por un lado, para la clasificación de la campaña 1 se diseñó un sistema capaz de clasificar las muestras pertenecientes a las cinco clases finales que se disponen en esta campaña: “grado II”, “grado III”, “grado IV”, “sustancia blanca” y “sustancia gris”. De manera similar a los sistemas basados en árboles, se establecieron números identificativos que se corresponden a cada una de las clases, para identificar a las diferentes muestras. En el caso de la campaña 1 las muestras se etiquetan desde “1” hasta “5”.

Por otro lado, el sistema multiclase diseñado para clasificar la campaña 2 tiene la capacidad de clasificar las seis clases finales de esta campaña. Estas son: “grado I”, “grado II”, “grado III”, “grado IV”, “tumor secundario” y por último la clase “no tumoral”. Como puede intuirse, en esta campaña las muestras se etiquetan desde “1” hasta “6”.

Como se ha visto en el apartado 5.2.1, la fusión de las campañas presenta las mismas clases finales que la campaña 2, por tanto los modelos SVM generados para clasificar estos conjuntos consisten del mismo número de *clusters*: seis. Ambos conjuntos presentan además el mismo espacio dimensional (1501 bandas). Por esta razón, se ha aprovechado en gran parte los algoritmos desarrollados en el diseño de los sistemas de clasificación multiclase para la campaña 2, para clasificar también la fusión de las campañas.

Por último, se debe destacar que se desarrollaron sistemas multiclase para clasificar la campaña 2 y la fusión de las campañas, tanto en absorbancia como en transmitancia, con el objetivo de cotejar los resultados en ambas magnitudes.

Los modelos SVM generados en estos sistemas se han implementado empleando distintos tipos de *kernel*. Algunos sistemas fueron diseñados utilizando el *kernel* de tipo lineal con sus parámetros por defecto, mientras que en otros se trabajó con el *kernel* RBF realizando ajustes en sus parámetros. Adicionalmente, se hicieron algunas pruebas con el tipo de *kernel* RBF utilizando sus parámetros por defecto, no obstante, como se verá en el Capítulo 6, estos experimentos solo proporcionaron buenos resultados en la discriminación de las muestras sanas de la campaña 1.

5.7 Optimización de parámetros en los sistemas multiclase

A pesar de que los sistemas de clasificación multiclase no presentan el problema de propagación de errores, inherente a los sistemas basados en árboles de clasificación jerárquica, los sistemas multiclase generan los hiperplanos de separación tratándo todos los *clusters* de manera independiente. Es decir, el sistema multiclase solamente cuenta con la

información de las clases finales, por esta razón no puede aprovecharse la información de agrupamientos mayores para establecer hiperplanos de separación más apropiados en la clasificación, en la generación del modelo.

Por ejemplo, en los sistemas multiclase no se toma en consideración que las muestras de tumores de grado I, II, III, IV y las muestras de tumores secundarios, pueden formar un nuevo *cluster* tumoral que cuenta con mayor número de muestras que cada uno de los *clusters* independientes de las clases finales.

Para afrontar esta “vulnerabilidad” intrínseca a este tipo de sistemas, se pueden aprovechar las funciones *kernels* y sus parámetros con el fin de generar hiperplanos de separación que proporcionen modelos SVM más robustos.

Cuando se abordó la optimización de parámetros en la elaboración de este proyecto, se tomó como punto de partida el siguiente artículo: “*Classification of ex-vivo breast cancer positive margins measured by hyperspectral imaging*” [110]. En el estudio realizado en dicho artículo se emplea la función *kernel* RBF (*Radial Basis Function*) para establecer el hiperplano de separación. En este trabajo se decide implementar los sistemas multiclase con este *kernel* en la búsqueda de los parámetros óptimos empleando el método “*grid search*”.

Como puede observarse en la ecuación (3.35), el *kernel* RBF cuenta con un parámetro *gamma* (γ) que puede ajustarse a la hora de establecer los hiperplanos, no obstante, hay que tener en cuenta que también se dispone de un parámetro de *coste* (C) para todos los *kernels*. El coste es un parámetro que penaliza los errores de las muestras mal clasificadas y que a su vez, determina la importancia de la maximización de la distancia de separación frente a la minimización de los errores, de modo que, a mayor valor del coste, mayor importancia se da a los errores de clasificación, y a menor valor del coste, mayor importancia se le da a la maximización de la distancia [4].

Por otra parte, se presentó la duda sobre la clase a considerar para cotejar los resultados. Esto se planteaba debido a que los parámetros óptimos, no tiene por qué ofrecer los mejores resultados a la hora de discriminar todas las clases. Es decir, puede suceder que un par de parámetros (γ, C) ofrezca los mejores resultados para clasificar una clase, y ofrecer resultados no tan buenos para clasificar otras clases.

En un principio se pensó en realizar el análisis tomando como clase bajo test, la clase “no tumoral” o “sana” (creando la clase “sanas” para la campaña 1) ya que, al obtener las métricas de sensibilidad y especificidad de estas clases en sus respectivas campañas, solo basta con intercambiar sus valores para conocer las métricas de la clase “tumor”. Es decir, la sensibilidad de la clase “no tumoral” es la especificidad de la clase “tumoral” y viceversa. De esta manera, al finalizar este estudio tendríamos los parámetros óptimos, (γ, C) , para la clase más relevante en cada campaña. Sin embargo, se tuvo que replantear la estrategia ya que las clases “no tumoral” y “sanas”, disponen de pocas muestras.

Por ello, se decidió realizar el análisis de parámetros óptimos tomando bajo test la clase que más muestras presenta, “Grado IV”, y una vez obtenidos los parámetros óptimos, comprobar los resultados de clasificación del resto de las clases con estos parámetros.

5.7.1. Optimización de parámetros en la clasificación de la campaña 1

Al empezar la búsqueda de los parámetros óptimos para clasificar la campaña 1, se presentaba la incertidumbre sobre con qué valores comenzar. La razón de esto es que al tratarse de datos espectrográficos con tantas bandas, se hace imposible conocer *a priori* los valores de los parámetros que se acercan a los mejores resultados.

Ante esta duda se decidió empezar con los mismos valores empleados en el artículo [110]. Este conjunto de valores se presenta a continuación:

- $\gamma = \{2^{-5}, 2^{-3}, 2^{-1}, 2^1, 2^3, 2^5, 2^7, 2^9, 2^{11}, 2^{13}\};$
- $C = \{2^{-5}, 2^{-3}, 2^{-1}, 2^1, 2^3, 2^5, 2^7, 2^9, 2^{11}, 2^{13}, 2^{15}\}.$

Una vez se realizaron las clasificaciones empleando todas las combinaciones lineales de estos valores en el *kernel* RBF, se representaron en dos gráficas diferentes los resultados de las clasificaciones para estas combinaciones de valores de parámetros. Por un lado, se representaron los promedios de sensibilidad, y por otro lado se representaron los promedios de especificidad. Al tratarse de dos rangos de valores en el plano (γ, C) , los promedios²⁴ se representan a través de superficies.

Como se refleja en la Figura 6.16, con estos valores, tomados como punto de partida, no se obtuvieron resultados satisfactorios. Por tanto, se continuó la búsqueda de parámetros

²⁴ Se trata de promedios porque en cada proceso de clasificación se realizan 50 repeticiones, como se comentó en el apartado 5.3.

óptimos extendiendo el rango de valores, al mismo tiempo que se observaba en las gráficas los resultados que se iban obteniendo. De este modo, se realizaba la búsqueda de forma visual en diversas zonas del plano tratando de acercarse a los rangos de valores que mejores resultados presentaban. Para ser exactos, el plano (γ, C) se representó en escalas logarítmicas ya que los *barridos* de valores se realizaron en potencias de 2, $(\log_2 \gamma, \log_2 C)$.

De los barridos de parámetros realizados con esta campaña, se obtuvieron tres pares de promedios de sensibilidad y especificidad que presentaron los mejores resultados (promedios más altos y más balanceados). Para obtener el mejor de los tres pares se aplica la expresión (5.1) como parámetro de diseño, y se escoge el par que proporciona el factor más elevado.

Para establecer numéricamente el mejor resultado ante diferentes pares de promedios de sensibilidad y especificidad en este proyecto, se distinguen los tres criterios siguientes:

1. El promedio de sensibilidad debe ser lo más alta posible para que haya el menor número de falsos negativos (muestras de grado IV que no se clasifican como tales).
2. La media de especificidad debe ser lo más alta posible para que el número de falsos positivos sea el menor posible (muestras de otras clases que se clasifican como muestras de grado IV).
3. Estos promedios deben ser lo más balanceados posibles para que exista un equilibrio en el clasificador entre la tasa de verdaderos positivos y la tasa de verdaderos negativos.

Atendiendo a estos criterios, se define el siguiente factor de diseño:

$$f = \frac{\text{Sensibilidad (media)} + \text{Especificidad (media)}}{|\text{Sensibilidad (media)} - \text{Especificidad (media)}| + 1} \quad (5.1)$$

El numerador se establece para cumplir con los dos primeros criterios, mientras que el valor absoluto del denominador satisface el tercer criterio. El “1” que se ha añadido al denominador tiene la función de evitar que el factor sea infinito para los casos en que los promedios sean iguales.

Cabe resaltar que el estudio de obtención de parámetros óptimos se puede realizar calculando este factor de diseño en todas las clasificaciones y automatizando la búsqueda

del factor más elevado en las diferentes regiones del plano (γ, C) que se estudian. De este modo, se tiene una única superficie de estudio en vez de dos.

Sin embargo, se optó por realizar el estudio visualizando los promedios de sensibilidad y especificidad como métricas individuales, puesto que de este modo se puede obtener información a nivel visual de su comportamiento por separado. Es decir, la sensibilidad y la especificidad son medidas estadísticas independientes.

Posteriormente, se realizó un “ajuste fino” en torno al par de métricas que presentaban los mejores resultados con el objetivo de comprobar si se pueden mejorar estos resultados. Tras realizar este ajuste fino se obtienen los parámetros óptimos.

Por último, se comprobó que los resultados en la clasificación del resto de las clases habían sido altos, empleando estos parámetros óptimos. Este estudio se realizó con las muestras de la campaña 1 expresadas en absorbancia.

Esta parte resultará más comprensible cuando se vea el estudio en el capítulo de resultados experimentales (apartado 6.4.1.1).

5.7.2. Optimización de parámetros en la clasificación de la campaña 2

Aprovechando el estudio presentado en el apartado 5.7.1, en este caso se trató de acelerar la búsqueda del par de parámetros óptimos. Para ello, en vez de tomar como punto de partida los valores del artículo, se tomaron los parámetros óptimos (γ, C) obtenidos en el estudio anterior. Aunque este hecho nos proporcionó mejores resultados de comienzo, también fue necesario realizar una búsqueda en diferentes zonas del plano $(\log_2 \gamma, \log_2 C)$.

Para empezar, se escogieron dos rangos de valores (γ, C) que tuvieran aproximadamente en su centro los parámetros óptimos de la campaña 1. En esta región del plano ya se observaban zonas que ofrecían muy buenos resultados, tanto de sensibilidad como de especificidad, no obstante, se aumentaron los valores de γ en un nuevo barrido tratando de mejorar los resultados. Al conseguirse con este barrido un punto de inflexión con los mejores resultados, en esta zona se realizó un ajuste fino.

Análogamente al estudio anterior, se obtuvieron tres pares de promedios de sensibilidad y especificidad que presentaron buenos resultados, y de estos se tomó el que proporcionaba el factor más alto (ecuación (5.1)).

En este caso, también se comprobó que los resultados fueron satisfactorios empleando estos parámetros óptimos, tomando al resto de las clases bajo test. A diferencia del estudio de parámetros óptimos de la campaña 1, este estudio se realizó con las muestras de la campaña 2 expresadas en transmitancia.

No obstante, se transformaron estas muestras a absorbancia y luego se clasificaron utilizando los parámetros óptimos obtenidos en transmitancia. Como se verá en el Capítulo 6, los resultados no fueron tan buenos como en transmitancia.

5.7.3. Optimización de parámetros en la clasificación de la fusión de las campañas

Para empezar este estudio, se realizaron clasificaciones de la fusión de las campañas empleando los parámetros óptimos obtenidos con la campaña 1 y los obtenidos en la campaña 2.

Luego, se realiza una búsqueda de parámetros óptimos empleando las muestras de la fusión de las campañas. En este caso, se toma el rango de valores de parámetros tomando como límites los parámetros óptimos obtenidos con la campaña 1, y los parámetros óptimos obtenidos con la campaña 2.

Posteriormente se realizan dos “ajustes finos” de las zonas del plano que mejores resultados se presentan y de estos últimos barridos de parámetros, se obtiene el par de promedios de sensibilidad y especificidad que proporciona el factor más elevado (ecuación (5.1)) y finalmente se escogen los parámetros (γ, C) con los que se obtiene ese par, tomando para el coste aquel que tiene el valor más bajo. Se finaliza este estudio observando que los resultados obtenidos en la clasificación considerando a cada una de las clases como clase bajo test, son satisfactorios. Este estudio se ha realizado con las muestras de la fusión de campañas expresadas en absorbancia.

5.8 Resumen

En este capítulo se han explicado los dos métodos de clasificación empleados en este proyecto: el método basado en árboles de clasificación jerárquica y el método de clasificación multiclase. Por un lado, se ha utilizado el método de clasificación jerárquica para desarrollar un sistema capaz clasificar las muestras de la campaña 1 y otro para clasificar las muestras de la campaña 2. Por otro lado, se ha desarrollado un sistema de

clasificación multiclase para clasificar las muestras de la campaña 1, otro para las muestras de la campaña 2 y un sistema para clasificar las muestras de la fusión de las campañas.

Hay que recalcar que los sistemas basados en árboles ofrecen la ventaja de disponer de mayor número de muestras en los clasificadores de los primeros niveles. Esto se debe a que los conjuntos de muestras se van dividiendo a medida que se avanza en el árbol. Disponer de mayor cantidad de muestras en el primer nivel, implica obtener un modelo de clasificador SVM más robusto para realizar la clasificación más relevante de este trabajo: la discriminación de las muestras tumorales de las que no lo son. Sin embargo, en estos sistemas de clasificación surge el inconveniente de los errores de propagación, que se producen cuando se “canalizan” incorrectamente las muestras previamente mal clasificadas.

Por otra parte, aunque los sistemas multiclase no contemplan errores de propagación, clasifican las muestras considerando únicamente las clases finales y discriminando entre estas de manera independiente. En general, esta desventaja implica disponer de menor número de muestras para cada una de sus clases, por tanto, se desaprovecha la posibilidad de agrupar clases para realizar clasificaciones más robustas en las clases más relevantes.

Este inconveniente, inherente a los sistemas multiclase, se ha tratado de atenuar desarrollando sistemas que emplean el *kernel* RBF en las clasificaciones. Además, se ha realizado un estudio para obtener los parámetros óptimos del *kernel* RBF en la clasificación de cada una de las campañas por separado, y para la fusión de ambas campañas.

Los sistemas desarrollados para la clasificación de la campaña 1 permiten clasificar muestras de grado II, grado III, grado IV y entre los tejidos sanos: diferenciar entre tejidos de sustancia blanca y sustancia gris. Mientras que los sistemas desarrollados para clasificar la campaña 2 y la fusión de las campañas, permiten discriminar muestras de los cuatro grados de tumores primarios, muestras de tumor secundarios, así como las muestras no tumorales.

Se ha decidido establecer los conjuntos de entrenamiento (*trainings*) y validación (*test*), al 70% y 30% respectivamente, de las muestras de cada clase en los diferentes sistemas desarrollados en este proyecto, por ser los valores recomendados en este tipo de aplicaciones. Además, las muestras se han agrupado en estos conjuntos previamente a la clasificación, empleando un método de asignación aleatoria de muestras que se realiza 50

veces (correspondientes a las 50 clasificaciones del proceso de clasificación) con y sin reemplazo (a diferencia del método *cross-validation*). Sin embargo, como se refleja en el siguiente capítulo, se han hecho algunas excepciones con la finalidad de cotejar los resultados.

Capítulo 6

Resultados experimentales

6.1 Introducción

En este capítulo se detallan los resultados obtenidos a partir de los diferentes experimentos realizados en el presente Proyecto Fin de Carrera. La metodología seguida en estos experimentos se ha descrito pormenorizadamente en el Capítulo 5, pero se ha optado por reflejar los resultados cuantitativos y sus respectivas explicaciones en un capítulo aparte, con la finalidad de disponer de una memoria más ordenada, de facilitar su comprensión y de hacer más cómoda su consulta.

Conjuntamente a los resultados cuantitativos, se satisfacen algunas cuestiones relacionadas con estos resultados. Es decir, en este capítulo se aportan diversas explicaciones y aclaraciones que no fueron realizadas en el capítulo anterior, por considerar más apropiado presentarlas junto a estos resultados, para lograr un mejor entendimiento de lo que se está tratando.

Además, se presentan algunas comparativas entre resultados de los diferentes sistemas diseñados en este proyecto, así como con algunos experimentos realizados en el PFC precedente [4].

Hay que recordar que, en el proceso de clasificación de todos los sistemas diseñados en este PFC, se tomó como parámetro de diseño realizar cincuenta repeticiones para generar los modelos SVM, por lo tanto, los resultados de validación vendrán expresados en sus valores promedios. Por otro lado, se decidió entrenar los modelos SVM de los sistemas de clasificación con el 70% de las muestras, dejando el 30% restante para validación. Estos porcentajes también se toman como parámetros de diseño en este proyecto, salvo en las dos excepciones en las que se indicarán porcentajes diferentes.

6.2 Firmas espectrales de las muestras preprocesadas

En las siguientes figuras se presentan las firmas espectrales de todas las muestras empleadas en los diferentes sistemas desarrollados en el proyecto. Se han agrupado en las clases finales para la clasificación de la fusión: “Grado I” (Figura 6.1), “Grado II” (Figura 6.2), “Grado III” (Figura 6.3), “Grado IV” (Figura 6.4), “Tumor secundario” (Figura 6.5) y “No tumor” (Figura 6.6). En las figuras están separadas por campaña (las muestras de sustancia blanca y gris se han agrupado en la clase “No tumor”).

Se trata de la representación de sus respuestas, tanto en absorbancia como en transmitancia, normalizadas y sus medias. El eje espectral de las muestras de la campaña 1 está adaptado a las de la campaña 2 (de 1000 nm a 4000 nm), y las muestras que presentaban altos niveles de ruido (Figura 4.13), han sido filtradas a través de un filtrado tipo *smooth*. No obstante, se considera que el nivel de ruido presente en otras muestras es tolerable en los sistemas desarrollados.

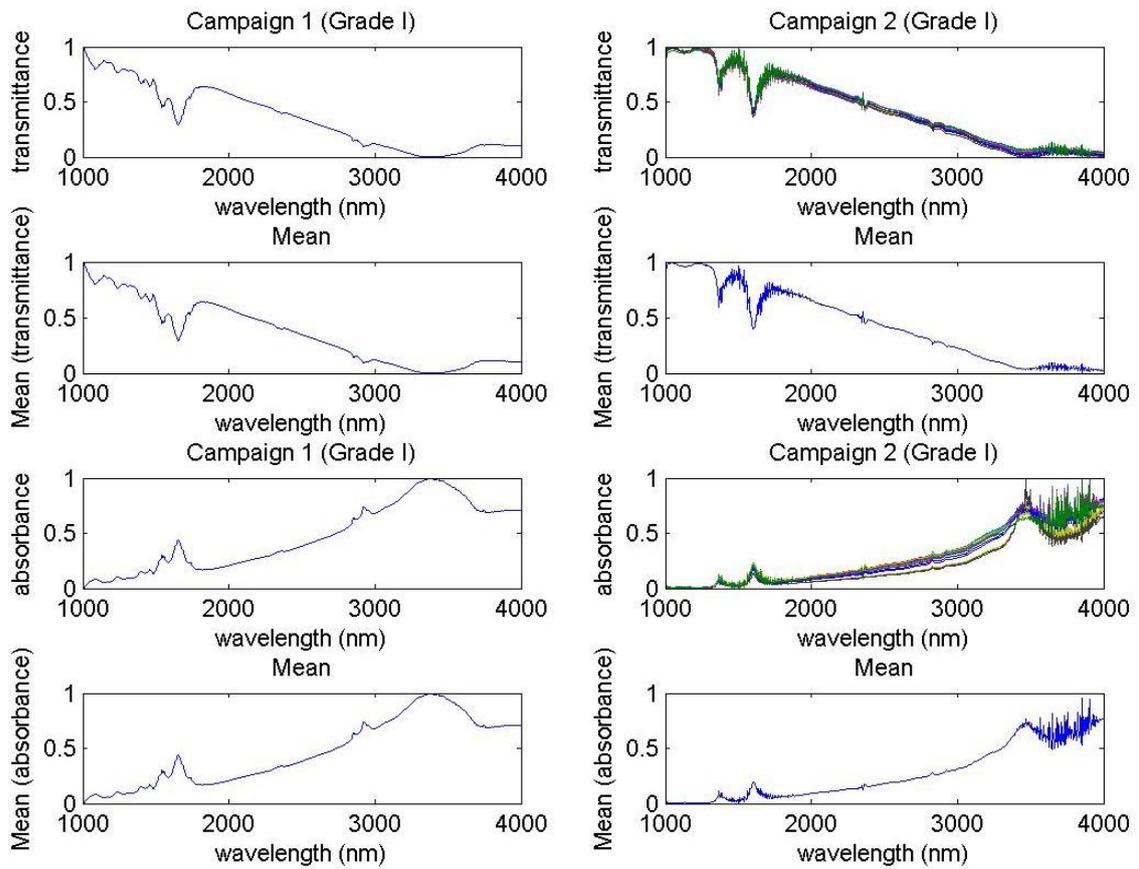


Figura 6.1: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de grado I (campana 1 y campana 2) y sus respectivas medias.

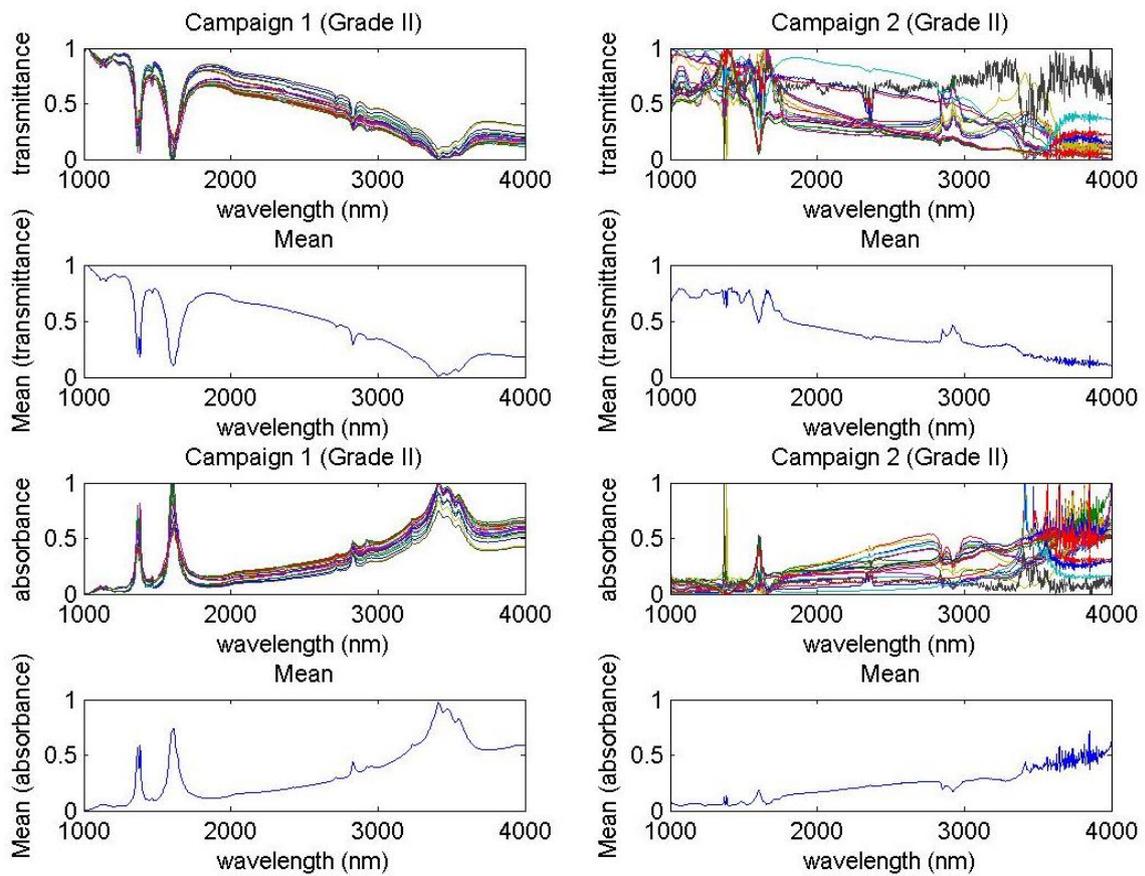


Figura 6.2: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de grado II (campana 1 y campana 2) y sus respectivas medias.

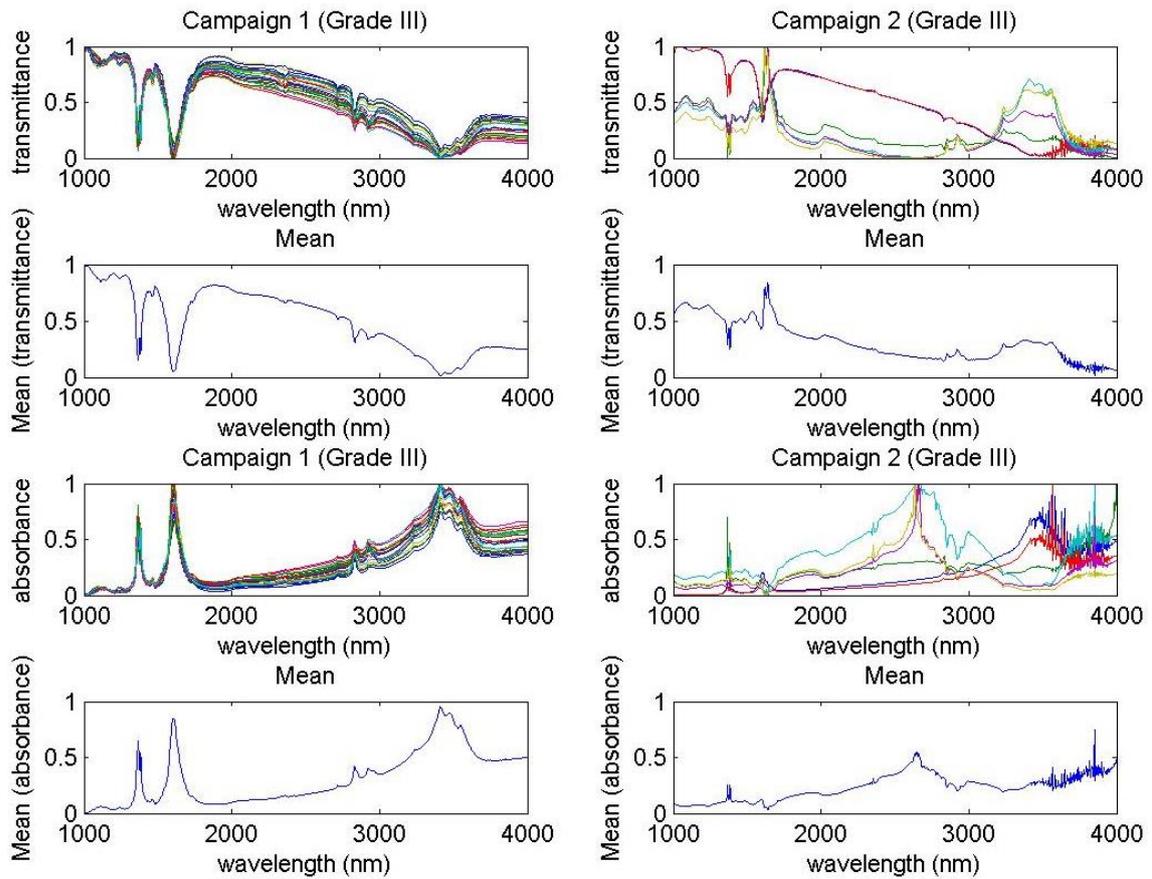


Figura 6.3: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de grado III (campana 1 y campana 2) y sus respectivas medias.

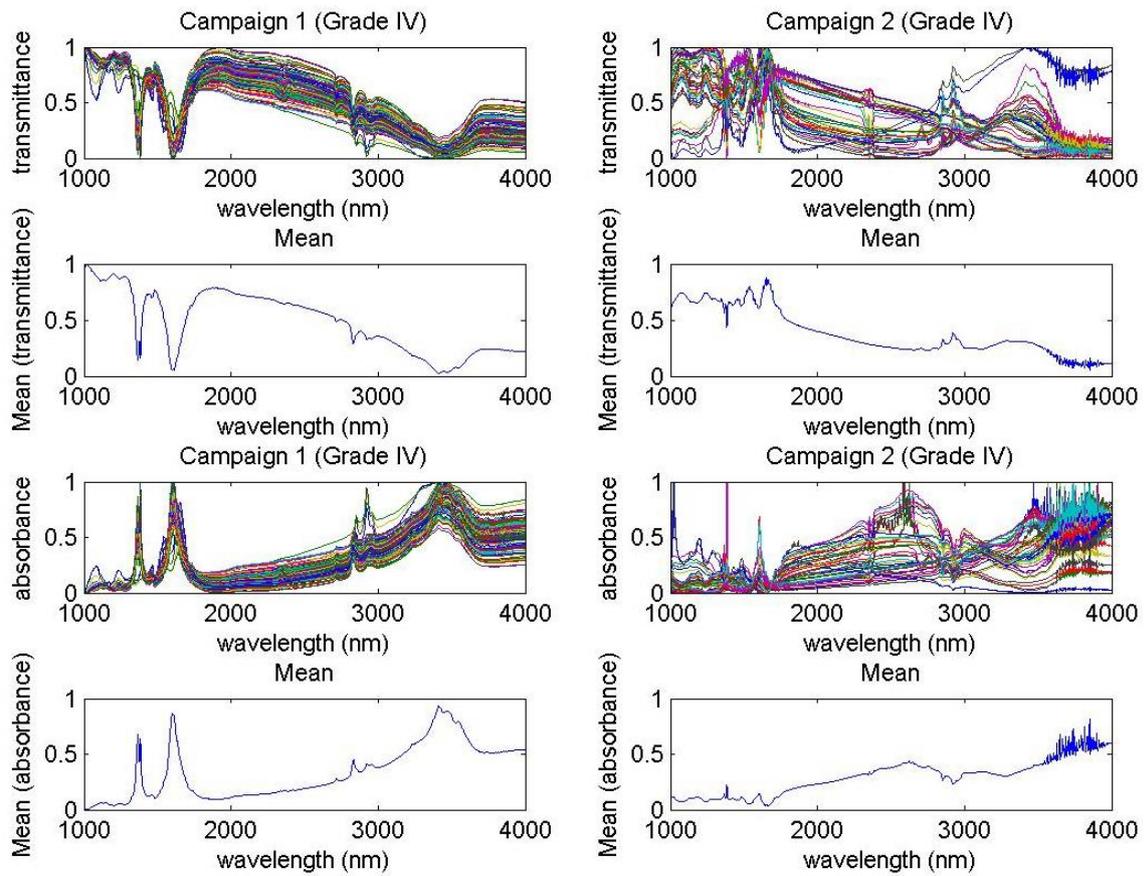


Figura 6.4: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de grado IV (campana 1 y campana 2) y sus respectivas medias.

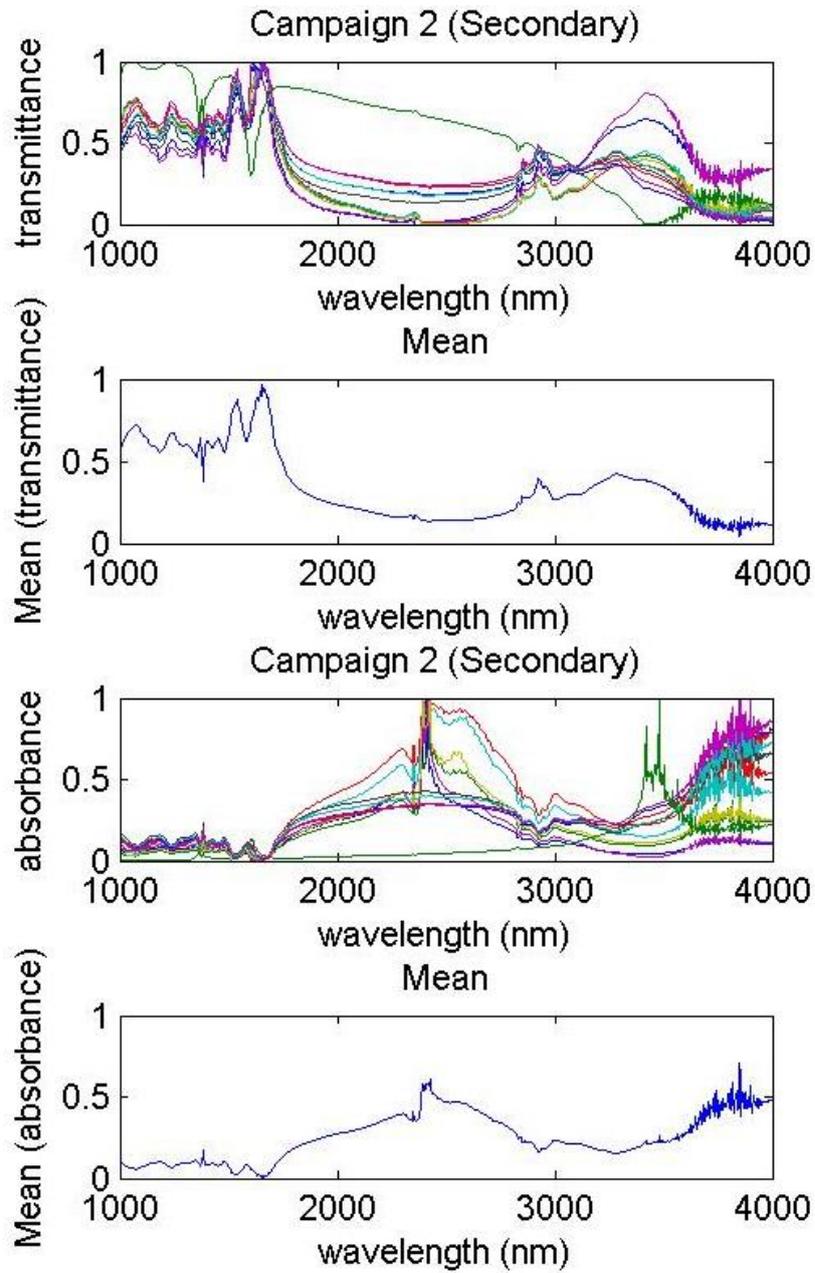


Figura 6.5: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras de metástasis (campaña 2) y sus respectivas medias.

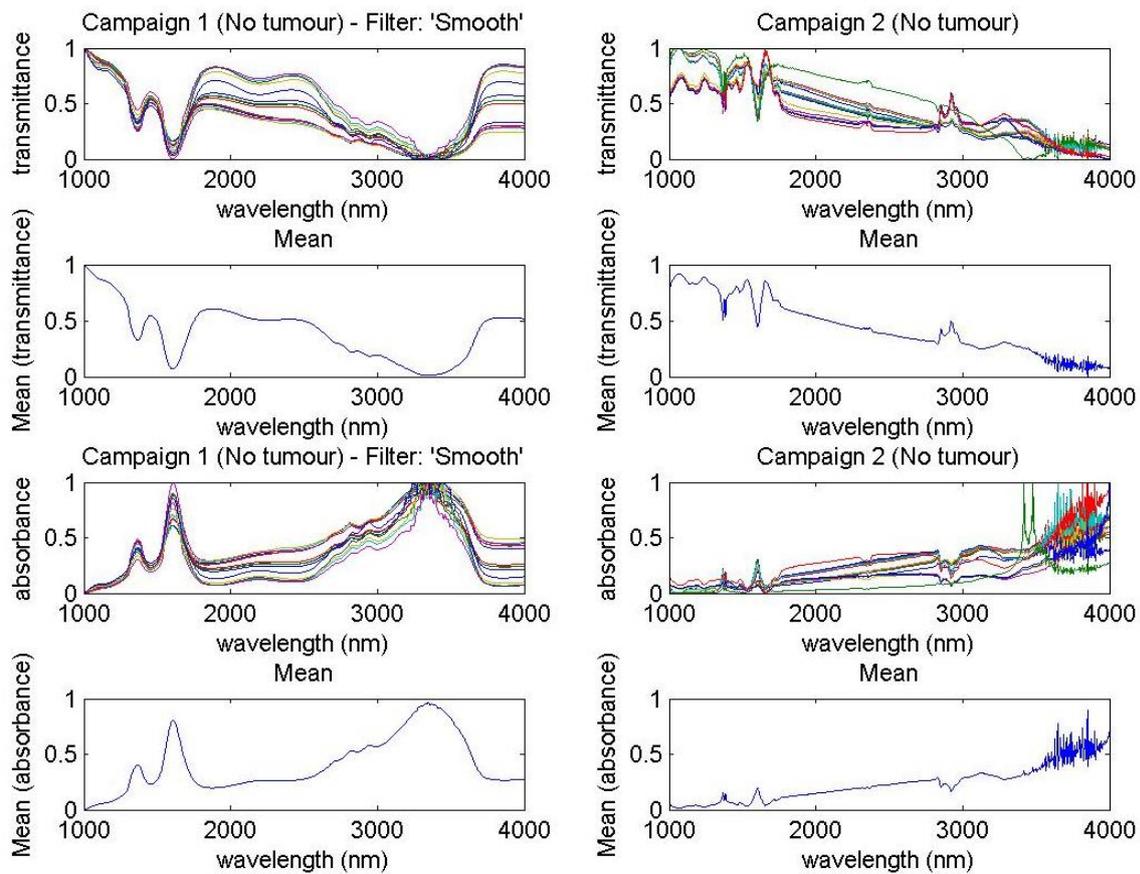


Figura 6.6: Respuesta en transmitancia y absorbancia de las firmas espectrales de las muestras no tumorales (campaña 1 y campaña 2) y sus respectivas medias.

6.3 Resultados en sistemas de árboles jerárquicos

Como se comentó en el apartado 5.5 de esta memoria, en este proyecto se diseñaron árboles de clasificación para clasificar la campaña 1 y la campaña 2. Se empleó únicamente el *kernel* lineal para todos los clasificadores de las dos estructuras diseñadas, y se tomaron los parámetros por defecto de dicho *kernel*.

Los experimentos realizados con los sistemas basados en árboles jerárquicos, se hicieron con las muestras de la campaña 1 en su respuesta en absorbancia, y con las muestras de la campaña 2, en su respuesta en transmitancia.

6.3.1. Resultados en sistemas de árboles jerárquicos (campaña 1)

6.3.1.1. Diseño preliminar y comparativa

El diseño preliminar de este árbol jerárquico para clasificar los datos de la campaña 1, consistió en desarrollar un sistema similar a uno de los diseños más básicos empleados en el PFC precedente [4].

Una condición en el diseño de este sistema fue que empleara los mismos conjuntos de entrenamiento y test que utilizaba el sistema desarrollado en el PFC anterior, con el objetivo de comparar sus resultados fehacientemente. Por tanto, este es uno de los casos excepcionales en los que no se siguen los parámetros de diseño establecidos en este proyecto, para entrenamiento y validación: 70% de muestras de entrenamiento y 30% de muestras de test. Por el contrario, en este caso se establecen los conjuntos de muestras de entrenamiento que se pueden ver en la Tabla 6.1.

Tabla 6.1: Número de muestras de entrenamiento en el diseño preliminar del árbol de clasificación (Campaña 1).

Número de muestras de entrenamiento								
Nivel	1		2		2		3	
Clasificador	<i>Clasificador 1</i>		<i>Clasificador 2.1</i>		<i>Clasificador 2.2</i>		<i>Clasificador 3</i>	
Clase	Tumor	Sanas	Grado bajo	Grado alto	Sustancia gris	Sustancia blanca	Grado III	Grado IV
Muestras	6	4	4	6	3	5	4	6

Cuando se estudiaron los conjuntos de entrenamiento empleados en el PFC anterior, se observó una particularidad en la estrategia de su diseño. Los conjuntos de entrenamiento y test que se emplean en los diferentes clasificadores, no se toman como conjuntos independientes en todo el árbol.

Debe tenerse en cuenta que esta decisión en el diseño implica que los conjuntos de *trainings* y *tests*, establecidos para un clasificador concreto de los niveles superiores, son independientes de los resultados obtenidos en las clasificaciones de los niveles previos al nivel de ese clasificador. Esto supone que las muestras de entrenamiento de un clasificador pueden haber sido usadas como muestras de test en otro clasificador y viceversa.

Por tanto, pueden ocurrir situaciones como la siguiente: muestras que se usan como *trainings* en un clasificador (por ejemplo el *Clasificador 2.1* del nivel 2), pueden haber sido clasificadas erróneamente por un clasificador de un nivel anterior hacia otra rama (por ejemplo por el *Clasificador 1* del nivel 1, “canalizando” las muestras hacia el *Clasificador*

2.2 del nivel 2). Este tipo de errores de propagación pueden evitarse si se establecen conjuntos de *trainings* y *tests* independientes para todo el árbol, de manera que los *trainings* de un clasificador no puedan ser mal canalizados por otros clasificadores.

A continuación, se presenta una comparativa de los resultados obtenidos por los sistemas diseñados en ambos PFCs, estableciéndose para ambos árboles de clasificación, los mismos *trainings* y *tests*.

Clasificador 1

En la Tabla 6.2 se exponen los valores medios y las desviaciones típicas de las métricas de validación explicadas en el apartado 5.4: *overall accuracy*, sensibilidad y especificidad; para el *Clasificador 1* (Figura 6.7).



Tabla 6.2: Resultados del árbol de clasificación en los PFCs actual y precedente en el *Clasificador 1* (campaña 1).

Clasificador 1			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
PFC precedente	100 (0)	100 (0)	100 (0)
PFC actual	100 (0)	100 (0)	100 (0)

Como puede observarse en la tabla, las clasificaciones en estos diseños básicos implementados en ambos PFCs aportan muy buenos resultados en la campaña 1. Obtener el 100% de acierto en las tres métricas promedio de este clasificador implica que el sistema discrimina correctamente todas las muestras sanas de las tumorales.

Clasificador 2.1

En la Tabla 6.3 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 2.1* (Figura 6.8).

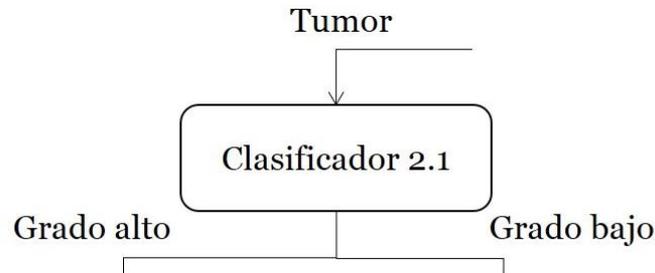


Figura 6.8: *Clasificador 2.1* del árbol jerárquico de la campaña 1.

Tabla 6.3: Resultados del árbol de clasificación en los PFCs actual y precedente en el *Clasificador 2.1* (campaña 1).

Clasificador 2.1			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
PFC precedente	76,91 (7,42)	78,17 (9,18)	66,24 (17,82)
PFC actual	77,39 (7,86)	78,20 (9,62)	69,82 (18,07)

Sin embargo, en este clasificador se cometen algunos errores en la predicción de muestras de grado alto, así como en la predicción de muestras de grado bajo. Los parámetros promedios de sensibilidad y especificidad se han calculado en este clasificador tomando como clase bajo test las muestras de grado alto, por tanto, la tasa de muestras de grado alto correctamente clasificadas es un poco mejor que la tasa de muestras de grado bajo en los dos diseños.

Estos resultados no balanceados pueden producirse por varias razones: o bien las muestras *trainings* escogidas aleatoriamente de grado alto eran más representativas, lo que originó un modelo SVM con un hiperplano más favorable para esta clase; o la distribución del *cluster* de esta clase presenta mejores condiciones de separabilidad que el *cluster* de las muestras de grado bajo; o porque el sistema se ha entrenado mejor para esta clase debido a que se cuenta con más muestras de entrenamiento. También pueden haberse dado varias de estas razones simultáneamente.

Hay que resaltar que en este clasificador no se ha originado ningún error por propagación de errores, puesto que el clasificador del nivel anterior clasifica correctamente todas las muestras.

Adicionalmente, puede apreciarse que el *Clasificador 2.1* implementado en este PFC presenta una ligera mejora respecto al desarrollado anteriormente.

Clasificador 2.2

En la Tabla 6.4 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 2.2* (Figura 6.9).

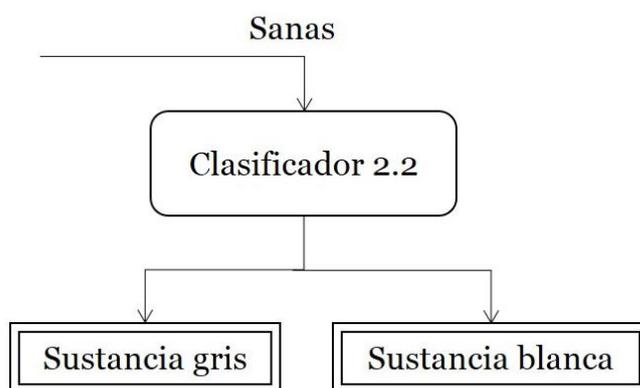


Figura 6.9: *Clasificador 2.2* del árbol jerárquico de la campaña 1.

Tabla 6.4: Resultados del árbol de clasificación en los PFCs actual y precedente en el *Clasificador 2.2* (campaña 1).

Clasificador 2.2			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
PFC precedente	98,29 (6,85)	100 (0)	97 (11,99)
PFC actual	70,22 (4,91)	100 (0)	58,4 (8,58)

En este clasificador se discrimina entre muestras de sustancia blanca y sustancia gris. En este caso se han obtenido la sensibilidad y la especificidad para la clase “sustancia gris” bajo test, por lo que, a tenor de los resultados, todas las muestras de sustancia gris están siendo correctamente clasificadas (la sensibilidad del 100% indica que no hay falsos negativos), mientras que se están produciendo errores al clasificar muestras de sustancia blanca como muestras de sustancia gris.

Hay que tener en cuenta que se han empleado 5 muestras de sustancia blanca frente a 3 de sustancia gris para entrenar los sistemas, de un total de 9 muestras de sustancia blanca y 6 de sustancia gris. En este clasificador, el sistema desarrollado en el PFC precedente ostenta unos porcentajes de acierto relativamente altos en relación a las muestras sanas que se disponen en esta campaña. Hecho que contrasta con los resultados obtenidos en el diseño de este proyecto, sobre todo en los errores al clasificar las muestras de sustancia blanca.

Por la misma razón dada para el *Clasificador 2.1*, tampoco se propagan errores a este clasificador desde el nivel anterior.

Clasificador 3

En la Tabla 6.5 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 3* (Figura 6.10).

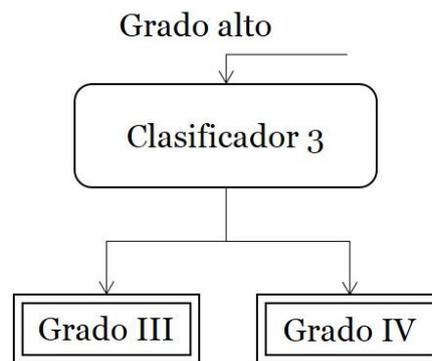


Figura 6.10: *Clasificador 3* del árbol jerárquico de la campaña 1.

Tabla 6.5: Resultados del árbol de clasificación en los PFCs actual y precedente en el *Clasificador 3* (campaña 1).

Clasificador 3			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
PFC precedente	65,62 (12,09)	66,53 (15,26)	60,73 (19)
PFC actual	63,62 (11,74)	67,73 (16,89)	44,33 (22,55)

En los resultados de este clasificador se pueden observar unas tasas de acierto global, OA, no especialmente favorables. No obstante, hay que resaltar que parte de los errores producidos en este clasificador son errores de propagación que proceden del *Clasificador 2.1*. El promedio de especificidad cercano al 70% en el *Clasificador 2.1* del diseño realizado

en este proyecto indica que se está produciendo un número considerable de falsos positivos al clasificar algunas muestras de grado bajo como si fueran de grado alto.

Además, las desviaciones típicas de las métricas de sensibilidad y especificidad reflejan respectivamente que las proporciones de falsos negativos y falsos positivos fluctúan bastante en las 50 repeticiones.

Por otro lado, al comparar los resultados de los dos sistemas, puede observarse que el *Clasificador 3* del diseño realizado en este PFC presenta una tasa de acierto global ligeramente por debajo a la del PFC precedente. También que las métricas promedio de sensibilidad y especificidad reflejan resultados menos balanceados en este diseño.

Al margen de los errores que se propagan hasta este clasificador y de las muestras de grado alto que no llegan a este clasificador, hay que tener en cuenta que en esta campaña se dispone de un total de 120 muestras de grado IV y 25 de grado III, de las que solamente se han tomado en total 10 muestras de *training*. Se trata de una cantidad amplia de muestras si se compara con las 15 que se dispone en la clase “sanas”. Esto implica que pequeñas variaciones en los porcentajes de sensibilidad y especificidad reflejan mayor número de muestras mal clasificadas. No ocurre lo mismo en un clasificador donde se discrimina entre pocas muestras, es decir, hay que tener presente que las métricas de sensibilidad y especificidad revelan información sobre la clasificación a nivel de proporciones. No obstante, también hay que tener en consideración que este clasificador no presenta tanta relevancia como el *Clasificador 1* por discriminar muestras sanas de tumorales del total de muestras de la campaña 1.

Por otra parte, aunque en esta comparativa el *Clasificador 2.1* de este PFC presenta mejores resultados, el *Clasificador 3* obtiene peores, y resalta el hecho de que el *Clasificador 2.2* se comporta peor, aún sin propagarse errores hacia este en los dos diseños.

Además, llama la atención que el *Clasificador 3* de este PFC clasifique peor, considerando que el promedio de especificidad del *Clasificador 2.1* es más alto en este diseño. Es decir, en los dos sistemas el *Clasificador 1* se comporta perfectamente, con lo que el *Clasificador 2.1* de este diseño propaga menos errores al *Clasificador 3* que el sistema del PFC precedente.

Tras estudiar los algoritmos de ambos diseños, se observó que el sistema preliminar desarrollado en este PFC presentaba una ventaja respecto al diseño del PFC precedente. Aunque los resultados de la comparativa son peores en este PFC (excepto en el *Clasificador 1 y 2.1*), en este diseño se contabilizan los errores de propagación. No obstante, se requiere realizar mejoras en este diseño puesto que los resultados del *Clasificador 2.2* son peores sin producirse propagación de errores a este clasificador.

6.3.1.2. Mejoras en el diseño

A colación de los resultados ofrecidos por el sistema preliminar, se trató de mejorar el diseño implementando un sistema que mantuviera los conjuntos de *training* y *test* independientes en todos los clasificadores del árbol. Además, para este nuevo sistema se establecieron los parámetros de diseño de 70% de *training* y 30% de *test*.

Esto se realiza del siguiente modo: se dispone del 70% de las muestras de cada una de las clases finales para entrenar los clasificadores de las ramas finales del árbol, y estas mismas muestras son reutilizadas, según a las clases no finales a las que pertenecen, para entrenar el resto de clasificadores del árbol.

Empleando esta estrategia en el diseño se reduce la cantidad de posibles errores de propagación. Esto es así porque se evita que las muestras empleadas como *trainings* en un clasificador, puedan ser empleadas como *tests* en otro.

Para aclarar esta cuestión, considérese el siguiente ejemplo: una muestra de grado IV empleada para entrenar el *Clasificador 3* nunca podrá ser clasificada incorrectamente por el *Clasificador 1* o el *Clasificador 2.1*, porque para estos clasificadores esta misma muestra también se emplea en sus respectivos *trainings*. Podría pensarse que esa muestra de grado IV, hipotéticamente mal clasificada, también aporta información sobre la capacidad del clasificador *1* ó *2.1*. Sin embargo, si ese error se propaga del *Clasificador 1* al *2.2*, la contribución que esta muestra aporta a la sensibilidad o especificidad del *Clasificador 2.2* no refleja la capacidad de discriminación de este clasificador por tratarse de un error de propagación. Entonces, este hipotético error de propagación es evitable empleando esta estrategia de diseño, independientemente del porcentaje de entrenamiento utilizado.

Por otro lado, para apreciar la influencia del porcentaje de *training* en los resultados, se realizó una comparación entre dos sistemas desarrollados en este proyecto que siguen el

mismo algoritmo, salvo el código que asigna el porcentaje de las muestras de validación y entrenamiento como es obvio.

Se trata del segundo caso excepcional en el que no se siguen los parámetros de diseño establecidos en este proyecto: el 70% de *training* y el 30% de *test*. Para esta excepción se ha introducido el 30% de *training* y el 70% de *test*.

A continuación, se presenta una comparación de los resultados obtenidos por este nuevo diseño tras haberse entrenado con el 70% y el 30% de muestras respectivamente.

Clasificador 1

En la Tabla 6.6 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 1* (Figura 6.7).

Tabla 6.6: Resultados del árbol de clasificación en el *Clasificador 1* para 70% y 30% de muestras de entrenamiento (campana 1).

Clasificador 1			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
Training = 70%	100 (0)	100 (0)	100 (0)
Training = 30%	100 (0)	100 (0)	100 (0)

Como puede verse en esta tabla, el *Clasificador 1* no discrimina mal ninguna muestra de la campana 1 para los dos porcentajes de *trainings* establecidos. Esto ya era de esperar considerando los resultados que se presentaron en la Tabla 6.2, puesto que con ninguno de los porcentajes de *trainings* empleados en estos experimentos (Tabla 6.6) se obtienen menor cantidad de muestras de entrenamiento que en el experimento anterior.

Clasificador 2.1

En la Tabla 6.7 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 2.1* (Figura 6.8).

Tabla 6.7: Resultados del árbol de clasificación en el *Clasificador 2.1* para 70% y 30% de muestras de entrenamiento (campana 1).

Clasificador 2.1			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
Training = 70%	95,85 (2,44)	99,26 (1,74)	72,62 (19,09)
Training = 30%	91,99 (2,44)	97,1 (3,02)	54,1 (18,09)

Por otro lado, en el *Clasificador 2.1* destaca la influencia del porcentaje de *training* sobre los resultados. Del mismo modo que en el diseño anterior, los errores de clasificación se producen en mayor proporción al clasificar muestras de grado alto como si fueran de grado bajo. Estos falsos positivos están implícitos en la cifra de la especificidad.

A tenor de los resultados obtenidos en el *Clasificador 1*, en este clasificador no se dan errores de propagación y tampoco ocurren en el *Clasificador 2.2*.

Clasificador 2.2

En la Tabla 6.8 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 2.2* (Figura 6.9).

Tabla 6.8: Resultados del árbol de clasificación en el *Clasificador 2.2* para 70% y 30% de muestras de entrenamiento (campana 1).

Clasificador 2.2			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
Training = 70%	100 (0)	100 (0)	100 (0)
Training = 30%	94,09 (9,96)	99,2 (3,96)	91 (16,21)

En esta tabla se refleja que, para el 70% de *training* se discriminan correctamente todas las muestras de sustancia blanca y gris, mientras que para el 30% se mantiene en un porcentaje de aciertos del orden del 90%. Hay que considerar que se cuenta con pocas muestras de tejido sano (9 de sustancia blanca y 6 de sustancia gris) en esta campaña, por

lo que la diferencia de una única muestra en la clasificación se manifiesta claramente en los porcentajes.

Clasificador 3

En la Tabla 6.3 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 3* (Figura 6.10).

Tabla 6.9: Resultados del árbol de clasificación en el *Clasificador 3* para 70% y 30% de muestras de entrenamiento (campaña 1).

<i>Clasificador 3</i>			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
<i>Training = 70%</i>	85,22 (3,25)	99,16 (2,41)	28,57 (10,29)
<i>Training = 30%</i>	81,18 (3,84)	97,59 (3,61)	18,06 (9,37)

Como se muestra en la tabla, aún con porcentajes de *training* muy diferentes se presentan tasas de acierto global muy parecidas. Hay que resaltar que el promedio de especificidad del *Clasificador 2.1*, está indicando que en dicho clasificador se produce una proporción importante de falsos positivos (muestras de grado bajo clasificadas como muestras de grado alto), como es evidente, estos errores de propagación también se reflejan en los resultados del *Clasificador 3*.

Por otra parte, se debe destacar que tras las clasificaciones realizadas empleando el *kernel* lineal, se realizaron pruebas empleando el *kernel* polinomial y el *kernel* RBF en todos sus clasificadores (tomando los parámetros de cada *kernel* por defecto). En los dos casos se obtuvieron los mismos resultados: con el *Clasificador 1* se clasifican todas las muestras correctamente obteniéndose los mismos resultados que se obtuvieron con el tipo de *kernel* lineal (Tabla 6.6); con el *Clasificador 2.2*, igualmente a los resultados obtenidos con *kernel* de tipo lineal, se clasifican correctamente todas las muestras (Tabla 6.8); en contraste, con el *Clasificador 2.1* se clasifican todas las muestras como muestras de grado alto; y con el *Clasificador 3* se clasificaron todas las muestras como muestras de grado IV.

Al finalizar los experimentos efectuados con esta estructura para clasificar la campaña 1, se procedió a desarrollar un sistema de clasificación basado en árboles capaz de clasificar las muestras de la campaña 2.

6.3.2. Resultados en sistemas de árboles jerárquicos (campaña 2)

Tomando como referencia los sistemas que se desarrollaron para clasificar la campaña 1, procedió a diseñar un sistema para clasificar las muestras de la campaña 2, atendiendo a la estructura ilustrada en la Figura 5.4.

A continuación, se presentan los resultados de validación obtenidos con dicho sistema, para cada uno de los clasificadores de este árbol.

Clasificador 1

En la Tabla 6.10 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 1* (Figura 6.11).

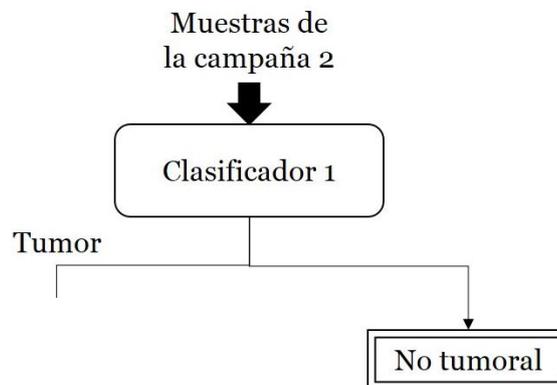


Figura 6.11: *Clasificador 1* del árbol jerárquico de la campaña 2.

Tabla 6.10: Resultados del árbol de clasificación en el *Clasificador 1* para el 70% de muestras de entrenamiento (campaña 2).

<i>Clasificador 1</i>			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
Training = 70%	90,47 (3,72)	99,27 (2,78)	37,93 (20,58)

Se trata del clasificador más relevante en este árbol puesto que discrimina entre muestras no tumorales y tumorales, y aunque el promedio de OA indica un buen porcentaje de acierto global, el promedio tan bajo de especificidad indica que el clasificador comete muchos errores en la clasificación. En este caso la clase bajo test es “tumor”, por tanto, el promedio de especificidad indica que se origina una proporción alta de falsos positivos, en otras palabras, que una proporción alta de muestras sanas se están clasificando como tumorales. Además, la desviación típica de la especificidad es de 20,58%, con lo que la especificidad fluctúa considerablemente en las 50 repeticiones. Sin embargo, a tenor de los

valores altos de sensibilidad (promedio cercano al 100%), son muy pocas las muestras tumorales que se clasifican como no tumorales, por lo que este clasificador genera muy pocos errores de propagación.

Clasificador 2

En la Tabla 6.11 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 2* (Figura 6.12).

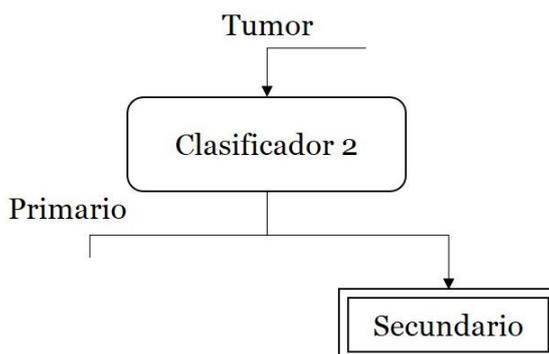


Figura 6.12: *Clasificador 2* del árbol jerárquico de la campaña 2.

Tabla 6.11: Resultados del árbol de clasificación en el *Clasificador 2* para el 70% de muestras de entrenamiento (campaña 2).

Clasificador 2			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
Training = 70%	89,22 (3,64)	99,38 (1,77)	48,24 (12,77)

El *Clasificador 2* discrimina entre muestras de tumor primario y secundario. Aunque presenta resultados muy altos de sensibilidad, la especificidad indica que existe una proporción importante de falsos positivos. En este caso la clase bajo test es “tumor primario”, por tanto, podría pensarse que una proporción alta de metástasis se está clasificando como tumores primarios. Esto implicaría un riesgo considerable *a priori*, puesto que en el ámbito médico podría pensarse que existe un tumor cerebral pasándose por alto que el origen puede estar en otra parte del organismo. No obstante, hay que recalcar que la inmensa mayoría de falsos positivos en este clasificador se debe a errores de propagación. Esto se conoce porque se realizaron pruebas omitiendo los errores de propagación en el cálculo de las métricas de validación, y en estas condiciones se obtuvo una especificidad del 91,17%.

Clasificador 3

En la Tabla 6.12 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 3* (Figura 6.13).

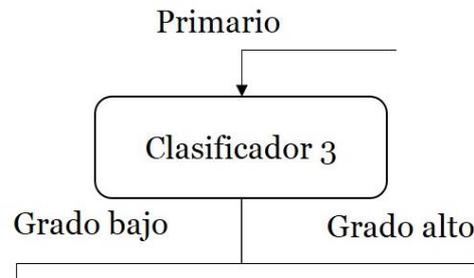


Figura 6.13: *Clasificador 3* del árbol jerárquico de la campaña 2.

Tabla 6.12: Resultados del árbol de clasificación en el *Clasificador 3* para el 70% de muestras de entrenamiento (campaña 2).

Clasificador 3			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
Training = 70%	77,24 (5,79)	66,41 (10,89)	87,68 (8,35)

Este clasificador desempeña la misma tarea que el *Clasificador 2.1* del sistema diseñado para clasificar la campaña 1: discriminar las muestras de grado alto de las muestras de grado bajo. Sin embargo, debe tenerse en cuenta que en este clasificador se ha tomado bajo test a la clase “grado bajo”. Puede observarse en las métricas de sensibilidad y especificidad que la proporción de falsos negativos es bastante mayor que la de falsos positivos (puesto que el promedio de sensibilidad es bastante menor que el promedio de especificidad). Análogamente a lo sucedido en el *Clasificador 2*, podría pensarse que la proporción de muestras de grado bajo que se está clasificando como muestras de grado alto es bastante mayor que la proporción de muestras de grado alto que se está clasificando como grado bajo. Sin embargo, la diferencia proporcional no es tanta si no se contabilizan los errores de propagación ya que en tal caso se obtiene un promedio de sensibilidad del 84,63% frente a un promedio de especificidad del 89,14% con desviaciones típicas muy parecidas a las de esta tabla.

Clasificador 4.1

En la Tabla 6.13 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 4.1* (Figura 6.14).

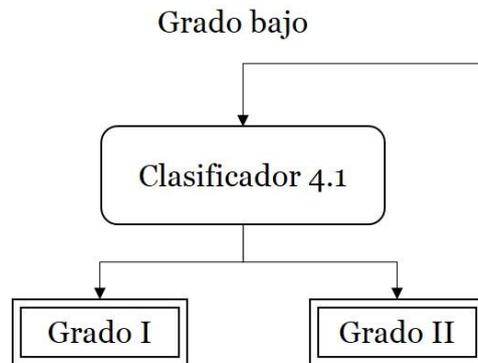


Figura 6.14: *Clasificador 4.1* del árbol jerárquico de la campaña 2.

Tabla 6.13: Resultados del árbol de clasificación en el *Clasificador 4.1* para el 70% de muestras de entrenamiento (campaña 2).

Clasificador 4.1			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
Training = 70%	84,26 (9,81)	78,31 (15,01)	91,06 (11,79)

Los resultados de esta tabla pueden interpretarse como que el *Clasificador 4.1* comete errores tanto al clasificar las muestras de grado I como las muestras de grado II. Sin embargo, se ha observado que todas las métricas presentan valores del 100% si no se contabilizan los errores de propagación. Por tanto, en contraste con los resultados de la tabla, se puede afirmar que este clasificador presenta excelentes resultados al discriminar las muestras de grado I y II del conjunto de datos espectrográficos de la campaña 2.

Clasificador 4.2

En la Tabla 6.14 se muestran los resultados de las clasificaciones realizadas por el *Clasificador 4.2* (Figura 6.15).

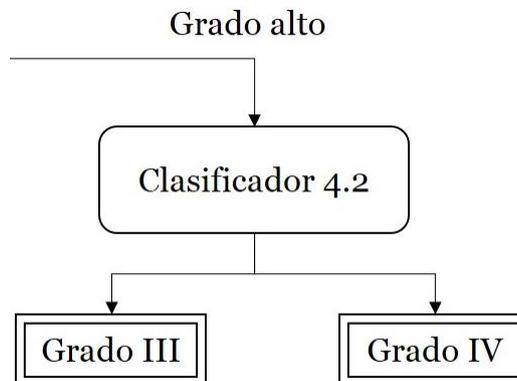


Figura 6.15: *Clasificador 4.2* del árbol jerárquico de la campaña 2.

Tabla 6.14: Resultados del árbol de clasificación en el *Clasificador 4.2* para el 70% de muestras de entrenamiento (campaña 2).

Clasificador 4.2			
Media % (Desviación típica %)	OA	Sensibilidad	Especificidad
Training = 70%	69,86 (6,95)	17,08 (14,94)	96,78 (6,84)

En el *Clasificador 4.2* se ha tomado la clase “grado III” bajo test para obtener los diferentes resultados que se presentan en esta tabla. Puede observarse que un promedio de sensibilidad muy bajo en comparación con el promedio de especificidad. En este caso también se podría incurrir en el error de pensar que el clasificador discrimina muestras de grado III como si fueran de grado IV (proporción alta de falsos negativos), y que discrimina correctamente casi todas las muestras de grado IV (proporción baja de falsos positivos). Por el contrario, si no se contabilizan los errores de propagación, los promedios de sensibilidad (70%) y especificidad (97,38%) indican que la discriminación en este clasificador no es tan drástica.

Del mismo modo que se procedió en la clasificación de la campaña 1 mediante sistemas de clasificación basados en árboles tras las pruebas con el *kernel* lineal, también se realizaron pruebas tratando de clasificar la campaña 2 empleando los *kernels* polinomial y RBF usando sus parámetros por defecto. Sin embargo, los resultados de tales experimentos no se reflejan en esta memoria por considerarse muy deficientes. Esto es

debido a que, con estos *kernels*, el *Clasificador 1* del árbol clasifica todas las muestras como tumorales.

6.3.3. Resumen de resultados más relevantes con sistemas de árboles jerárquicos

En este apartado se recogen los resultados más importantes que se han obtenido en este PFC, en referencia a los diferentes sistemas desarrollados mediante árboles de clasificación jerárquica.

En la Tabla 6.15 se presentan los promedios y las desviaciones típicas de OA, sensibilidad y especificidad para *trainings* del 70% y 30% en la campaña 1, mientras que en la Tabla 6.16 se reflejan estas métricas para un *training* del 70% en la campaña 2.

Tabla 6.15: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad, en cada uno de los clasificadores para *trainings* del 30% y 70% (campaña 1).

Sistema de clasificación jerárquico (campaña 1)					
Training	Clase bajo test	OA	Sensibilidad	Especificidad	Clasificador (Nivel)
70%	Tumor	100 (0)	100 (0)	100 (0)	<p>Muestras de la campaña 1</p> <p>↓</p> <p>Clasificador 1</p> <p>Tumor Sanas</p> <p>Clasificador 1 (Nivel 1)</p>
30%		100 (0)	100 (0)	100 (0)	
70%	Grado alto	95,85 (2,44)	99,26 (1,74)	72,62 (19,09)	<p>Tumor</p> <p>↓</p> <p>Clasificador 2.1</p> <p>Grado alto Grado bajo</p> <p>Grado II</p> <p>Clasificador 2.1 (Nivel 2)</p>
30%		91,99 (2,44)	97,1 (3,02)	54,1 (18,09)	
70%	Sustancia gris	100 (0)	100 (0)	100 (0)	<p>Sanas</p> <p>↓</p> <p>Clasificador 2.2</p> <p>Sustancia gris Sustancia blanca</p> <p>Clasificador 2.2 (Nivel 2)</p>
30%		94,09 (9,96)	99,2 (3,96)	91 (16,21)	
70%	Grado IV	85,22 (3,25)	99,16 (2,41)	28,57 (10,29)	<p>Grado alto</p> <p>↓</p> <p>Clasificador 3</p> <p>Grado III Grado IV</p> <p>Clasificador 3 (Nivel 3)</p>
30%		81,18 (3,84)	97,59 (3,61)	18,06 (9,37)	

Tabla 6.16: Promedio y desviación típica de las métricas: overall accuracy, sensibilidad y especificidad, en cada uno de los clasificadores para un training del 70% (campaña 2).

Sistema de clasificación jerárquico (campaña 2)				
Clase bajo test	OA	Sensibilidad	Especificidad	Clasificador (Nivel)
Tumor	90,47 (3,72)	99,27 (2,78)	37,93 (20,58)	<p>Muestras de la campaña 2</p> <p>↓</p> <p>Clasificador 1</p> <p>Tumor No tumoral</p> <p>Clasificador 1 (Nivel 1)</p>
Tumor primario	89,22 (3,64)	99,38 (1,77)	48,24 (12,77)	<p>Tumor</p> <p>↓</p> <p>Clasificador 2</p> <p>Primario Secundario</p> <p>Clasificador 2 (Nivel 2)</p>
Grado bajo	77,24 (5,79)	66,41 (10,89)	87,68 (8,35)	<p>Primario</p> <p>↓</p> <p>Clasificador 3</p> <p>Grado bajo Grado alto</p> <p>Clasificador 3 (Nivel 3)</p>
Grado I	84,26 (9,81)	78,31 (15,01)	91,06 (11,79)	<p>Grado bajo</p> <p>↓</p> <p>Clasificador 4.1</p> <p>Grado I Grado II</p> <p>Clasificador 4.1 (Nivel 4)</p>
Grado III	69,86 (6,95)	17,08 (14,94)	96,78 (6,84)	<p>Grado alto</p> <p>↓</p> <p>Clasificador 4.2</p> <p>Grado III Grado IV</p> <p>Clasificador 4.2 (Nivel 4)</p>

Al compararse los resultados obtenidos entre los niveles más bajos (los más relevantes) de las dos campañas, se puede dilucidar que se ha conseguido clasificar mucho

mejor las muestras de la campaña 1 que las de la campaña 2. Estos resultados no sorprenden si nos fijamos en las formas de las firmas espectrales de las muestras (apartado 6.2), puesto que las firmas espectrales de cada clase de las muestras de la campaña 1 manifiestan a simple vista, en su forma, un patrón mucho más definido que las muestras de la campaña 2.

6.4 Resultados en sistemas de clasificación multiclase

Con el objetivo de obtener unos resultados más alentadores que los obtenidos con el método basado en árboles de decisión, se trató de afrontar este problema desde otra perspectiva. Es por esta razón por la que se decide diseñar los sistemas empleando el método de clasificación multiclase.

Los experimentos efectuados con los sistemas de clasificación multiclase, se realizaron con la campaña 1 en su respuesta en absorbancia y, con la campaña 2 y la campaña fusión, en sus respuestas en absorbancia y transmitancia.

A diferencia de los árboles de decisión, en los que se hacen clasificaciones binarias, en este método se realizan clasificaciones de más de dos clases simultáneamente. Por tanto, se hace necesario calcular las métricas de sensibilidad y especificidad tomando a cada una de las clases como clase bajo test. Por esta razón, se presentan promedios de sensibilidad, promedios de especificidad, y sus respectivas desviaciones típicas, para cada una de las clases del sistema.

Debe recalarse para el caso multiclase lo siguiente: se considera un positivo, a una muestra de la clase bajo test; y un negativo, a una muestra que pertenece a cualquiera de las otras clases (apartado 5.4). De este modo, un verdadero positivo indica una muestra correctamente clasificada, sin embargo, a diferencia de los sistemas basados en árboles de decisión, un verdadero negativo indica que se ha rechazado correctamente de la clase bajo test, pero esto no implica que se haya clasificado correctamente en la clase que le pertenece. Este sesgo en la información es complementado al considerar el resto de clases bajo test.

No obstante, esto también supone que se deban hacer consideraciones diferentes para cada clase, incluso ante valores similares en sus métricas. Por ejemplo: en la campaña 1 se cuenta con 120 muestras de grado IV frente a 62 muestras que no son de grado IV; en contraste, se cuenta con 6 muestras de sustancia gris frente a 176 muestras que no son de sustancia gris.

Por tanto, ante un 50% de sensibilidad y un 50% de especificidad; tomando la clase “grado IV” como clase bajo test, se indican 60 verdaderos positivos (muestras de grado IV correctamente clasificadas) y 31 verdaderos negativos (muestras de otras clases que son correctamente rechazadas de esta clase); mientras que tomando la clase “sustancia gris” como clase bajo test, se indican 3 de verdaderos positivos (muestras de sustancia gris clasificadas como tales) y 88 verdaderos negativos (muestras de otras clases correctamente rechazadas de esta clase).

En resumen: con los mismos porcentajes de ejemplo en las métricas, para la clase “grado IV” bajo test se tienen más verdaderos positivos (60) que verdaderos negativos (31), y para la clase “sustancia gris” bajo test, se tienen menos verdaderos positivos (3) que verdaderos negativos (88).

Teniendo en cuenta lo anterior, en las explicaciones que se dan sobre los resultados obtenidos en los diferentes experimentos con sistemas multiclase, las comparaciones se hacen en relación al número de muestras de la clase que se esté considerando en cada momento y al número de muestras del resto de las clases que no están tomando bajo test.

6.4.1. Resultados en sistemas de clasificación multiclase (campana 1)

En la Tabla 6.17 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase, empleando el *kernel* de tipo lineal, para las muestras definidas en absorbancia.

Tabla 6.17: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* lineal (campaña 1 - absorbanca).

Sistema de clasificación multiclase (campaña 1 - absorbanca)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado II	85,24 (17,04)		98,02 (2,14)
Grado III	33,68 (12,18)		99,5 (1,66)
Grado IV	97,62 (2,81)		68,72 (9,14)
Sustancia gris	100 (0)		100 (0)
Sustancia blanca	100 (0)		100 (0)
Overall Accuracy (OA)	87,64 (3,22)	Tipo de kernel:	Lineal

Como puede deducirse del valor promedio de OA, se obtiene un porcentaje de acierto en las clasificaciones bastante alto.

Además, las medias de especificidad y sensibilidad de las clases de muestras sanas ostentan los máximos valores. De esto se deduce que ninguna muestra sana se confunde con tumoral ni viceversa, y que tampoco se confunde ninguna muestra de sustancia blanca o sustancia gris.

Por otro lado, se tiene un promedio de especificidad de “grado III” muy alto. Esto indica que una proporción muy baja de muestras que no pertenecen a esta clase se están clasificando como tal. Sin embargo, la sensibilidad de esta clase indica que una proporción importante de muestras de grado III se están clasificando en otra/s categoría/s.

Por otra parte, el alto promedio de sensibilidad de la clase “grado IV”, indica que la tasa de muestras de este grado que se clasifican con otra etiqueta es muy baja (falsos negativos). Esto es importante porque se reduce la probabilidad de que un tumor avanzado pase desapercibido, sin embargo, el promedio de su especificidad indica que tejidos de otro tipo se pueden confundir con tumor de este grado (68,72%).

Por último, destacan las métricas de la clase “grado II” por ser altas y las más balanceadas de entre las clases tumorales. Esto implica que el *cluster* de esta clase está bien delimitado, aunque no del todo, por los hiperplanos creados por este sistema.

Posteriormente, se realizaron experimentos empleando el *kernel* RBF tomando sus parámetros por defecto.

En la Tabla 6.17 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase, empleando el *kernel* de tipo RBF para las muestras definidas en absorbancia.

Tabla 6.18: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* RBF (campaña 1 - absorbancia).

Sistema de clasificación multiclase (campaña 1 - absorbancia)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado II	0 (0)		100 (0)
Grado III	0 (0)		100 (0)
Grado IV	100 (0)		21,84 (3,34)
Sustancia gris	100 (0)		100 (0)
Sustancia blanca	100 (0)		100 (0)
Overall Accuracy (OA)	74,57 (1,17)	Tipo de kernel:	RBF (parámetros por defecto)

Aunque el promedio de OA puede aparentar una buena clasificación de las muestras *a priori*, al contrastarse los promedios de sensibilidad y especificidad resaltan algunos aspectos negativos que deben tomarse en consideración.

Análogamente al experimento con el *kernel* lineal, de las métricas para las muestras sanas se deduce que el sistema ha discriminado perfectamente las muestras sanas de las tumorales.

Sin embargo, las métricas para “grado II” y “grado III” indican que para estas clases no se predice ningún positivo, en otras palabras, ninguna muestra se clasifica como grado II o grado III.

Por otro lado, en las métricas de las muestras de grado IV se refleja que todas las muestras, a excepción de las sanas, se clasifican como grado IV. De la especificidad de las clases de muestras sanas se deduce que ninguna muestra tumoral se clasifica como sana,

por tanto, la única opción es que las muestras de grado II y III se hayan clasificado en la clase “grado IV”, así como todas las muestras que pertenecen a esta clase.

Se puede concluir de los experimentos que se reflejan en la Tabla 6.18 (*kernel* RBF con parámetros por defecto), que estas clasificaciones clasifican correctamente a las muestras sanas y a todas las tumorales las clasifica como muestras de grado IV.

6.4.1.1. Resultados de la optimización de parámetros del *kernel* RBF (campaña 1)

Puesto que el empleo de otros tipos de *kernels* con sus parámetros por defecto no proporcionó buenos resultados para las clases tumorales, se decidió realizar una búsqueda de los parámetros óptimos del *kernel* RBF.

Como se explicó en el apartado 5.7.1, se hizo el primer barrido partiendo de los mismos valores del artículo [110].

Barrido 1:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-5 \leq n \leq 13$ y $-5 \leq m \leq 15$ con n y m impar.

A continuación, se presentan las superficies promedio de sensibilidad y especificidad tomando como clase bajo test la clase “grado IV”, para este barrido de valores de parámetros (Figura 6.16)²⁵.

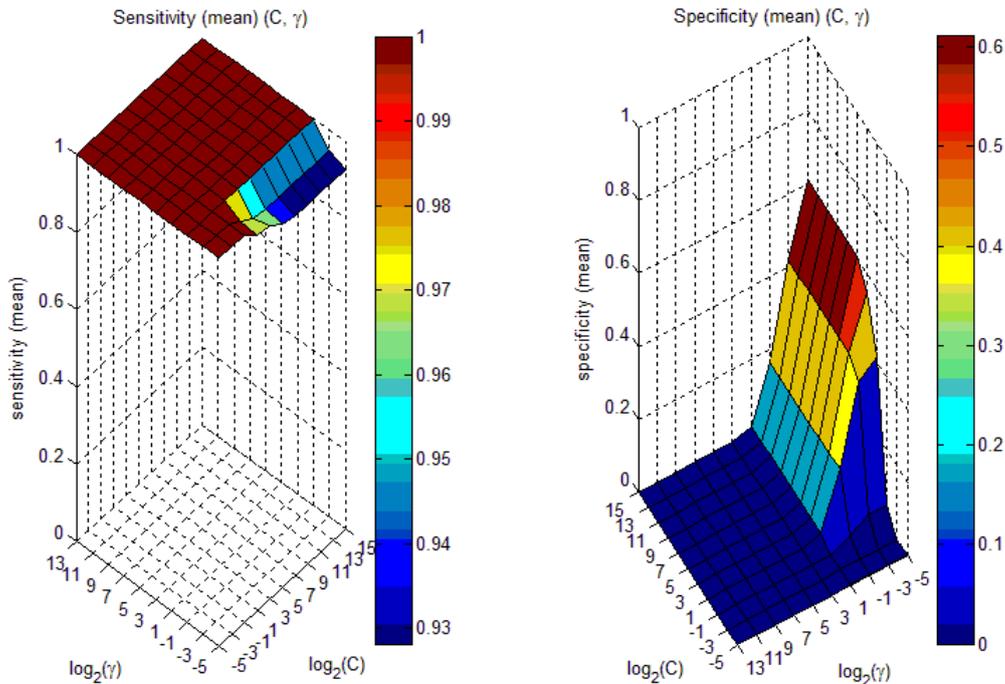


Figura 6.16: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 1).

Como puede observarse en las gráficas, para este primer barrido en (γ, C) se obtienen promedios de sensibilidad muy altos, mientras que las medias de especificidad más altas llegan en torno al 60%. Esto es: aunque se está clasificando correctamente una buena proporción de muestras de grado IV, también se está clasificando como muestras de grado IV una gran proporción de muestras que no pertenecen a esta clase, con lo que la capacidad de rechazo del clasificador está en detrimento, en esta zona del plano (γ, C) .

Además, puede apreciarse que los valores medios de especificidad comienzan a crecer considerablemente entorno a valores de $\log_2(\gamma)$ inferiores a 1. Sin embargo, el promedio de sensibilidad comienza a decaer para valores de $\log_2(\gamma)$ inferiores de -1. Nótese que las gráficas ilustradas en la Figura 6.16 se han representado con cierta rotación relativa entre ellas (en un eje vertical), para ver mejor la tendencia de ambas superficies.

²⁵ Se ha decidido expresar los parámetros en su escala logarítmica, $\log_2(\blacksquare)$, y, la especificidad y la sensibilidad en tanto por uno, para hacer más cómoda la visualización de las gráficas en este estudio.

Ante estos resultados, surge la idea de buscar en una zona del plano (γ, C) en el que las métricas están más balanceadas. Según los valores representados, se intuye que la tendencia de ambas superficies que propicia este hecho, ocurre para valores decrecientes de $\log_2(\gamma)$.

No obstante, se decide comenzar extendiendo el barrido hacia valores crecientes de $\log_2(\gamma)$ y $\log_2(C)$, con el fin de asegurar que la tendencia de las superficies se mantiene constante en regiones crecientes cercanas. Para ello, se establece el “barrido 2”.

Barrido 2:

- $\gamma = \{2^n\}$ y $C = \{2^n\}$ donde $-5 \leq n \leq 31$ con n impar.

En la Figura 6.17 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 2.

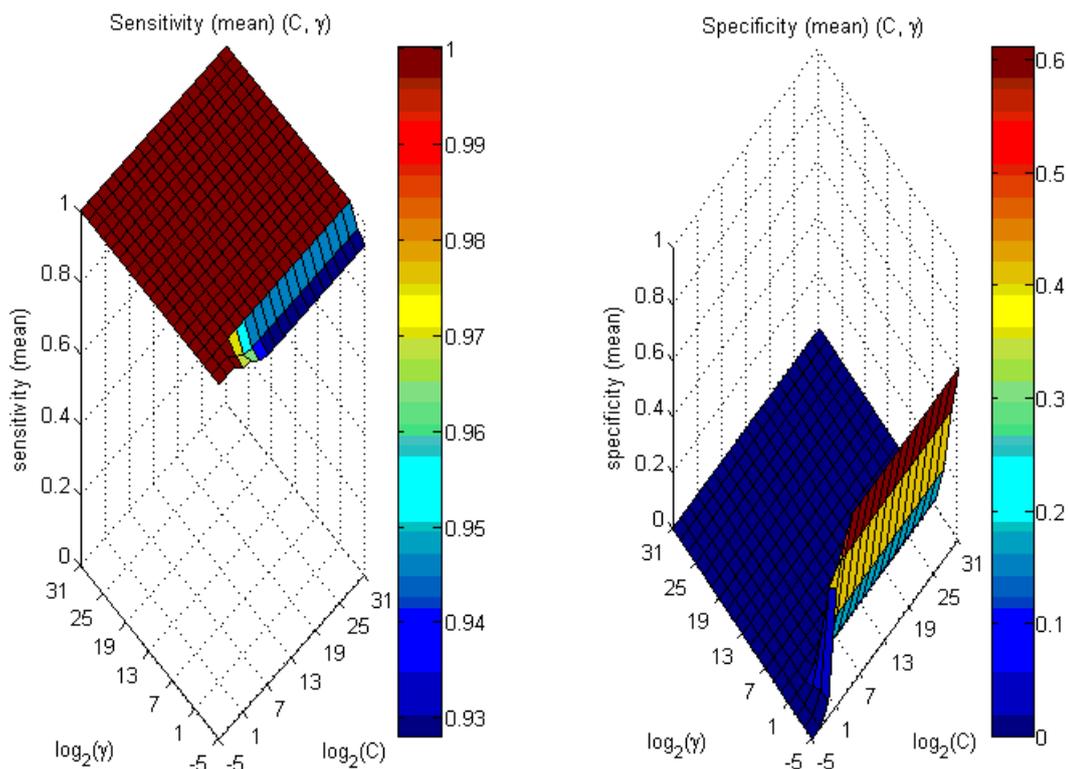


Figura 6.17: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 2).

Tras comprobarse que para esta región del plano (γ, C) las métricas se mantienen constantes, se comprobó además que la misma tendencia se mantiene para valores decrecientes de $\log_2(C)$ (barrido 3).

Barrido 3:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-5 \leq n \leq 11$ y $-21 \leq m \leq 1$ con n y m impar.

En la Figura 6.18 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 3.

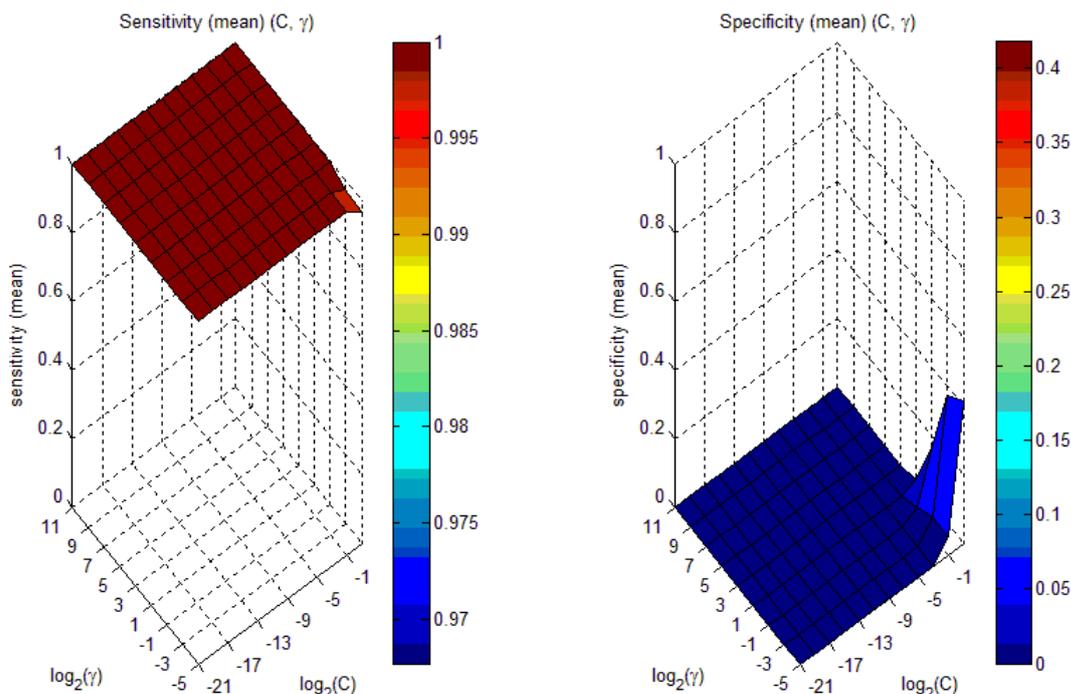


Figura 6.18: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 3).

Puede observarse que en esta región del plano de parámetros también se obtiene promedios de especificidad nulos.

Llegados a este punto cabe preguntarse si puede haber alguna zona del plano no explorada en este proyecto, donde exista una combinación de parámetros, (γ, C) , que propicie los mejores resultados de sensibilidad y especificidad. En este PFC se ha optado por tratar de buscar los parámetros óptimos en función de las superficies de sensibilidad y especificidad obtenidas, en las regiones exploradas.

Por otro lado, a la vista de que los resultados en los barridos 2 y 3 se mantienen constantes, se trató de averiguar si para valores de $\log_2(\gamma) < -5$ y valores de $\log_2(C) > 3$, las superficies de promedios de sensibilidad y especificidad nos indican regiones de parámetros que mejoren las clasificaciones. Este razonamiento propone el siguiente barrido.

Barrido 4:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-19 \leq n \leq -5$ y $3 \leq m \leq 13$ con n y m impar.

En la Figura 6.19 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 4.

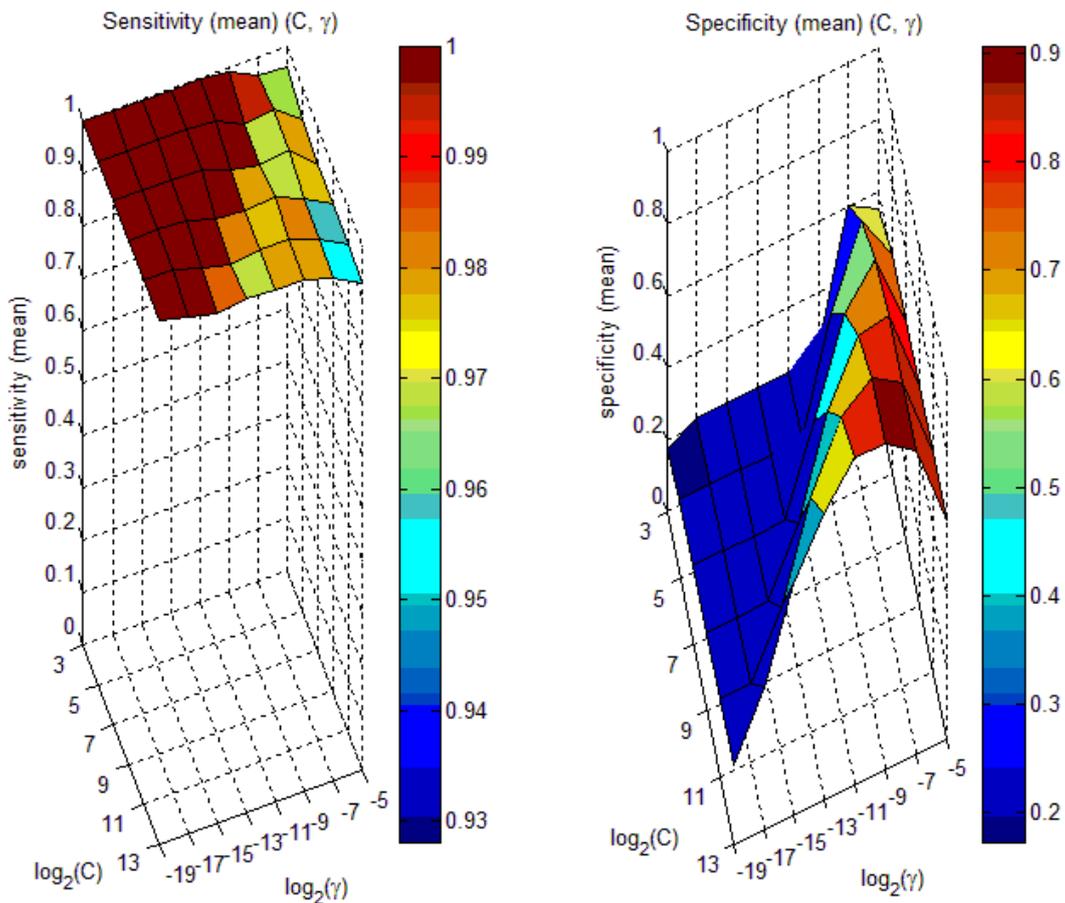


Figura 6.19: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 4).

En la superficie de promedio de especificidad puede observarse que en torno a valores de $\log_2(\gamma)$ inferiores a -13, se tiene una caída de esta superficie. Por tanto, se hace otro barrido a partir de ese valor de modo que quede centrada la "cresta" de la superficie promedio de especificidad.

Barrido 5:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-13 \leq n \leq -3$ y $3 \leq m \leq 15$ con n y m impar.

En la Figura 6.20 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 5.

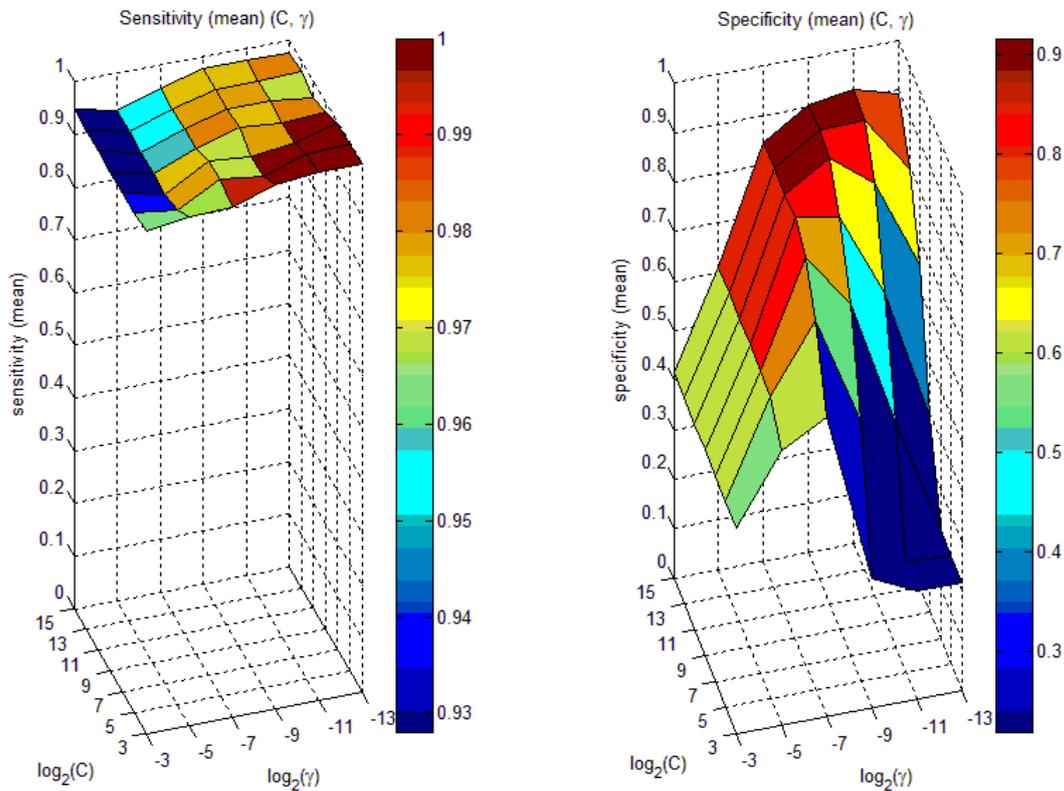


Figura 6.20: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 5).

En estas gráficas puede observarse que hay zonas comunes del plano (γ, C) en las que tanto el promedio de sensibilidad como el de especificidad presentan valores altos. Además, en la superficie promedio de especificidad se aprecia una tendencia a aumentar con el coste sin que vaya en detrimento de la sensibilidad. Para observar lo que ocurre al aumentar aún más el coste, se establece el siguiente barrido.

Barrido 6:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-19 \leq n \leq -1$ y $3 \leq m \leq 43$ con n y m impar.

En la Figura 6.21 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 6.

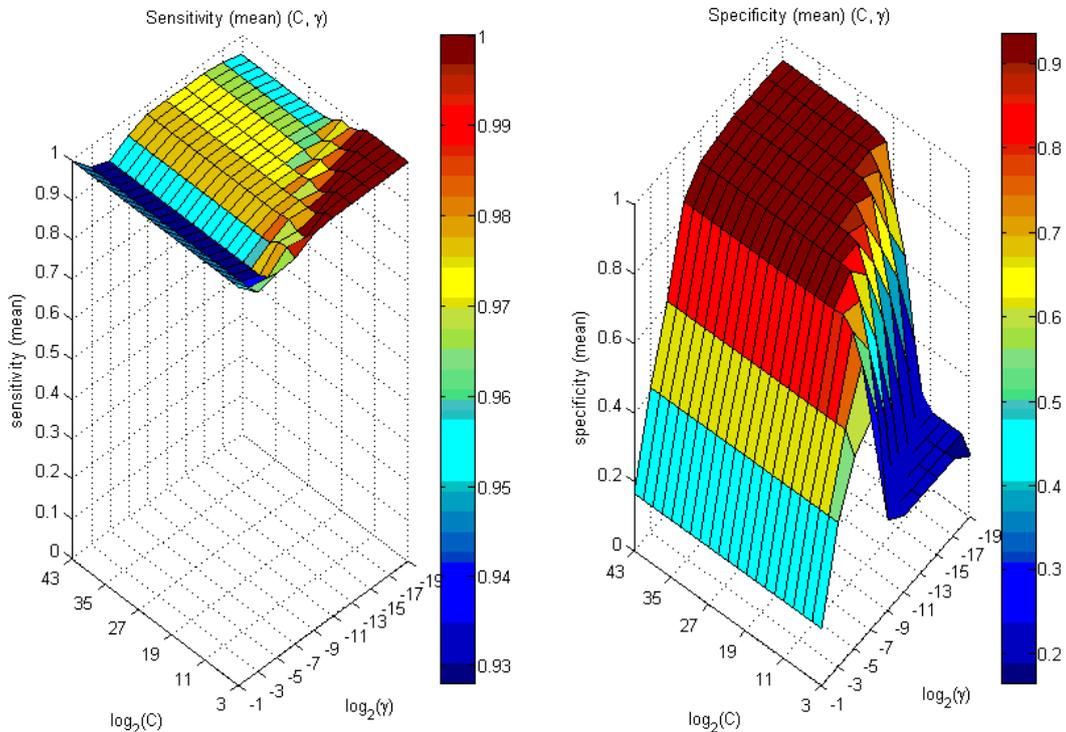


Figura 6.21: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 6).

Como puede observarse en las gráficas, al aumentar el coste se obtienen promedios de especificidad muy altos sin verse muy afectada la métrica de sensibilidad. Esta situación propone ampliar el rango de coste en un nuevo barrido.

Barrido 7:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-19 \leq n \leq -1$ y $43 \leq m \leq 75$ con n y m impar.

En la Figura 6.22 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 7.

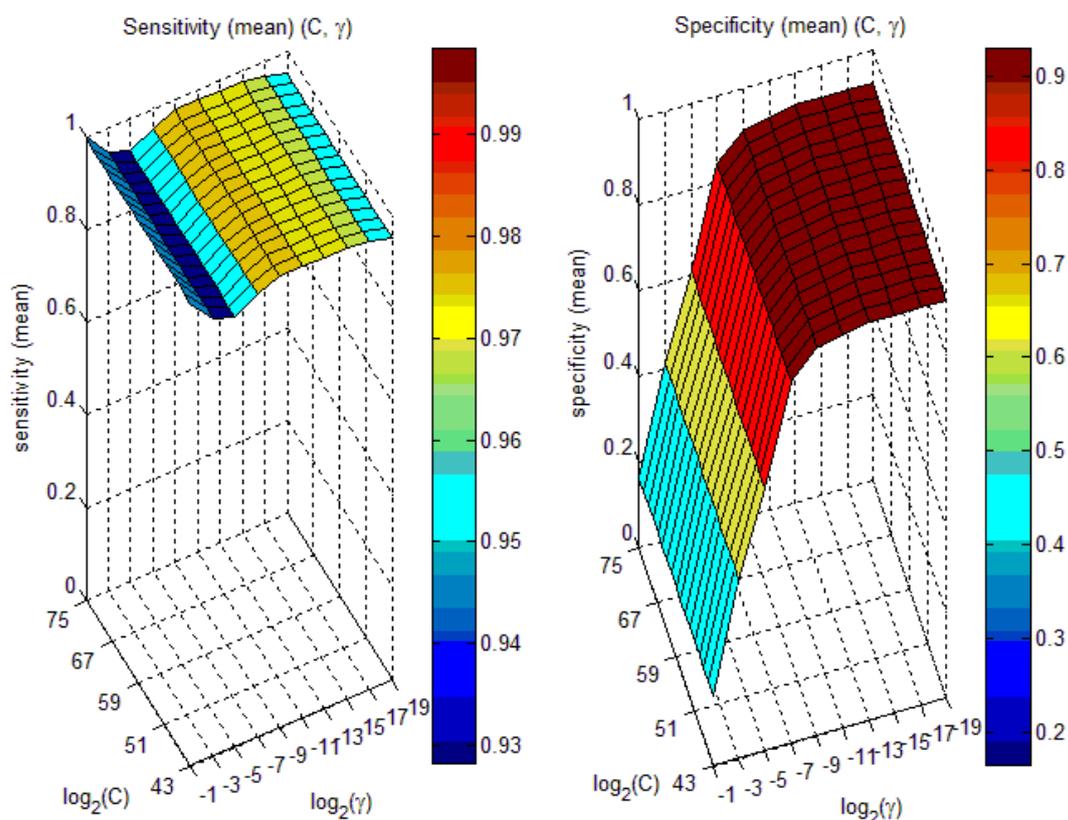


Figura 6.22: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 7).

En estas gráficas se puede observar que los promedios de sensibilidad y especificidad varían mínimamente para valores de γ constantes. Nótese que el límite superior en el rango del coste ya toma un valor muy alto, $C = 2^{75}$, y la tendencia de las superficies al incrementar este parámetro permanece igual, constante. Por esta razón, se ha decidido restringir la búsqueda en esta zona del plano (γ, C), ya que en estos dos últimos barridos es donde se han obtenidos los mejores promedios de sensibilidad y especificidad para la clase de grado IV.

Al analizar en detalle los resultados obtenidos en estos dos barridos, destacan tres pares de valores de medias de sensibilidad y especificidad por su alto porcentaje en ambas magnitudes.

En la Tabla 6.19 se exponen los tres pares de promedios de sensibilidad y especificidad, tomando las muestras de grado IV como clase bajo test, que mejores

resultados presentan tras realizarse la búsqueda de parámetros óptimos, y los valores de los parámetros que los proporcionan.

Tabla 6.19: Pares de medias de sensibilidad y especificidad con la clase "grado IV" bajo test, que mejores resultados presentan, y los parámetros que los proporcionan.

	Sensibilidad (Media)	Especificidad (Media)	Parámetro γ	Parámetro C
PAR 1	97,4%	92,77%	$\{2^{-13}\}$	$\{2^n\} \ 17 \leq n \leq 75, n \text{ impar.}$
PAR 2	96,67%	92,89%	$\{2^{-17}\}$	$\{2^n\} \ 23 \leq n \leq 75, n \text{ impar.}$
PAR 3	96,51%	93,33%	$\{2^{-17}\}$	$\{2^{21}\}$

Al aplicar la expresión (5.1) para cada uno de los tres pares de promedios de sensibilidad y especificidad, se obtiene el factor más alto con el "par 3". Y como puede verse en la tabla, este par es proporcionado unívocamente por los parámetros siguientes:

$$(\gamma, C) = (2^{-17}, 2^{21}) \quad (6.1)$$

Una vez obtenidos estos valores, se realiza un ajuste fino mediante un último barrido de parámetros para comprobar si en zonas aledañas es posible mejorar el resultado.

Barrido 8 (ajuste fino):

- $\gamma = \{2^{-18}, 2^{-17,5}, 2^{-17}, 2^{-16,5}, 2^{-16}\};$
- $C = \{2^{20}, 2^{20,5}, 2^{21}, 2^{21,5}, 2^{22}\}.$

En la Figura 6.23 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 8.

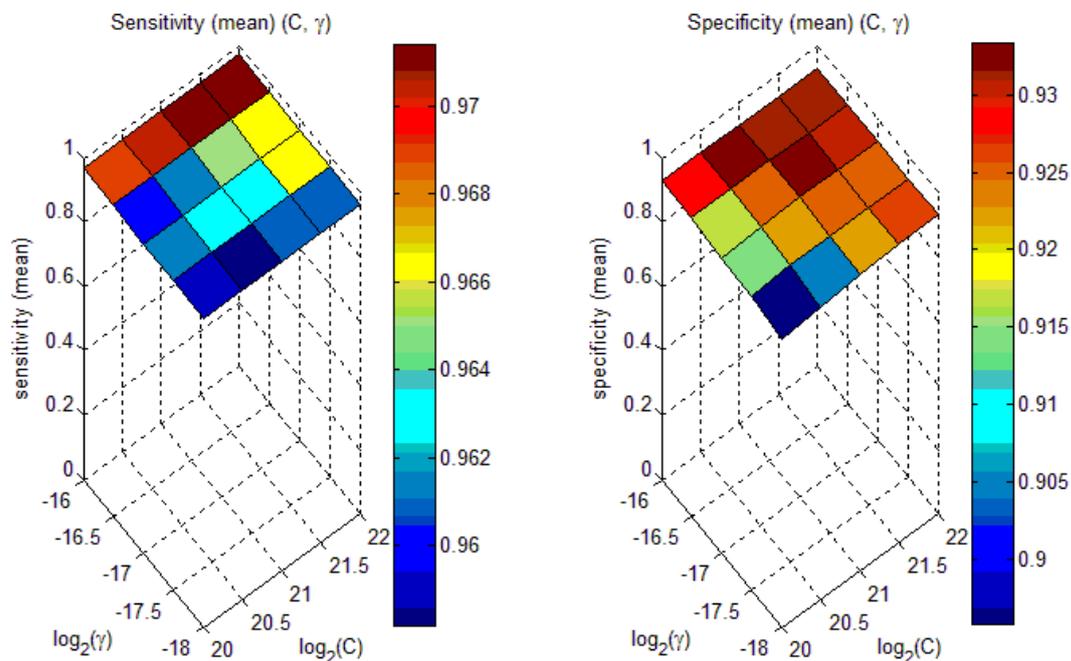


Figura 6.23: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 8 - ajuste fino).

Tras realizar el ajuste fino de los parámetros, se pudo comprobar que, con ninguna de las combinaciones de parámetros empleadas en este barrido, se obtuvo mejores resultados de promedios de sensibilidad y especificidad que el “par 3”. Nótese que los promedios de sensibilidad y especificidad en estas gráficas están por encima del 90%.

De este estudio se deduce que los parámetros $(\gamma, C) = (2^{-17}, 2^{21})$ son los parámetros óptimos del *kernel* RBF tomando la clase “grado IV” bajo test. No obstante, queda comprobar si con estos parámetros óptimos se obtienen buenos resultados tomando como clase bajo test al resto de las clases de esta campaña.

Finalmente, en la Tabla 6.20 se exponen los resultados de las medias de OA, sensibilidad y especificidad, al tomar como clase bajo test a cada una las clases de la campaña 1, tras establecer los parámetros óptimos en el *kernel* RBF de este sistema de clasificación multiclase.

Tabla 6.20: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* RBF y parámetros óptimos (campaña 1 - absorbancia).

Sistema de clasificación multiclase (campaña 1 - absorbancia)					
Media % (Desviación típica %)	Sensibilidad			Especificidad	
Clase bajo test					
Grado II	91,48 (12,86)			98,38 (1,87)	
Grado III	83,61 (12,4)			97,68 (2,32)	
Grado IV	96,51 (3,1)			93,33 (6,53)	
Sustancia gris	100 (0)			100 (0)	
Sustancia blanca	100 (0)			100 (0)	
Overall Accuracy (OA)	94,39 (2,91)	Tipo de kernel:	RBF	(γ, C)	$(2^{-17}, 2^{21})$

Como puede observarse en estos resultados, no solo se ha logrado obtener unos buenos porcentajes promedio de sensibilidad y especificidad para la clase “grado IV”, sino también para el resto de las clases. Esto indica que con este sistema se consiguen buenos niveles de discriminación para todas las clases, lográndose una discriminación perfecta entre las muestras tumorales y sanas de esta campaña; y en las muestras sanas, entre las muestras de sustancia blanca y gris. Esto se deduce de los promedios del 100% en las clases: “sustancia blanca” y “sustancia gris”.

6.4.2. Resultados en sistemas de clasificación multiclase (campaña 2 - transmitancia)

Después de realizar el estudio de los parámetros óptimos para el *kernel* RBF del sistema de clasificación multiclase de la campaña 1, se procedió a diseñar un sistema multiclase para clasificar a las muestras de la campaña 2.

En la Tabla 6.21 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase de la campaña 2, empleando el *kernel* de tipo lineal, para las muestras definidas en transmitancia.

Tabla 6.21: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* lineal (campaña 2 - transmitancia).

Sistema de clasificación multiclase (campaña 2 - transmitancia)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado I	100 (0)		98,08 (2,69)
Grado II	79,43 (14,7)		93,37 (5,16)
Grado III	93 (20,23)		98,09 (2,35)
Grado IV	76,31 (12,9)		85,36 (8,86)
Tumor secundario	83,83 (18,85)		99,42 (1,31)
No tumoral	62,4 (22,5)		96,64 (4,26)
Overall Accuracy (OA)	79,05 (6,22)	Tipo de kernel:	Lineal

En este caso, el valor promedio de OA sugiere que en general se consigue un porcentaje de acierto en las clasificaciones bastante satisfactorio.

Sin embargo, al observar los promedios de especificidad y sensibilidad de la clase “no tumoral” destaca que este sistema clasifica algunas muestras no tumorales como tumorales (falsos negativos) y viceversa (falsos positivos).

Por otro lado, del promedio de sensibilidad de las muestras de grado I se deduce que todas las muestras de grado I se clasifican correctamente, y de la media de especificidad, que una muy baja proporción de muestras que no pertenecen a esta clase, se clasifica como grado I.

De las muestras de grado II, III y IV se hace una clasificación relativamente buena ya que de sus promedios de sensibilidad se deduce que las proporciones de falsos negativos no son muy considerables.

Por otra parte, si observamos el promedio de las métricas de tumor secundario, podemos deducir que la proporción de muestras con metástasis que no se detectan es relativamente baja, mientras que la proporción de muestras que no son metástasis y se confunden con metástasis, es muy baja.

De la misma manera que se procedió con las muestras de la campaña 1, después de efectuar los experimentos con el *kernel* lineal, se realizaron experimentos empleando el *kernel* RBF tomando sus parámetros por defecto.

En la Tabla 6.22 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase, empleando el *kernel* de tipo RBF, para las muestras definidas en transmitancia.

Tabla 6.22: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* RBF (campaña 2 - transmitancia).

Sistema de clasificación multiclase (campaña 2 - transmitancia)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado I	0 (0)		100 (0)
Grado II	0 (0)		100 (0)
Grado III	0 (0)		100 (0)
Grado IV	100 (0)		0 (0)
Tumor secundario	0 (0)		100 (0)
No tumoral	0 (0)		100 (0)
Overall Accuracy (OA)	40,42 (1,28)	Tipo de kernel:	RBF (parámetros por defecto)

Como puede verse, el porcentaje de acierto global indica que este sistema presenta gran deficiencia en las discriminaciones realizadas, y al analizarse los promedios de sensibilidad y especificidad, se concluye que se trata de un sistema completamente ineficiente. Este sistema ni siquiera es válido para discriminar entre muestras tumorales y no tumorales.

De las métricas para el grado IV se deduce que todas las muestras de esta clase son correctamente clasificadas (no hay falsos negativos al ser la sensibilidad del 100% en las 50 clasificaciones) y que el clasificador no rechaza correctamente a ninguna muestra (no hay verdaderos negativos ya que la especificidad es cero en las 50 clasificaciones). En otras palabras, el sistema clasifica a todas las muestras de test como muestras de grado IV.

“Grado IV” es la clase que dispone de más muestras, 53 de las 128 totales de esta campaña. Por esta razón se obtiene un valor promedio de OA un tanto elevado para tratarse de un sistema tan ineficiente.

Por otro lado, resalta el hecho de que para la métrica de OA se obtenga una desviación típica de 1,28% en un sistema que en las 50 repeticiones clasifica todas las muestras como grado IV. Esto se debe a que el algoritmo que establece el *training* y el *test* para un porcentaje especificado, presenta una componente aleatoria al escoger el número de muestras de cada clase.

Esto ocurre por ejemplo a la hora de establecer el *training* de la clase “grado IV”, puesto que el 70% de las 53 muestras no resulta un valor entero, con lo que hay una pequeña variación en relación a la cantidad de muestras escogidas en las 50 repeticiones para cada clase.

En definitiva, no se emplea exactamente la misma cantidad de muestras de grado IV en el *test* de cada repetición. En algunas repeticiones se emplean 15 muestras de grado IV, en otras 16, etc. (el 30% de 53 es 15,9). Por eso al aplicar la ecuación (3.39) no se obtiene el mismo resultado en todas las repeticiones.

6.4.2.1. Resultados de la optimización de parámetros del *kernel* RBF (campaña 2)

Del mismo modo que se realizó el estudio de la campaña 1 para obtener los parámetros óptimos, se procedió a realizar un estudio similar para obtener los parámetros óptimos con las muestras de la campaña 2; con el objetivo de mejorar los resultados obtenidos en las clasificaciones de esta campaña.

Como se explicó en el apartado 5.7.2, se tomó como punto de partida los parámetros óptimos obtenidos para la campaña 1: $(\gamma, C) = (2^{-17}, 2^{21})$

Antes de hacer los diferentes barridos de parámetros con los datos de esta campaña, se realizó una clasificación con estos valores para comprobar si ofrecían buenos resultados de partida. En la Tabla 6.23 se muestran los promedios de las métricas con la clase de grado IV bajo test para estos valores.

Tabla 6.23: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (para la clase "grado IV"), con tipo de kernel RBF y parámetros óptimos de la campaña 1 (campaña 2 - transmitancia).

Sistema de clasificación multiclase (campaña 2 - transmitancia)					
Media % (Desviación típica %)	Sensibilidad			Especificidad	
Clase bajo test					
Grado IV	89,76 (9,61)			94,33 (4,51)	
Overall Accuracy (OA)	89,53 (5,09)	Tipo de kernel:	RBF	(γ, C)	$(2^{-17}, 2^{21})$

Los resultados de la tabla reflejan que los valores de partida escogidos, $(2^{-17}, 2^{21})$, presentan promedios de sensibilidad y especificidad altos. Estos resultados sugieren que la búsqueda de los valores óptimos de estos parámetros, será más rápida con esta campaña puesto que se parte de mejores condiciones.

En el primer barrido de parámetros para esta campaña, se toma los parámetros óptimos de la campaña 1 como punto central del barrido y se establece un rango de valores escogido sin muchos preámbulos.

Barrido 1:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-27 \leq n \leq -7$ y $11 \leq m \leq 31$ con n y m impar.

En la Figura 6.24 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 1.

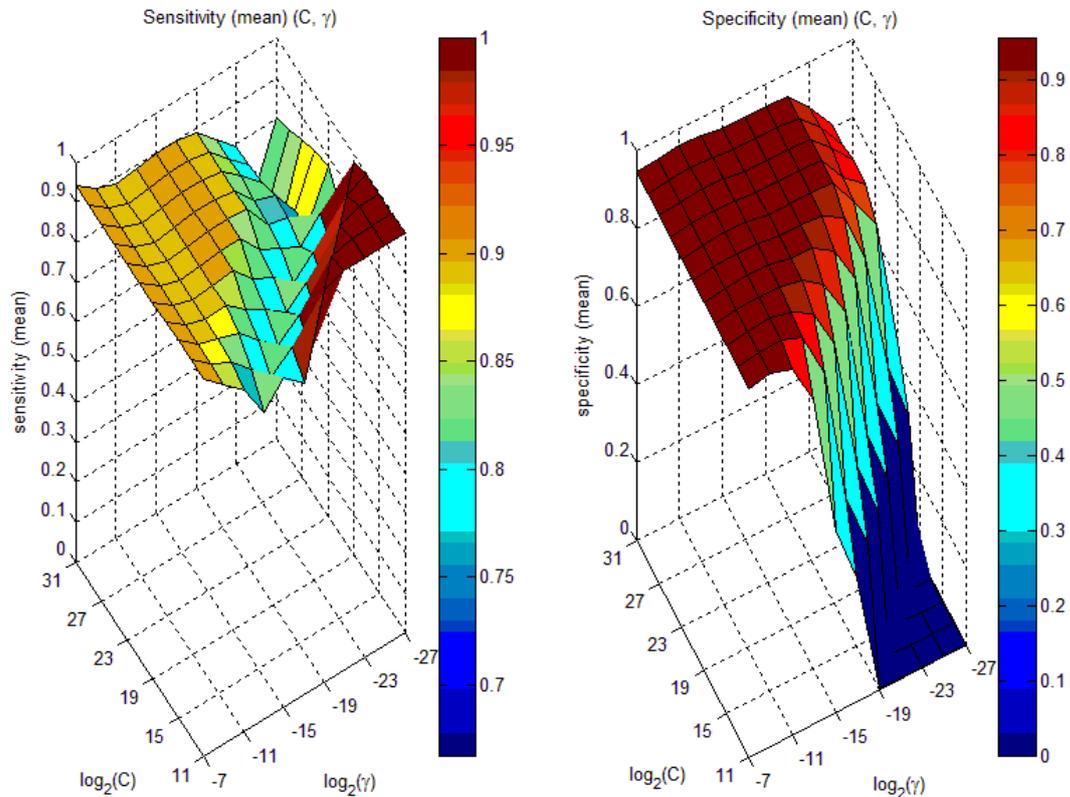


Figura 6.24: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 1).

Como puede observarse en estas gráficas, a medida que decrementa γ la media de sensibilidad presenta una tendencia a aumentar, sin que por ello se vean comprometidos los valores medios de especificidad puesto que estos no reflejan un decremento en esta zona.

Por tanto, el siguiente paso es comprobar lo que ocurre con las medias de sensibilidad y de especificidad cuando se aumenta aún más los valores de γ , manteniendo el coste, C , en el mismo rango. Para ello se realiza el siguiente barrido.

Barrido 2:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-7 \leq n \leq 13$ y $11 \leq m \leq 31$ con n y m impar.

En la Figura 6.25 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 2.

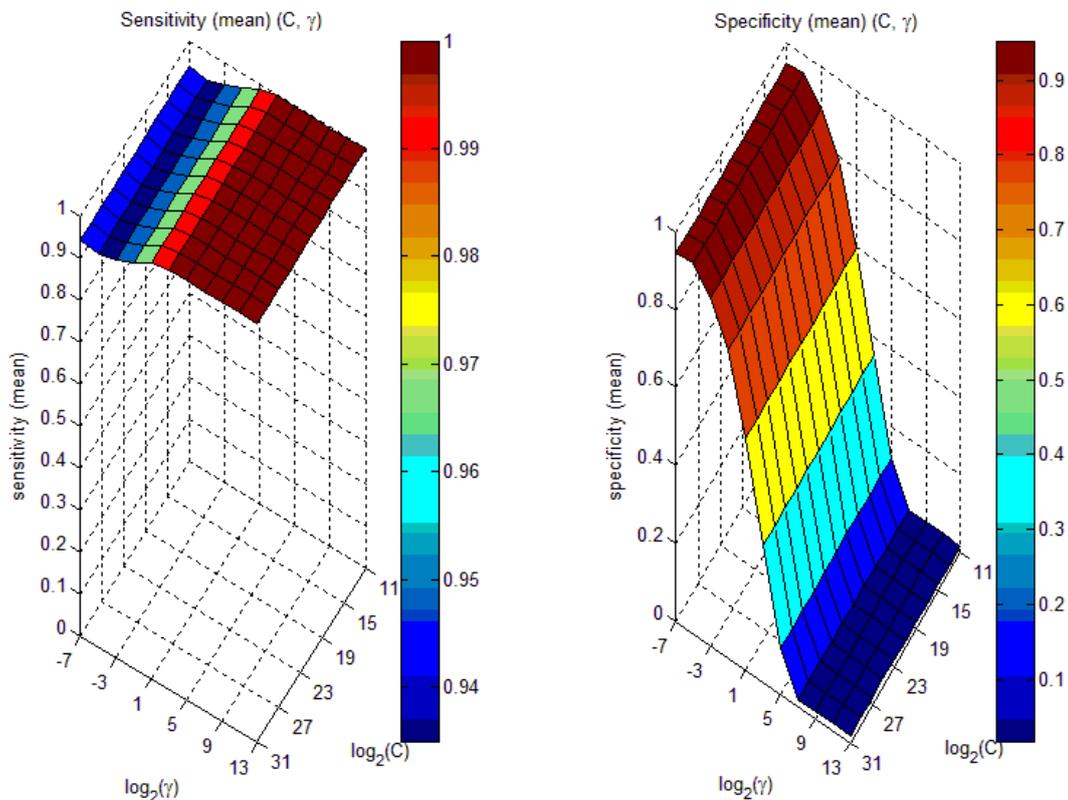


Figura 6.25: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 2).

En estas gráficas puede observarse que los valores medios de especificidad sufren una caída a partir de valores mayores a $\gamma = 2^{-5}$, y entorno a esta zona tenemos los valores más altos de especificidad. Esta situación sugiere continuar el estudio en esta zona haciendo un ajuste fino de los parámetros. El ajuste fino se realiza con el siguiente barrido.

Barrido 3 (ajuste fino):

- $\gamma = \{2^{-8}, 2^{-7,5}, 2^{-7}, 2^{-6,5}, 2^{-6}, 2^{-5,5}, 2^{-5}, 2^{-4,5}, 2^{-4}\};$
- $C = \{2^{14}, 2^{14,5}, 2^{15}, 2^{15,5}, 2^{16}, 2^{16,5}, 2^{17}\}.$

En la Figura 6.26 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 3.

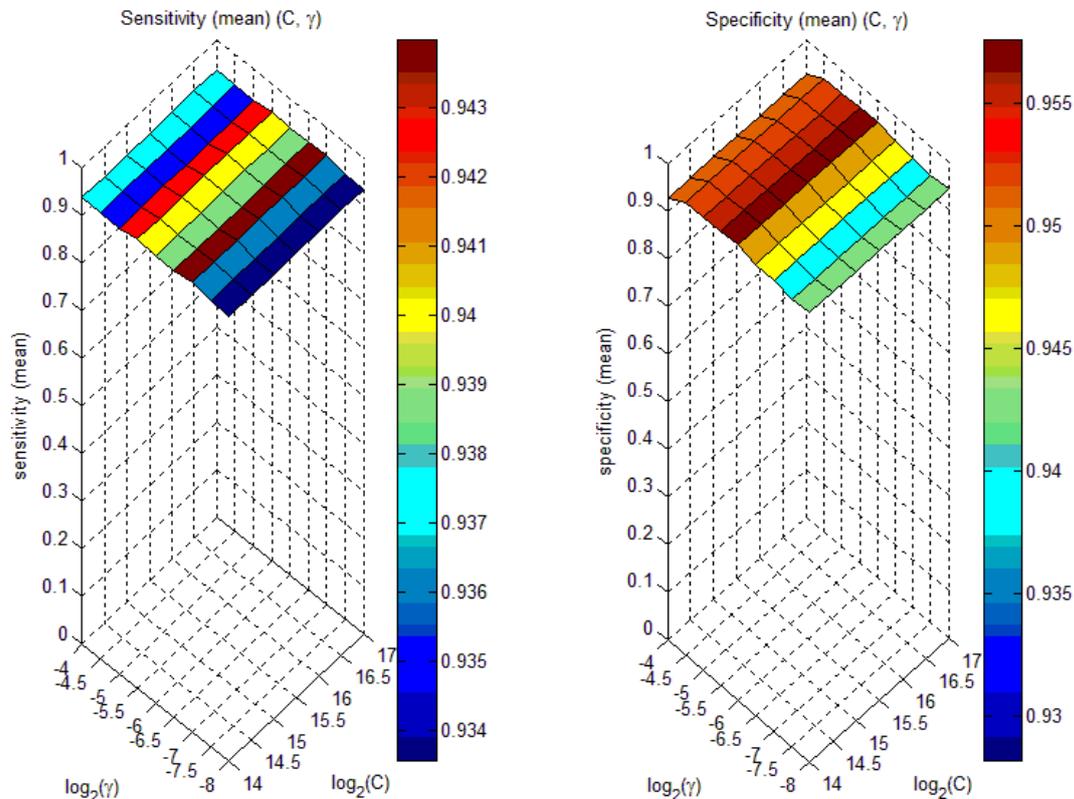


Figura 6.26: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 3 - ajuste fino).

Como puede observarse con este ajuste fino, a partir de valores de $\gamma = 2^{-4.5}$, el promedio de especificidad adopta una sutil tendencia a decrementar, por lo que tomamos esta región del plano para comprobar los pares de promedios de sensibilidad y especificidad en detalle. Nótese que los promedios de sensibilidad y especificidad en esta región del plano están por encima del 93%.

Al compararse los pares de estas métricas en esta región, del mismo modo que se hizo con la campaña 1, se escogen tres pares que presentan los mejores resultados.

Tabla 6.24: Pares de medias de sensibilidad y especificidad con la clase "grado IV" bajo test, que mejores resultados presentan, y los parámetros que los proporcionan.

	Sensibilidad (Media)	Especificidad (Media)	Parámetro γ	Parámetro C
PAR 1	94,4%	94,59%	$\{2^{-7}\}$	$\{2^{14}, 2^{14,5}, 2^{15}, 2^{15,5}, 2^{16}, 2^{16,5}, 2^{17}\}$
PAR 2	94%	95,76%	$\{2^{-6}\}$	$\{2^{14}, 2^{14,5}, 2^{15}, 2^{15,5}, 2^{16}, 2^{16,5}, 2^{17}\}$
PAR 3	94,27%	95,49%	$\{2^{-5,5}\}$	$\{2^{14}, 2^{14,5}, 2^{15}, 2^{15,5}, 2^{16}, 2^{16,5}, 2^{17}\}$

Tras aplicar la expresión (5.1) a los tres pares de promedios, en este caso se obtiene el factor más elevado con el “par 1”. A diferencia del estudio realizado con la campaña 1, no se realiza posteriormente el “ajuste fino”, por haberse realizado previamente.

Por otro lado, en el apartado 5.7 de esta memoria se comentó que el parámetro coste, C , involucra un compromiso entre la penalización de los errores y la maximización de la distancia al establecer el hiperplano de separación. Hay que tener en cuenta que el “par 1” ya considera una cantidad de errores (falsos positivos y falsos negativos) que están implícitos en los valores de estas métricas, y que estos valores se obtienen para los diferentes costes que se presentan en la tabla, por tanto, se escogerá el menor coste puesto que será el que maximiza la distancia del hiperplano de separación.

Por tanto, de este estudio se deduce que los parámetros óptimos del *kernel* RBF tomando la clase “grado IV” bajo test para la campaña 2 son los siguientes:

$$(\gamma, C) = (2^{-7}, 2^{14}) \quad (6.2)$$

Por último, queda comprobar si con estos parámetros óptimos se obtienen buenos resultados tomando como clase bajo test al resto de las clases de esta campaña.

Finalmente, en la Tabla 6.25 se exponen los resultados de las medias de OA, sensibilidad y especificidad, al tomar como clase bajo test a cada una las clases de la campaña 2, tras establecer los parámetros óptimos en el *kernel* RBF de este sistema de clasificación multiclase.

Tabla 6.25: Promedio y desviación típica de las métricas: overall accuracy, sensibilidad y especificidad (por clase), con tipo de kernel RBF y parámetros óptimos (campaña 2 - transmitancia).

Sistema de clasificación multiclase (campaña 2 - transmitancia)					
Media % (Desviación típica %)					
Clase bajo test	Sensibilidad		Especificidad		
Grado I	100 (0)			98,86 (2,1)	
Grado II	88 (11,69)			98,89 (2,35)	
Grado III	81 (34,83)			99,02 (2,38)	
Grado IV	94,4 (7,34)			94,59 (4,06)	
Tumor secundario	81,83 (20,18)			98,49 (2,05)	
No tumoral	89,47 (13,99)			98,65 (2,5)	
Overall Accuracy (OA)	91,42 (5,07)	Tipo de kernel:	RBF	(γ, C)	$(2^{-7}, 2^{14})$

Como puede observarse en esta tabla, empleando los parámetros óptimos en el *kernel* RBF, no solo se consiguen buenos resultados para la clase “grado IV”. En general se reflejan promedios de sensibilidad y especificidad muy altas, y el porcentaje de acierto global supera el 90% indicando un gran número de muestras clasificadas correctamente.

Destaca el promedio de sensibilidad de las muestras de grado I ya que indica que todas las muestras de esta clase se han clasificado correctamente, mientras que el alto promedio de su especificidad refleja que se ha conseguido un gran rechazo de las muestras que no pertenecen a esta clase.

No obstante, aunque en general los resultados ostentan unos porcentajes satisfactorios, hay que tomar en especial consideración que las métricas de la clase “no tumoral” no son del 100%, indicando que se han producido errores al clasificar algunas muestras no tumorales como tumorales y viceversa.

6.4.3. Resultados en sistemas de clasificación multiclase (campaña 2 - absorbancia)

Una vez realizado el estudio de los parámetros óptimos para el *kernel* RBF del sistema de clasificación multiclase de la campaña 2, se procedió a diseñar un sistema multiclase para clasificar a las muestras de la campaña 2 en absorbancia.

En la Tabla 6.26 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase de la campaña 2, empleando el *kernel* de tipo lineal, para las muestras definidas en absorbancia.

Tabla 6.26: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* lineal (campaña 2 - absorbancia).

Sistema de clasificación multiclase (campaña 2 - absorbancia)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado I	94,2 (13,07)		95,02 (2,55)
Grado II	78,11 (14,05)		95,54 (4,54)
Grado III	32 (41,35)		99,19 (2,27)
Grado IV	87,03 (7,49)		93,66 (4,94)
Tumor secundario	87,33 (20,91)		97,92 (2,33)
No tumoral	79,13 (18,94)		96,62 (3,31)
Overall Accuracy (OA)	82,68 (5,1)	Tipo de kernel:	Lineal

El valor promedio de OA indica que en general se consigue un porcentaje de acierto en las clasificaciones optimista.

No obstante, en este sistema también se clasifican algunas muestras no tumorales como tumorales (falsos negativos) y viceversa (falsos positivos).

De las muestras de grado II y IV se hace una clasificación aceptable ya que de sus promedios de sensibilidad se desprende que las proporciones de falsos negativos no son muy considerables.

Por otro lado, del promedio de sensibilidad de las muestras de grado I se deduce que gran parte de las muestras de grado I se clasifican correctamente, y de la media de especificidad, que una baja proporción de muestras que no pertenecen a esta clase, se clasifica como grado I.

Sin embargo, para la clase “grado III” se produce una proporción alta de falsos negativos ya que el promedio de sensibilidad es del 32%.

Por otra parte, si observamos el promedio de las métricas de tumor secundario, podemos deducir que la proporción de muestras con metástasis que no se detectan es baja, aunque su desviación típica indica que fluctúa bastante en las 50 repeticiones. Del promedio de especificidad se desprende que la proporción de muestras que no son metástasis y se confunden con metástasis, es muy baja.

Análogamente a los experimentos realizados con esta campaña en transmitancia, se realizaron experimentos en absorbancia, empleando el *kernel* del tipo RBF utilizando sus parámetros por defecto.

En la Tabla 6.27 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase, empleando el *kernel* de tipo RBF, para las muestras definidas en absorbancia.

Tabla 6.27: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* RBF (campaña 2 - absorbancia).

Sistema de clasificación multiclase (campaña 2 - absorbancia)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado I	0 (0)		100 (0)
Grado II	0 (0)		100 (0)
Grado III	0 (0)		100 (0)
Grado IV	100 (0)		0 (0)
Tumor secundario	0 (0)		100 (0)
No tumoral	0 (0)		100 (0)
Overall Accuracy (OA)	40,42 (1,28)	Tipo de kernel:	RBF (parámetros por defecto)

Como se refleja en los resultados de la tabla, el sistema resultante de emplear este tipo de *kernel* sin hacer uso los parámetros es un sistema ineficiente.

Este sistema realiza exactamente la misma clasificación que su sistema homólogo en transmitancia. El hecho de que incluso refleje el mismo promedio y desviación típica en OA se debe a que se ha tomado el mismo *training* y *test* para los dos sistemas.

Tras este experimento, se realizó un experimento que empleaba el *kernel* RBF introduciendo los parámetros óptimos obtenidos en el estudio en transmitancia.

En la Tabla 6.28 se presentan los resultados de las medias de OA, sensibilidad y especificidad, al tomar como clase bajo test a cada una las clases de la campaña 2 en absorbancia, tomando los parámetros óptimos en el *kernel* RBF obtenidos en el estudio en transmitancia.

Tabla 6.28: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de kernel RBF y parámetros óptimos obtenidos para esta campaña en transmitancia (campaña 2 - absorbancia).

Sistema de clasificación multiclase (campaña 2 - absorbancia)					
Media % (Desviación típica %)					
Clase bajo test	Sensibilidad		Especificidad		
Grado I	91,7 (15,93)			94,42 (2,91)	
Grado II	78,11 (14,78)			95,61 (3,99)	
Grado III	9 (21,88)			98,15 (2,32)	
Grado IV	91,41 (9,96)			95,25 (5,15)	
Tumor secundario	85,67 (20,96)			96,58 (2,88)	
No tumoral	81,07 (15,83)			96,63 (3,34)	
Overall Accuracy (OA)	83,58 (5,28)	Tipo de kernel:	RBF	(γ, C)	$(2^{-7}, 2^{14})$

Puede observarse que para las muestras de la campaña 2 en absorbancia, empleando los parámetros óptimos obtenidos con las muestras en transmitancia, se consiguen buenos resultados excepto para las muestras de grado III. Al conseguirse un promedio de 9% de la sensibilidad para esta clase, se indica que hay una gran proporción de falsos negativos, es decir, de muestras de grado III que no se clasifican como tales.

6.4.4. Resultados en sistemas de clasificación multiclase (fusión de las campañas)

Posteriormente, se realiza una fusión de ambas campañas tal como se explica en el apartado 5.2.2, disponiéndose la fusión de las campañas en las dos respuestas: absorbancia y transmitancia (apartado 6.2). Una vez realizada la fusión, se procede a desarrollar los

sistemas clasificadores multiclase para clasificar la fusión en ambas magnitudes. En todos estos experimentos se emplean los mismos *training* y *test datasets*.

En la Tabla 6.29 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase de la fusión de las campañas, empleando el *kernel* de tipo lineal, para las muestras definidas en absorbancia.

Tabla 6.29: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* lineal (campaña fusión - absorbancia).

Sistema de clasificación multiclase (campaña fusión - absorbancia)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado I	88,47 (16,96)		97,68 (1,17)
Grado II	44,44 (13,86)		97,55 (1,95)
Grado III	6,58 (7,73)		99,6 (0,91)
Grado IV	95,08 (2,68)		62,05 (5,64)
Tumor secundario	86,67 (18,52)		99,38 (1,32)
No tumoral	84,36 (10,76)		98,96 (1,32)
Overall Accuracy (OA)	73,38 (2,85)	Tipo de kernel:	Lineal

En la Tabla 6.30 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase de la fusión de las campañas, empleando el *kernel* de tipo lineal, para las muestras definidas en transmitancia.

Tabla 6.30: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* lineal (campaña fusión - transmitancia).

Sistema de clasificación multiclase (campaña fusión - transmitancia)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado I	94,87 (8,76)		98,87 (1,21)
Grado II	42,48 (14,22)		97,25 (1,78)
Grado III	6,51 (6,25)		99,98 (0,17)
Grado IV	96,81 (2,9)		56,9 (6,25)
Tumor secundario	84,33 (19,46)		99,8 (0,62)
No tumoral	77,4 (11,73)		100 (0)
Overall Accuracy (OA)	77,7 (3,58)	Tipo de kernel:	Lineal

Como puede observarse tanto en la Tabla 6.29 como en la Tabla 6.30, al clasificar la fusión de las campañas en absorbancia y transmitancia, utilizando el *kernel* lineal, se obtienen buenos resultados excepto en la clasificación de las muestras de grado II y III. Al clasificar estas clases en ambas magnitudes se obtiene unos promedios de sensibilidad bajos, siendo muy bajos en el caso de “grado III”. De estos valores se deduce que una gran proporción de estas muestras de estas clases, pero especialmente de “grado III”, están siendo clasificadas como otra clase.

Por otro lado, de los promedios de sensibilidad y especificidad de la clase “no tumoral” se deduce que una proporción baja de estas muestras se están clasificando como tumorales, mientras que muy pocas muestras tumorales se están clasificando como no tumorales, llegando incluso a no clasificarse ninguna para el caso de las muestras en transmitancia ya que la especificidad de la clase no tumoral es en este caso 100% en las 50 repeticiones.

De la misma manera que se hizo con ambas campañas por separado, también se desarrolló un sistema para clasificar la fusión de las campañas empleando el *kernel* RBF con los parámetros por defecto. Estos experimentos se realizaron tanto en transmitancia como en absorbancia, sin embargo, los resultados fueron los mismos.

En la Tabla 6.31 se presentan los promedios y las desviaciones típicas de OA, sensibilidades y especificidades respectivas de cada clase, empleando el *kernel* de tipo RBF, para las muestras definidas tanto en absorbancia como transmitancia.

Tabla 6.31: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* RBF (campaña fusión - absorbancia y transmitancia).

Sistema de clasificación multiclase (campaña fusión – absorbancia y transmitancia)			
Media % (Desviación típica %)	Sensibilidad		Especificidad
Clase bajo test			
Grado I	0 (0)		100 (0)
Grado II	0 (0)		100 (0)
Grado III	0 (0)		100 (0)
Grado IV	100 (0)		0 (0)
Tumor secundario	0 (0)		100 (0)
No tumoral	0 (0)		100 (0)
Overall Accuracy (OA)	56,7 (0,49)	Tipo de kernel:	RBF (parámetros por defecto)

Como se refleja en la tabla, una vez más se vuelve a obtener el resultado de un sistema ineficiente al clasificarse todas las muestras como “grado IV”. Este resultado no es muy sorprendente ya que al realizarse este experimento con ambas campañas por separado y obtenerse este resultado, era de esperar que con la fusión ocurriera lo mismo. Solamente quedaba comprobarlo en transmitancia ya que este experimento no se hizo con la campaña 1 en su respuesta en transmitancia.

Luego se procedió a realizar las clasificaciones de la campaña fusión en absorbancia empleando el *kernel* RBF introduciendo los parámetros óptimos obtenidos para la campaña 1 y la campaña 2.

En la Tabla 6.32 se presentan los resultados de las medias de OA, sensibilidad y especificidad, al tomar como clase bajo test a cada una las clases de la fusión de las campañas en absorbancia, tomando los parámetros óptimos en el *kernel* RBF obtenidos en el estudio en absorbancia de la campaña 1.

Tabla 6.32: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* RBF y parámetros óptimos obtenidos para la campaña 1 en absorbancia (campaña fusión - absorbancia).

Sistema de clasificación multiclase (campaña fusión - absorbancia)					
Media % (Desviación típica %)					
Clase bajo test	Sensibilidad		Especificidad		
Grado I	86,33 (16,22)			98,89 (1,01)	
Grado II	87,87 (9,74)			96,85 (2,13)	
Grado III	56,2 (17,07)			97,43 (1,84)	
Grado IV	93,36 (4,16)			88,89 (5,63)	
Tumor secundario	89,83 (16,0)			99,09 (1,5)	
No tumoral	85,16 (13,35)			99,27 (0,86)	
Overall Accuracy (OA)	87,6 (3,1)	Tipo de kernel:	RBF	(γ , C)	(2^{-17} , 2^{21})

Como puede verse en esta tabla, al emplear los parámetros óptimos obtenidos en el estudio realizado en la campaña 1 para clasificar la fusión de las campañas en absorbancia, se obtienen en general unos promedios de sensibilidad y especificidad relativamente altos. Cabe mencionar que el peor de los resultados se obtiene en el promedio de la sensibilidad tomando las muestras de grado III bajo test. Este promedio indica que un poco menos de la mitad de las muestras de grado III se clasifican como si fueran de otras clases.

También hay que resaltar que se producen algunos errores al clasificar las muestras no tumorales puesto que algunas de estas se clasifican como tumorales, mientras que son muy pocas las que pertenecen a algún tipo de tumor y se clasifican como no tumorales.

En la Tabla 6.33 se presentan los resultados de las medias de OA, sensibilidad y especificidad, al tomar como clase bajo test a cada una las clases de la fusión de las campañas en absorbancia, tomando los parámetros óptimos en el *kernel* RBF obtenidos en el estudio en transmitancia de la campaña 2.

Tabla 6.33: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* RBF y parámetros óptimos obtenidos para la campaña 2 en transmitancia (campaña fusión - absorbancia).

Sistema de clasificación multiclase (campaña fusión - absorbancia)					
Media % (Desviación típica %)					
Clase bajo test	Sensibilidad		Especificidad		
Grado I	84,47 (17,57)			98,89 (1,11)	
Grado II	85,04 (10,71)			96,45 (2,29)	
Grado III	73,91 (14,27)			97,81 (1,8)	
Grado IV	92,72 (3,55)			93,27 (4,53)	
Tumor secundario	86,67 (20,96)			98,76 (1,67)	
No tumoral	89,44 (8,02)			98,82 (1,27)	
Overall Accuracy (OA)	88,77 (2,35)	Tipo de kernel:	RBF	(γ, C)	$(2^{-7}, 2^{14})$

Al emplear los parámetros óptimos obtenidos en el estudio realizado en la campaña 2 para clasificar la fusión de las campañas en absorbancia, también se obtienen resultados favorables destacándose que con estos parámetros incluso se mejoran los resultados de clasificación de las muestras de grado III que se exponen en la Tabla 6.32.

Posteriormente, se realizan los mismos experimentos empleando los parámetros óptimos de la campaña 1 y los parámetros óptimos de la campaña 2, para clasificar la fusión de las campañas en transmitancia.

En la Tabla 6.34 se presentan los resultados de las medias de OA, sensibilidad y especificidad, al tomar como clase bajo test a cada una las clases de la fusión de las campañas en transmitancia, tomando los parámetros óptimos en el *kernel* RBF obtenidos en el estudio en absorbancia de la campaña 1.

Tabla 6.34: Promedio y desviación típica de las métricas: overall accuracy, sensibilidad y especificidad (por clase), con tipo de kernel RBF y parámetros óptimos obtenidos para la campaña 1 en absorbancia (campaña fusión - transmitancia).

Sistema de clasificación multiclase (campaña fusión - transmitancia)					
Media % (Desviación típica %)	Sensibilidad			Especificidad	
Clase bajo test					
Grado I	95,27 (8,52)			99,28 (1,02)	
Grado II	89,79 (9,32)			98,49 (1,57)	
Grado III	51 (17,18)			98,54 (1,76)	
Grado IV	94,98 (3,33)			85,69 (4,28)	
Tumor secundario	85,33 (17,54)			99,03 (1,39)	
No tumoral	93,58 (7,39)			99,84 (0,53)	
Overall Accuracy (OA)	89,43 (2,44)	Tipo de kernel:	RBF	(γ, C)	$(2^{-17}, 2^{21})$

Una vez obtenidos los resultados de la fusión de la campaña en transmitancia, puede observarse que los resultados mejoran un poco respecto a los resultados obtenidos en absorbancia, excepto para las muestras de tumor secundario y grado III. De esta última clase, siguen clasificándose mal aproximadamente la mitad de las muestras.

6.4.4.1. Resultados de la optimización de parámetros del *kernel* RBF (fusión de las campañas)

Para tratar de mejorar los resultados de la fusión de las campañas, se decidió realizar una búsqueda de los parámetros óptimos del *kernel* RBF para la campaña fusión.

Como se explicó en el apartado 5.7.3, se hizo el primer barrido partiendo de los parámetros óptimos obtenidos con cada campaña por separado.

En el primer barrido de parámetros para la fusión de las campañas, se toma los parámetros óptimos de las campañas 1 y 2 como puntos límites del barrido.

Barrido 1:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-17 \leq n \leq -7$ y $14 \leq m \leq 21$ con n impar y m natural.

En la Figura 6.27 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 1.

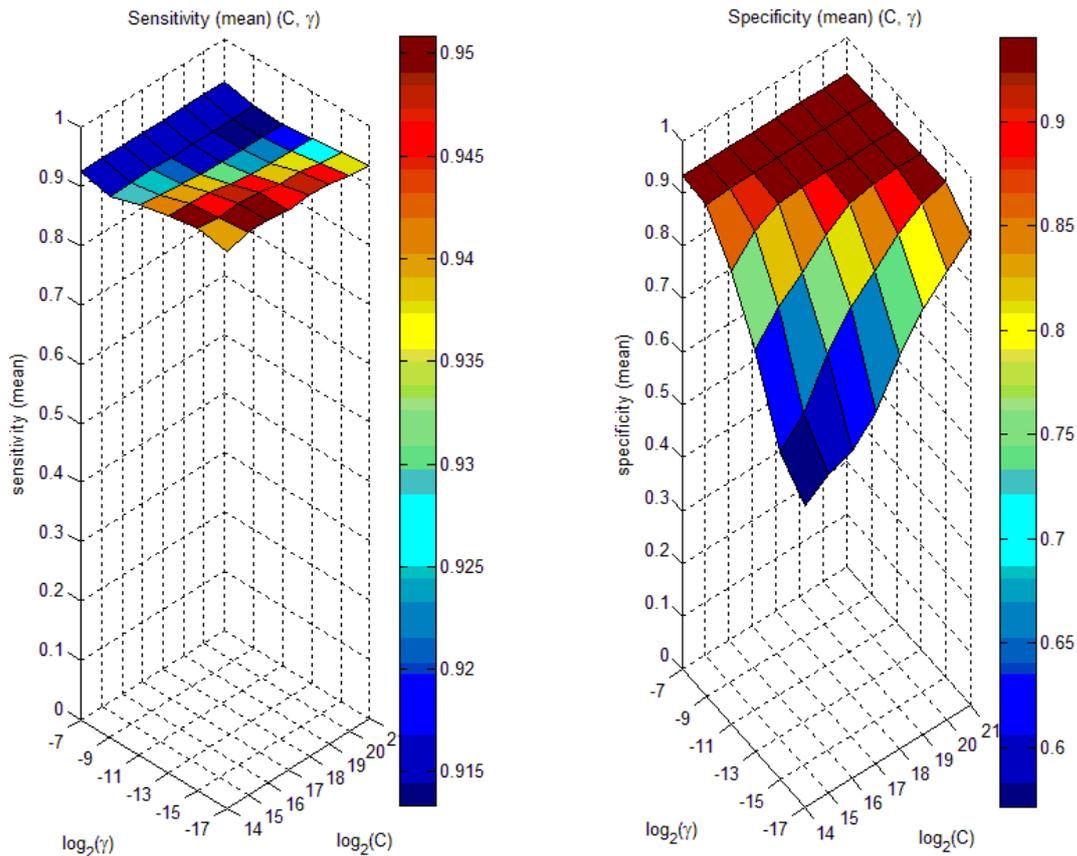


Figura 6.27: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test barrido 1.

Como puede observarse en estas gráficas, los valores de promedios altos se encuentran para valores de $\gamma > 2^{-9}$. Por tanto se realiza un segundo barrido con la intensidad de ver si se pueden obtener mejores resultados para valores crecientes de γ .

Barrido 2:

- $\gamma = \{2^n\}$ y $C = \{2^m\}$ donde $-9 \leq n \leq -1$ y $14 \leq m \leq 22$ con n impar y m par.

En la Figura 6.28 se presentan las superficies promedio de sensibilidad y especificidad para el barrido 2.

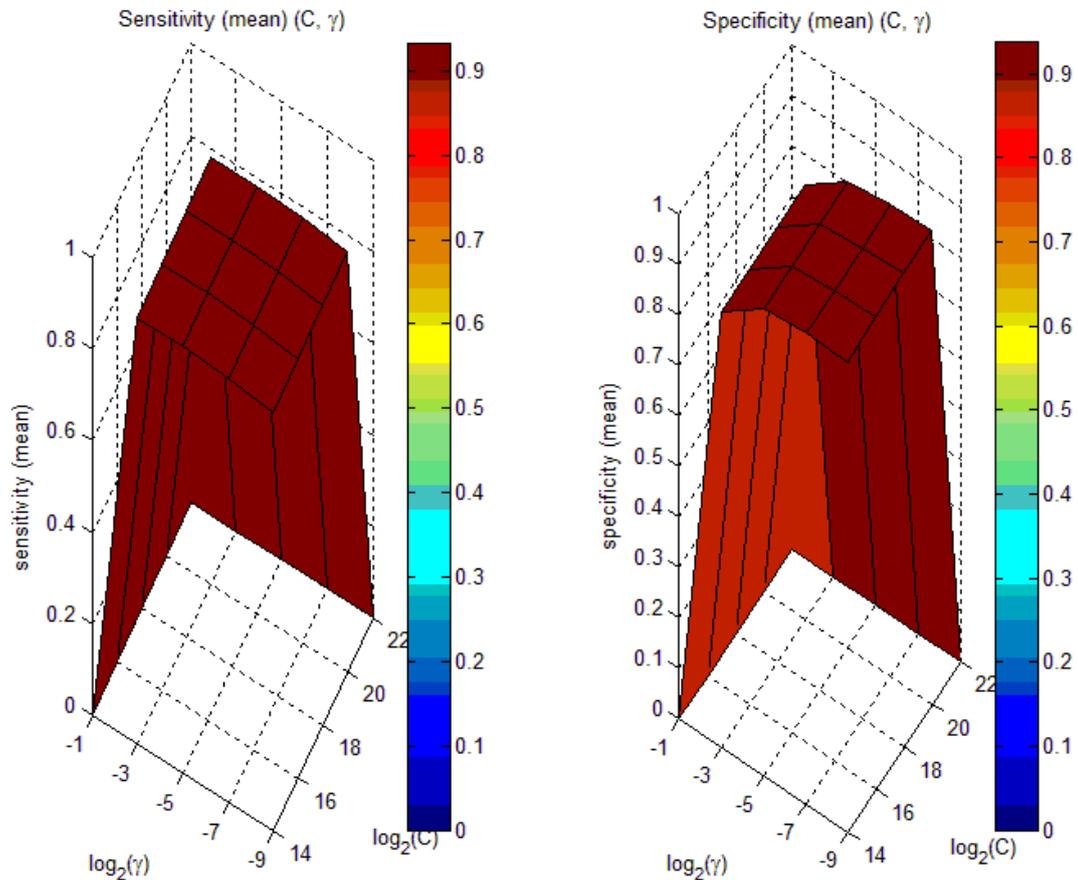


Figura 6.28: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test barrido 2.

Tras realizar este barrido, se comprobó que los valores promedio de sensibilidad y especificidad eran menores que los obtenidos en determinadas zonas del plano cubiertas por el barrido 1. Por tanto, se realiza un ajuste fino en esa zona del plano.

Barrido 3 (ajuste fino):

- $\gamma = \{2^{-9}, 2^{-8,5}, 2^{-8}, 2^{-7,5}, 2^{-7}\};$
- $C = \{2^{18}, 2^{18,5}, 2^{19}, 2^{19,5}, 2^{20}, 2^{20,5}, 2^{21}\}.$

En la Figura 6.29 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 3.

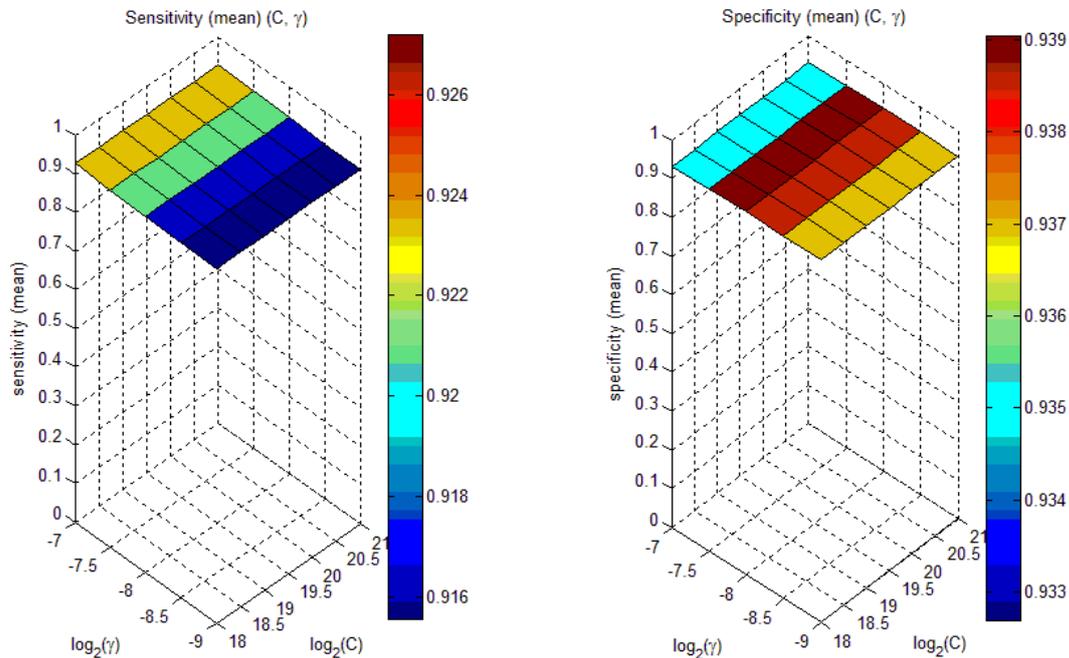


Figura 6.29: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 3 - ajuste fino).

Después de realizar este ajuste fino, se plantea la duda de saber si es posible mejorar estos resultados para valores inferiores de C ya que en la Figura 6.27 se observan zonas favorables para estos valores de coste, que no se han cubierto en el último barrido. Por tanto, se plantea el siguiente ajuste fino.

Barrido 4 (ajuste fino):

- $\gamma = \{2^{-9}, 2^{-8,5}, 2^{-8}, 2^{-7,5}, 2^{-7}\};$
- $C = \{2^{14}, 2^{14,5}, 2^{15}, 2^{15,5}, 2^{16}, 2^{16,5}, 2^{17}, 2^{17,5}\}.$

En la Figura 6.30 se ilustran las superficies promedio de sensibilidad y especificidad para el barrido 4.

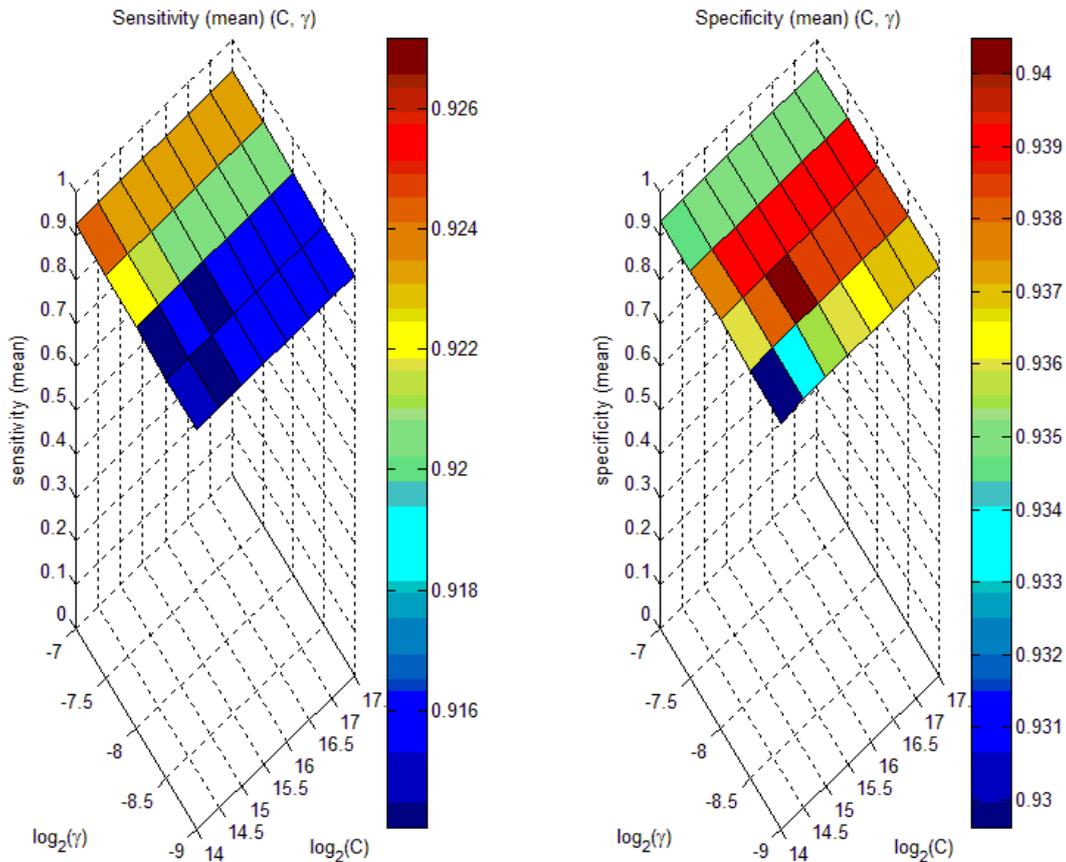


Figura 6.30: Superficies promedio de sensibilidad y especificidad con la clase "grado IV" bajo test (barrido 4 - ajuste fino).

Después de obtener los resultados en los dos ajustes finos se aplicó la expresión (5.1) a todos los pares de promedios de estos dos barridos. El par de promedios de sensibilidad y especificidad que proporcionaron el factor más elevado fueron los siguientes valores:

- Sensibilidad (media) = 92,72%.
- Especificidad (media) = 93,27%.

Estos resultados se obtienen empleando los siguientes parámetros:

- $\gamma = \{2^{-7}\}$.
- $C = \{2^{14}, 2^{14,5}, 2^{15}, 2^{15,5}, 2^{16}, 2^{16,5}, 2^{17}, 2^{17,5}, 2^{18}, 2^{18,5}, 2^{19}, 2^{19,5}, 2^{20}, 2^{20,5}, 2^{21}\}$.

Una vez más se escoge el de menor coste para que el margen del hiperplano de separación sea máximo. Por tanto, los parámetros óptimos de la fusión de las campañas en absorbancia tomando como clase bajo test las muestras de grado IV son los siguientes:

$$(\gamma, C) = (2^{-7}, 2^{14}) \quad (6.3)$$

Finalmente, tras la realización de este estudio se llega a la conclusión de que los parámetros óptimos para la fusión de las campañas en absorbancia son exactamente los mismos que se obtuvieron en el estudio de parámetros óptimos para la campaña 2 en transmitancia. Los resultados de la clasificación multiclase de la fusión de las campañas en absorbancia empleando el *kernel* RBF y estos parámetros óptimos, fueron mostrados en la Tabla 6.33.

Por último, se realizó una última clasificación empleando los parámetros óptimos obtenidos en el último estudio para clasificar la fusión de las campañas en transmitancia.

En la Tabla 6.35 se presentan los resultados de las medias de OA, sensibilidad y especificidad, al tomar como clase bajo test a cada una las clases de la fusión de las campañas en transmitancia, tomando los parámetros óptimos en el *kernel* RBF obtenidos en el estudio en absorbancia de la fusión de las campañas.

Como puede contrastarse en la Tabla 6.33 y la Tabla 6.35, al realizar la clasificación con los parámetros óptimos obtenidos en el último estudio, se logran una mejora de resultados en transmitancia. El promedio del porcentaje de acierto global indica que se ha llevado una clasificación con más aciertos en general.

Las muestras de grado I presentan una media de especificidad del que se desprende que una proporción de muestras muy baja de otra/s clase/s se clasifica incorrectamente como grado I, de igual modo, muy pocas de grado I se clasifican en otra clase.

Para las clases: “grado II” y “grado III” también se consiguen muy buenos resultados ya que los promedios de sensibilidad y especificidad de ambos son del orden del 90% (clasificación en transmitancia).

Para la clase “tumor secundario, el promedio de especificidad ronda el 100% produciéndose solamente algunos errores al confundirse muy pocas muestras que no son de esta clase. Sin embargo, el promedio de la sensibilidad de esta clase baja un poco con

respecto al resultado obtenido en absorbancia; de un 86,67% a un 82,33% con la desviación típica que más fluctúa de las clases.

Por otro lado, para la “clase III”, se ha conseguido un aumento en el promedio de sensibilidad, llegándose hasta el 80,27%. Téngase en cuenta que esta clase presentaba promedios de sensibilidad muy bajos en diferentes sistemas.

Por último, en esta clasificación de la fusión de las campañas en transmitancia, se ha logrado obtener el valor medio más alto de sensibilidad para la clase “no tumor”. Además, el valor medio de su especificidad casi llega al 100% obteniéndose una proporción muy baja de muestras tumorales que se clasifican como no tumorales.

Tabla 6.35: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad (por clase), con tipo de *kernel* RBF y parámetros óptimos obtenidos para la fusión de las campañas en absorbancia (campaña fusión - transmitancia).

Sistema de clasificación multiclase (campaña fusión - transmitancia)					
Media % (Desviación típica %)	Sensibilidad		Especificidad		
Clase bajo test					
Grado I	95,67 (8,25)			99,25 (1,11)	
Grado II	90,24 (9,16)			98,14 (1,78)	
Grado III	80,27 (13,54)			98,46 (1,22)	
Grado IV	94,86 (3,0)			93,11 (3,91)	
Tumor secundario	82,33 (20,94)			99,27 (0,96)	
No tumoral	93,36 (6,78)			99,48 (0,86)	
Overall Accuracy (OA)	92,17 (2,7)	Tipo de kernel:	RBF	(γ , C)	(2^{-7} , 2^{14})

6.4.5. Resumen de resultados más relevantes con sistemas multiclase

En todas las clasificaciones realizadas en este proyecto mediante el uso de sistemas multiclase, se estableció un mismo *training* para el tipo de sistema que clasifica el mismo conjunto de datos. Es decir: se estableció un mismo *training* de las muestras de la campaña 1, para los sistemas encargados de clasificar esta campaña; un mismo *training* de la campaña 2, para los sistemas encargados de clasificar la campaña 2 en absorbancia y transmitancia (cada *training* consta de las mismas muestras pero en su correspondiente

magnitud); y el mismo *training* de muestras de la fusión de las campañas, para los sistemas encargados de clasificar las muestras de las dos campañas, en absorbancia y transmitancia.

Hay que mencionar que, de todas las clasificaciones multiclase que se han realizado en este proyecto empleando el *kernel* RBF con sus parámetros por defecto, solamente presentan cierto interés las realizadas con la campaña 1. En la Tabla 6.18 se reflejaba que con las muestras de esta campaña se conseguía discriminar perfectamente las muestras sanas (y discernir entre el tipo de sustancia), pero que las muestras tumorales se clasificaban todas como muestras de grado IV. Sin embargo, para el caso de la campaña 2 y para la campaña fusión, tanto en absorbancia como en transmitancia, se obtenía sistemas ineficientes puesto que todas las muestras se clasificaban como muestras de grado IV.

Por otra parte, en la Tabla 6.36 se presentan las clasificaciones realizadas con sistemas que emplean el *kernel* lineal. De los resultados obtenidos por estos sistemas, destaca que se obtienen medias de sensibilidad muy bajas para la clase “grado III”, especialmente en los sistemas que clasifican la fusión de las campañas. Solamente se obtiene un promedio alto de esta métrica con el sistema que clasifica la campaña 2 en transmitancia, y en este caso, su desviación típica indica que dicho promedio fluctúa considerablemente en las 50 repeticiones.

Además, destaca el resultado del promedio de especificidad de la clase “no tumor” para el sistema que clasifica la fusión de las campañas en transmitancia. El 100% de especificidad indica que ninguna muestra tumoral es clasificada por ese sistema como tumoral. También hay que señalar que el sistema con este tipo de *kernel* que clasifica la campaña 1, discrimina perfectamente las muestras sanas de las tumorales, y entre las clases sanas tampoco se cometen errores, no obstante, debe recordarse que solo se cuenta con 15 muestras sanas en esta campaña, con lo que, si tomamos el 75% como *training*, lo cierto es que quedan pocas para *test*.

Por otro lado, en la Tabla 6.37 se presentan los resultados obtenidos por los sistemas que emplean el *kernel* de tipo RBF, y hacen uso de los parámetros óptimos proporcionados en los diferentes estudios. Los parámetros óptimos de la campaña 1 se obtuvieron en absorbancia, mientras que los de la campaña 2 y la campaña fusión, que resultaron los mismos, se alcanzaron en sus respectivos estudios en transmitancia.

Cotejando los porcentajes de acierto global obtenidos con los parámetros óptimos (RBF), y los obtenidos con el *kernel* de tipo lineal, se puede observar una mejoría respecto al lineal. Esta mejoría se refleja incluso para la clasificación de la campaña 2 en absorbancia, aunque el OA es ligeramente superior (alrededor de un 1%). Adviértase que para esta clasificación (campaña 2 en absorbancia con parámetros óptimos, RBF) se obtiene una clasificación de las muestras de grado III muy deficiente, puesto que el promedio de sensibilidad indica que una proporción muy baja de estas muestras se clasifican correctamente. También con este *kernel* se consigue discriminar correctamente las muestras sanas de las tumorales en la campaña 1.

Por último, debe recalcarse que de los diferentes sistemas multiclase desarrollados para clasificar la fusión de las campañas, el que mejores resultados ha proporcionado es el sistema que emplea el *kernel* RBF con los parámetros óptimos: $(\gamma, C) = (2^{-7}, 2^{14})$, y clasifica las muestras en su respuesta en transmitancia.

En este proyecto no se ha conseguido discriminar correctamente el 100% de las muestras tumorales de las muestras no tumorales de la campaña 2, sin embargo, se considera que se han conseguido porcentajes de acierto muy elevados para todas las clases.

Tabla 6.36: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad; en las clasificaciones realizadas por los sistemas multiclase con tipo de *kernel* lineal.

Sistemas de clasificación multiclase - <i>kernel</i>: Lineal											
Media % (D. Típica %)	Campaña 1 (absorbancia)		Campaña 2 (transmitancia)		Campaña 2 (absorbancia)		Fusión (absorbancia)		Fusión (transmitancia)		
Clase bajo test	Sens.	Espec.	Sens.	Espec.	Sens.	Espec.	Sens.	Espec.	Sens.	Espec.	
Grado I	-	-	100 (0)	98,08 (2,69)	94,2 (13,07)	95,02 (2,55)	88,47 (16,96)	97,68 (1,17)	94,87 (8,76)	98,87 (1,21)	
Grado II	85,24 (17,04)	98,02 (2,14)	79,43 (14,7)	93,37 (5,16)	78,11 (14,05)	95,54 (4,54)	44,44 (13,86)	97,55 (1,95)	42,48 (14,22)	97,25 (1,78)	
Grado III	33,68 (12,18)	99,5 (1,66)	93 (20,23)	98,09 (2,35)	32 (41,35)	99,19 (2,27)	6,58 (7,73)	99,6 (0,91)	6,51 (6,25)	99,98 (0,17)	
Grado IV	97,62 (2,81)	68,72 (9,14)	76,31 (12,9)	85,36 (8,86)	87,03 (7,49)	93,66 (4,94)	95,08 (2,68)	62,05 (5,64)	96,81 (2,9)	56,9 (6,25)	
Tumor Secundario	-	-	83,83 (18,85)	99,42 (1,31)	87,33 (20,91)	97,92 (2,33)	86,67 (18,52)	99,38 (1,32)	84,33 (19,46)	99,8 (0,62)	
No tumor	Sustancia gris	100 (0)	100 (0)	62,4 (22,5)	96,64 (4,26)	79,13 (18,94)	96,62 (3,31)	84,36 (10,76)	98,96 (1,32)	77,4 (11,73)	100 (0)
	Sustancia blanca	100 (0)	100 (0)								
Overall Accuracy		87,64 (3,22)		79,05 (6,22)		82,68 (5,1)		73,38 (2,85)		77,7 (3,58)	

Tabla 6.37: Promedio y desviación típica de las métricas: *overall accuracy*, sensibilidad y especificidad; en las clasificaciones realizadas por los sistemas multiclase con tipo de *kernel* RBF y los parámetros óptimos obtenidos en sus respectivos estudios.

Sistemas de clasificación multiclase - kernel: RBF (Parámetros Óptimos)											
Media % (D. Típica %)		Campaña 1 (absorbancia)		Campaña 2 (transmitancia)		Campaña 2 (absorbancia)		Fusión (absorbancia)		Fusión (transmitancia)	
Clase bajo test		Sens.	Espec.	Sens.	Espec.	Sens.	Espec.	Sens.	Espec.	Sens.	Espec.
Grado I		-	-	100 (0)	98,86 (2,1)	91,7 (15,93)	94,42 (2,91)	84,47 (17,57)	98,89 (1,11)	95,67 (8,25)	99,25 (1,11)
Grado II		91,48 (12,86)	98,38 (1,87)	88 (11,69)	98,89 (2,35)	78,11 (14,78)	95,61 (3,99)	85,04 (10,71)	96,45 (2,29)	90,24 (9,16)	98,14 (1,78)
Grado III		83,61 (12,4)	97,68 (2,32)	81 (34,83)	99,02 (2,38)	9 (21,88)	98,15 (2,32)	73,91 (14,27)	97,81 (1,8)	80,27 (13,54)	98,46 (1,22)
Grado IV		96,51 (3,1)	93,33 (6,53)	94,4 (7,34)	94,59 (4,06)	91,41 (9,96)	95,25 (5,15)	92,72 (3,55)	93,27 (4,53)	94,86 (3,0)	93,11 (3,91)
Tumor Secundario		-	-	81,83 (20,18)	98,49 (2,05)	85,67 (20,96)	96,58 (2,88)	86,67 (20,96)	98,76 (1,67)	82,33 (20,94)	99,27 (0,96)
No tumor	Sustancia gris	100 (0)	100 (0)	89,47 (13,99)	98,65 (2,5)	81,07 (15,83)	96,63 (3,34)	89,44 (8,02)	98,82 (1,27)	93,36 (6,78)	99,48 (0,86)
	Sustancia blanca	100 (0)	100 (0)								
Overall Accuracy		94,39 (2,91)		91,42 (5,07)		83,58 (5,28)		88,77 (2,35)		92,17 (2,7)	
P. Óptimos (γ, C)		$(2^{-17}, 2^{21})$		$(2^{-7}, 2^{14})$		$(2^{-7}, 2^{14})$		$(2^{-7}, 2^{14})$		$(2^{-7}, 2^{14})$	

6.5 Resumen

En este capítulo se reflejan los resultados de los experimentos que se han ido realizando en el transcurso de este Proyecto de Fin de Carrera. En una primera parte se muestran, en su respuesta en absorbancia y transmitancia, las firmas espectrales de todas las muestras de las dos campañas (con las muestras de la campaña 1 adaptadas) agrupadas según sus clases. Puede comprobarse en estas gráficas que las firmas de la campaña 2 presentan mayor diferencia en su forma, mientras que las de la campaña 1 se ajustan más a un patrón común según sus clases. Es precisamente este hecho el que motiva el desarrollo de diferentes tipos de sistemas con el objetivo de lograr buenos resultados al clasificar, no solo las muestras de la campaña 2, sino además las de la campaña 1.

Por esta razón, se desarrolla un sistema de clasificación basado en el método de árboles jerárquicos para clasificar las muestras de la campaña 1, y posteriormente se desarrolla otro sistema con la idea de clasificar las muestras de la campaña 2. Ambos sistemas emplean modelos SVM con el tipo de *kernel* lineal, para realizar clasificaciones biclase en los nodos de los diferentes niveles de sus respectivos árboles.

Mientras que en el sistema que clasifica la campaña 1 se consiguen buenos resultados, puesto que se consigue una discriminación perfecta entre muestras sanas y muestras tumorales, en el que clasifica la campaña 2 se obtienen peores resultados.

Por tanto, posteriormente se trata de mejorar los resultados empleando los sistemas de clasificación multiclase. A través de este método se desarrollan diversos sistemas para clasificar cada una de las campañas por separado y la fusión de las campañas. Se comienza realizando experimentos con sistemas que emplean *kernels* de tipo lineal, hasta llegar a diseños más complejos que emplean *kernels* de tipo RBF utilizando distintos parámetros. Para lograr obtener los parámetros óptimos se hacen diferentes estudios en el plano de parámetros efectuando diversos barridos hasta llegar a obtener los valores de los parámetros que mejores resultados de clasificación ofrecen.

Por último, se realizan clasificaciones con estos parámetros óptimos con las diferentes campañas en sus dos magnitudes, hasta llegar a obtener resultados en las métricas muy favorables.

Capítulo 7

Conclusiones y líneas futuras

7.1 Introducción

Para terminar, en este último capítulo de la memoria se examinan los objetivos alcanzados con la finalización de este proyecto y se contrasta con los objetivos que se habían fijado en un principio.

Por otra parte, se exponen las conclusiones generales a las que se ha llegado con la implementación de los sistemas de clasificación jerárquica basada en árboles de decisión, de los sistemas de clasificación multiclase y con la realización de los diversos experimentos.

Por otro lado, desde la experiencia que se ha adquirido con el desarrollo de este proyecto, se proponen posibles líneas futuras de trabajo para orientar a quienes deseen continuar en la investigación de estos sistemas de clasificación hiperespectral.

Finalmente, se comenta la aplicabilidad que tienen los sistemas desarrollados en este proyecto, de modo que pueda aprovecharse el conocimiento adquirido para integrarlo a las técnicas de diagnóstico médico que siguen esta línea de investigación.

7.2 Análisis de los objetivos alcanzados

Como se comentó en el apartado 1.3 de esta memoria, el principal objetivo al inicio de este proyecto era desarrollar un sistema clasificador que permitiera diferenciar entre muestras tumorales y no tumorales, e identificar el tipo y el grado del tumor de las muestras tumorales, a partir de la información contenida en las firmas espectrales de los tejidos a clasificar. Además, se pretendía validar los algoritmos de clasificación implementados empleando las muestras proporcionadas por el Hospital Universitario de Southampton, y comparar las prestaciones de los sistemas a tenor de los resultados obtenidos al clasificar cada una de las campañas por separado y la fusión de ambas campañas.

En el Capítulo 6 se han reflejado los resultados de validación que se fueron obteniendo en los diversos sistemas desarrollados a medida que se avanzaba en este proyecto. Simultáneamente, en el transcurso del mismo capítulo se realizan comparativas continuamente entre los sistemas cuando se exponen y explican sus respectivos resultados. Adicionalmente, se exponen y comparan los resultados de validación de los sistemas más importantes en los apartados destinados para ello en ese capítulo: apartados 6.3.3 y 6.4.5.

Como puede consultarse en sendos apartados, tanto con los sistemas de árboles jerárquicos como en los sistemas de clasificadores multiclase, se consigue una diferenciación perfecta de las muestras tumorales y sanas de la campaña 1; mientras que para la campaña 2 y la fusión de las campañas, no se llega a discriminar estas clases con el 100% de las muestras, aunque sí se llega a tasas muy altas empleando sistemas multiclase con el *kernel* de tipo RBF y parámetros optimizados.

Debe tenerse en cuenta que, a diferencia de la campaña 1, en la campaña 2 no se hace eliminación de muestras mediante el algoritmo de mapeo de ángulo espectral, por lo tanto, se hace más complicado lograr una discriminación perfecta de las clases en los conjuntos que contienen muestras de la campaña 2. La razón de no eliminar ninguna muestra de la campaña 2 en este proyecto, es que se pretendió abarcar el problema tratando de ser lo más fiel posible a la heterogeneidad de firmas espectrales que pueden encontrarse en los diversos tipos de tejido. No obstante, se decidió mantener la selección final de muestras de la campaña 1 que se había realizado en el proyecto titulado “*Clasificación de tumores cerebrales mediante espectroscopia utilizando Máquinas de Vectores Soporte (SVM)*” [4], para seguir con su línea de trabajo y poder cotejar adecuadamente los resultados.

7.3 Conclusiones

En el transcurso de este proyecto se han ido obteniendo diferentes conclusiones relacionadas con los sistemas desarrollados y con los resultados que estos proporcionaron. En este apartado se destacan las que se consideran más relevantes.

En primer lugar, deben considerarse las principales ventajas y desventajas inherentes a los métodos de clasificación desarrollados en este PFC para afrontar la clasificación de muestras de tejido cerebral.

Por un lado, si se decide emplear un sistema basado en árboles de clasificación, debe tenerse en cuenta que los resultados de los clasificadores en determinados niveles, pueden verse afectados por la propagación de errores. Es decir, si a un clasificador se le propagan errores, estas muestras se clasificarán mal independientemente de la predicción que se realice en el clasificador. Por tanto, si se propagan muchos errores, las métricas de dicho clasificador no reflejarán información veraz en relación a su capacidad de discriminación o robustez. Este es un hándicap que no manifiestan los sistemas multiclase, ya que no presentan propagación de errores.

Por otro lado, si se decide emplear un sistema de clasificación multiclase debe tenerse presente que el entrenamiento del sistema se enfoca a las clases finales. Este hecho supone que se disponga de menor cantidad de muestras de entrenamiento para determinadas clases, por tanto, se obtendrá un sistema menos robusto para esas clases. Es decir, en los sistemas multiclase los *clusters* de las diferentes clases se consideran *clusters* independientes, mientras que, en los árboles jerárquicos, los *clusters* de las clases: “grado I”, “grado II”, “grado III”, “grado IV” y “tumor secundario” formarán parte del *cluster* “tumoral”, y al tener mayor cantidad de muestras, será más propicio discriminar las muestras tumorales de las no tumorales.

Por otra parte, considerando los resultados reflejados en la Tabla 6.15, la Tabla 6.16, la Tabla 6.36 y la Tabla 6.37, los sistemas jerárquicos y multiclases diseñados en este proyecto con cualquier tipo de *kernel* discriminan perfectamente las muestras sanas de las tumorales de la campaña 1. Sin embargo, es el sistema multiclase con tipo de *kernel* RBF y parámetros óptimos: $(\gamma, C) = (2^{-17}, 2^{21})$, el que presenta mejores resultados en la clasificación de esta campaña.

En la Figura 7.1 se ilustra el promedio de sensibilidad para las clases finales de la campaña 1, según los sistemas más relevantes diseñados en este proyecto. Para obtener las métricas promedio de sensibilidad de las clases: “grado II”, “sustancia blanca” y “grado III”, del sistema de clasificación jerárquica; basta con tomar los promedio de especificidad de las clases bajo test: “grado alto”, “sustancia gris” y “grado IV” respectivamente, de la Tabla 6.15, puesto que al tratarse de clasificadores biclase, las métricas son complementarias. Se ha tomado los resultados de los experimentos realizados con 70% de *training* de esta tabla.

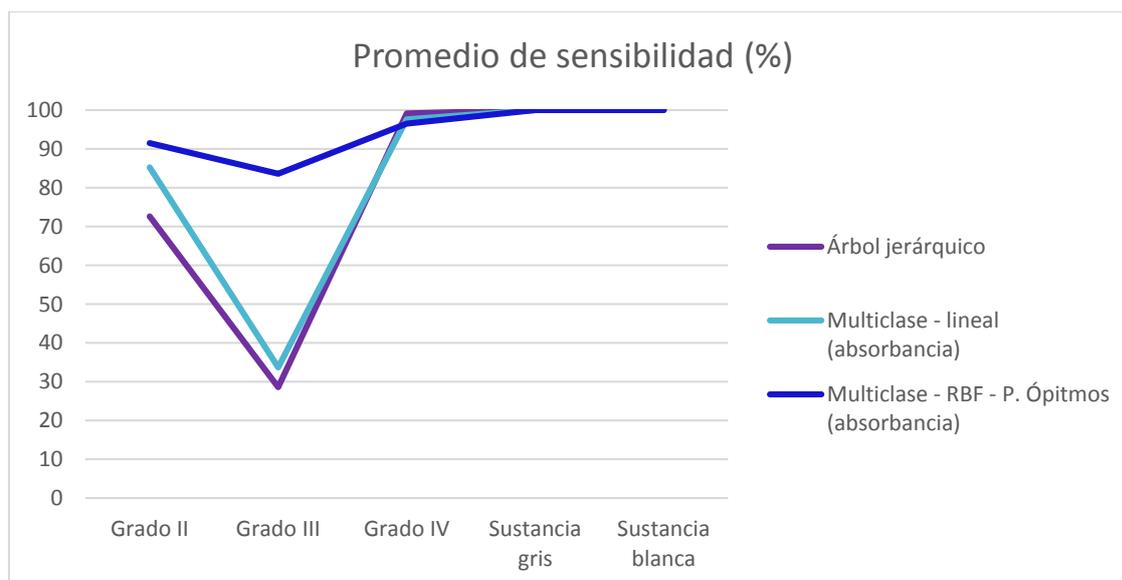


Figura 7.1: Promedio de sensibilidad de las clases finales en los sistemas más relevantes del proyecto (campaña 1).

En cuanto a la campaña 2, en cualquiera de los dos métodos desarrollados se cometen errores al discriminar la clase tumoral de la no tumoral. No obstante, es a través de los clasificadores multiclase con los que se consiguen mejores resultados, y con respecto a todas las clases de esta campaña, el sistema que mejores resultados ofrece es el sistema multiclase con tipo de *kernel* RBF y con parámetros óptimos $(\gamma, C) = (2^{-7}, 2^{14})$.

En la Figura 7.2 se ilustra el promedio de sensibilidad para las clases finales de la campaña 2, según los sistemas más relevantes diseñados en este proyecto. Para obtener las métricas promedio de sensibilidad de las clases: “no tumoral”, “tumor secundario”, “grado II” y “grado IV”, del sistema de clasificación jerárquica; basta con tomar los promedio de especificidad de las clases bajo test: “tumor”, “tumor primario”, “grado I” y “grado III” respectivamente, de la Tabla 6.16, puesto que al tratarse de clasificadores biclase, las métricas son complementarias.

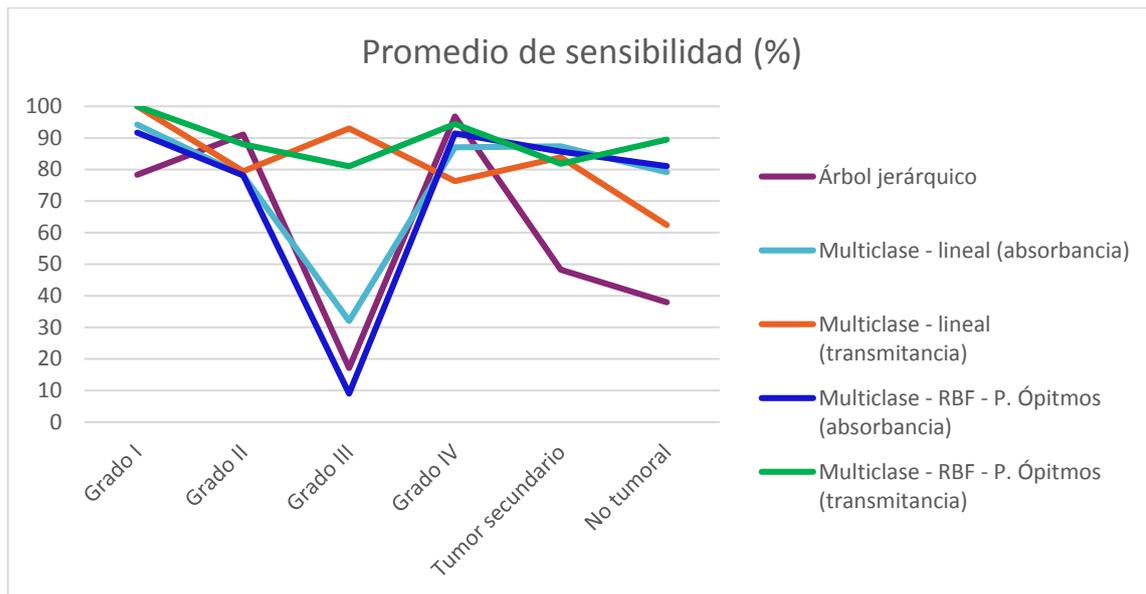


Figura 7.2: Promedio de sensibilidad de las clases finales en los sistemas más relevantes del proyecto (campaña 2).

De igual manera, con la fusión de las campañas los mejores resultados se obtuvieron empleando el sistema de clasificación multiclase, con tipo de *kernel* RBF y los mismos parámetros óptimos de la campaña 2. En ambos casos (clasificación de la campaña 2 y de la fusión de las campañas), se logra una mejor clasificación con la respuesta de los datos en transmitancia.

En la Figura 7.3 se ilustra el promedio de sensibilidad para las clases finales de la fusión de las campañas, según los sistemas más relevantes diseñados en este proyecto.

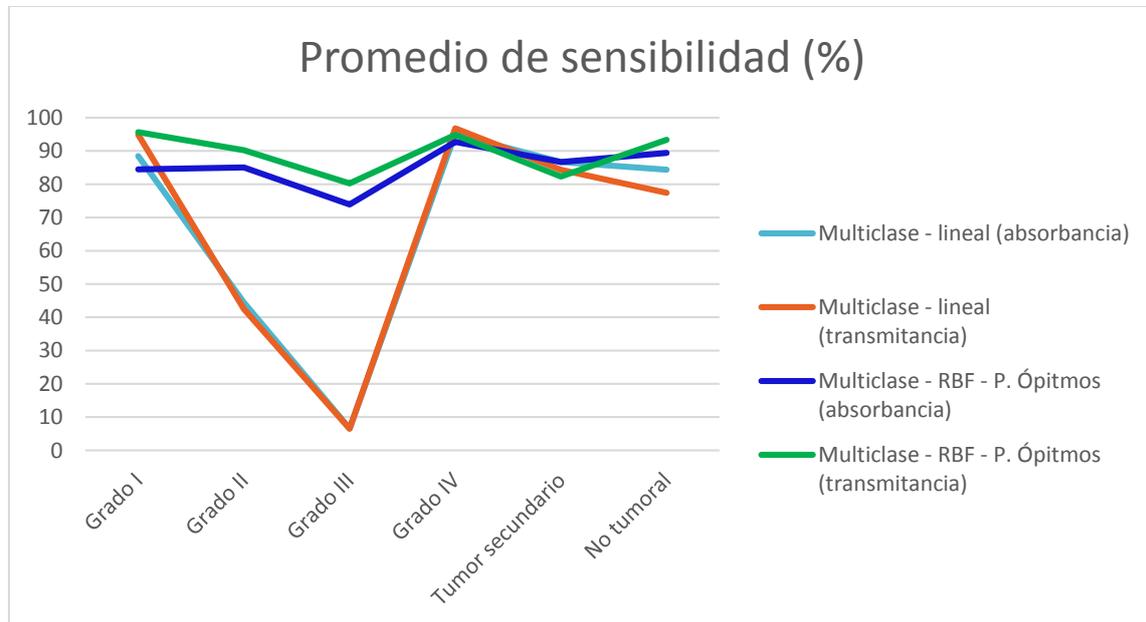


Figura 7.3: Promedio de sensibilidad de las clases finales en los sistemas más relevantes del proyecto (campaña fusión).

Se ha optado por prescindir de las gráficas de promedios de especificidad en este apartado de conclusiones, puesto que en los sistemas multiclase, la tasa de verdaderos negativos no refleja una proporción de muestras correctamente clasificadas sino correctamente rechazadas de la clase considerada bajo test.

7.4 Líneas futuras

Tomando como base de desarrollo el trabajo realizado en el presente Proyecto Fin de Carrera, se realizan las siguientes propuestas, referentes a nuevas líneas de trabajo que pueden ser seguidas para dar continuidad a esta línea de investigación:

- Validación de los sistemas desarrollados empleando muestras de otros tipos de tumores y tejidos sanos.
- Desarrollar sistemas multiclase que utilicen otros tipos de *kernels* y estudios de optimización de parámetros, con la finalidad de tratar de mejorar los resultados.
- Desarrollar sistemas jerárquicos que utilicen otros tipos de *kernels* y estudios de optimización de parámetros, que también permitan clasificar la fusión de las campañas.

- Implementar nuevos sistemas basados en otras técnicas de clasificación para clasificar la base de datos empleada en este proyecto, y realizar comparativas.
- Integrar los diferentes sistemas de clasificación mediante el método “el más votado”.

7.5 Aplicabilidad del proyecto

La principal aplicación de este proyecto se encuadra en el diagnóstico médico de tumores cerebrales e identificación del grado o tipo (primario o secundario) de dichos tumores. Se pretende que los sistemas desarrollados en este proyecto, que mejores resultados proporcionan, puedan ser optimizados e integrados con otros sistemas que clasifiquen tumores cerebrales mediante cámaras de imágenes hiperespectrales.

Los sistemas de clasificación de tumores cerebrales mediante imágenes, permiten a los médicos identificar de forma automática los límites de un tumor para conseguir una completa resección del mismo, dañándose el tejido cerebral sano lo menos posible. No obstante, debido a la repercusión que puede tener una identificación errónea de los tejidos, siempre será el médico el que tome la decisión final, empleando el sistema desarrollado como una herramienta de ayuda al diagnóstico final.

Bibliografía

- [1] D. Glotsos, P. Spyridonos, P. Petalas, D. Cavouras, V. Zolota, P. Dadioti, I. Lekka, and G. Nikiforidis, "A hierarchical decision tree classification scheme for brain tumour astrocytoma grading using support vector machines," in *3rd International Symposium on Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the*, vol. 2, pp. 1034–1038, IEEE.
- [2] WebMD, "Brain cancer and gliomas." Disponible en: www.webmd.com/cancer/brain-cancer/malignant-gliomas, 2017. [En línea; visitado el 24 de Abril de 2017].
- [3] N. B. T. Society, "Glioblastoma multiforme." Disponible en: www.braintumor.org/patients-family-friends/about-brain-tumors/tumortypes/glioblastoma-multiforme.html, 2016. [En línea; visitado el 8 de Abril de 2016].
- [4] M. C. D. Falcón, "Clasificación de tumores cerebrales mediante espectroscopia utilizando máquinas de vectores soporte (svm)," PFC, Universidad de Las Palmas de Gran Canaria, 2016.
- [5] A. T. N. Facility, "Obtaining astronomical spectra - spectrographs." Disponible en: www.atnf.csiro.au/outreach//education/senior/astrophysics/spectrographs.html, 2017. [En línea; visitado el 25 de Abril de 2017].
- [6] C.-I. Chang, *Hyperspectral Data Processing: Algorithm Design and Analysis*. John Wiley & Sons, Inc., mar 2013.
- [7] B. Fei, H. Akbari, and L. V. Halig, "Hyperspectral imaging and spectral-spatial classification for cancer detection," in *2012 5th International Conference on BioMedical Engineering and Informatics*, IEEE, oct 2012.
- [8] R. Onmeda, "Anatomía del cerebro." Disponible en: www.onmeda.es/anatomia/-anatomia_cerebro, 2012. Revisión médica: Dra. María Fernanda Pedrero, [En línea; visitado el 26 de Abril de 2017].
- [9] A. R. Martín, "Cerebro humano: Funciones y partes." Disponible en: www.lifeder.com/funciones-partes-del-cerebro-humano/, 2012. [En línea; visitado el 26 de Abril de 2017].
- [10] Sabelotodo.org, "Anatomía del cerebro humano." Disponible en: www.sabelotodo.org/anatomia/cerebro.html. [En línea; visitado el 26 de Abril de 2017].
- [11] M. J. N. Borge, "Organización funcional del sistema nervioso." Disponible en: ocw.unican.es/ciencias-de-la-salud/fisiologia-humana-2011-g367/material-de-clase/bloque-tematico-6.-fisiologia-del-sistema-nervioso/tema-1.-organizacion-funcional-del-sistema/tema-1.-organizacion-funcional-del-sistema, 2011. [En línea; visitado el 26 de Abril de 2017].

- [12] D. Matamala, "Introducción al estudio de la anatomía macroscópica del sistema nervioso central." Disponible en: es.slideshare.net/AlanZurbrigk/introduccion-al-estudio-del-la-anatoma-macroscopica-del-sistema-nervioso-central, 2012. [En línea; visitado el 26 de Abril de 2017].
- [13] O. College, *Biology*, ch. 35 | THE NERVOUS SYSTEM, p. 1026. May 2013. [En línea; visitado el 26 de Abril de 2017], Disponible en: cnx.org/contents/-GFy_h8cu@9.85:c9j4poj@3/Neurons-and-Glial-Cells.
- [14] aecc, "Tipos de tumores del sistema nervioso central." Disponible en: www.aecc.es/-SobreElCancer/CancerPorLocalizacion/Sistemanerviosocentral/Paginas/-Tipos.aspx, 2012. [En línea; visitado el 26 de Abril de 2017].
- [15] D. W. Kufe, J. F. Holland, R. E. Pollock, R. C. Bast, and R. R. Weichselbaum, *Holland-Frei Cancer Medicine*. BC Decker, 6ª ed., 2003.
- [16] F. R. Oliveros and M. L. Bouzas, *Gliomas del encéfalo*. Universidad de Santiago de Compostela, 2007.
- [17] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvett, B. W. Scheithauer, and P. Kleihues, "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropathologica*, vol. 114, pp. 97–109, jul 2007.
- [18] Sanitas, "Tipos de carcinoma." Disponible en: www.sanitas.es/sanitas/seguros/es-particulares/biblioteca-de-salud/cancer/carcinoma.html. [En línea; visitado el 26 de Abril de 2017].
- [19] A. C. Society, "Tipos de cáncer de pulmón no microcítico." Disponible en: www.cancer.org/es/cancer/cancer-de-pulmon-no-microcitico/acerca/que-es-cancer-de-pulmon-no-microcitico.html, 2016. [En línea; visitado el 26 de Abril de 2017].
- [20] aecc, "Cáncer de pulmón." Disponible en: www.aecc.es/SobreElCancer/-CancerPorLocalizacion/cancerdepulmon/Paginas/tipos.aspx, 2012. [En línea; visitado el 26 de Abril de 2017].
- [21] tuotromedico.com, "Cáncer gástrico." Disponible en: www.tuotromedico.com/-temas/cancer-gastrico.htm. [En línea; visitado el 26 de Abril de 2017].
- [22] conexioncancer.es, "Cáncer gástrico." Disponible en: conexioncancer.es/tipos-de-cancer/informacion-general-sobre-el-cancer-gastrico/. [En línea; visitado el 26 de Abril de 2017].
- [23] A. E. N. Frómeta, "La espectroscopía." Disponible en: www.navarrof.orgfree.com/-Docencia/QuimicaAnalitica/Espectro/espectroo.htm. [En línea; visitado el 27 de Abril de 2017].
- [24] espectrometria.com, "Espectrometría." Disponible en: www.espectrometria.com. [En línea; visitado el 27 de Abril de 2017].
- [25] N. A. Díaz, J. A. B. Ruiz, E. F. Reyes, A. G. Cejudo, J. J. Novo, J. P. Peinado, F. T. Meléndez-Valdés, and I. T. Fiñana, "Espectrofometría: Espectros de absorción y

cuantificación colorimétrica de biomoléculas.” Disponible en: www.uco.es/dptos/-bioquimica-biol-mol/pdfs/o8_ESPECTROFOTOMETR%C3%8DA.pdf. [En línea; visitado el 27 de Abril de 2017].

- [26] M. R. Muriel, “Caracterización de imágenes hiperespectrales utilizando support vector machines y técnicas de extracción de características,” proyecto fin de carrera, Escuela Politécnica de la Universidad de Extremadura, Diciembre 2009.
- [27] R. N. Clark, *Spectroscopy of rocks and minerals, and principles of spectroscopy, in Manual of Remote Sensing*, vol. 3 of *Remote Sensing for the Earth Sciences*, ch. 1, pp. 3–58. New York: John Wiley and Sons, 3 ed., Junio 1999.
- [28] D. Landgrebe, “Hyperspectral image data analysis,” *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 17–28, 2002.
- [29] J. W. Boardman, F. A. Kruse, and R. O. Green, *Mapping target signatures via partial unmixing of AVIRIS data*. Summaries of the VI JPL Airborne Earth Science Workshop, 1995.
- [30] P. fuei Hsieh and D. Landgrebe, *Classification of high dimensional data*. PhD thesis, Electrical and Computer Engineering, Purdue University, 1998.
- [31] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, M. R. Olah, and O. Williams, “Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS),” *Remote Sensing of Environment*, vol. 65, pp. 227–248, sep 1998.
- [32] G. Lu and B. Fei, “Medical hyperspectral imaging: a review,” *Journal of Biomedical Optics*, vol. 19, p. 010901, jan 2014.
- [33] Q. Li, X. He, Y. Wang, H. Liu, D. Xu, and F. Guo, “Review of spectral imaging technology in biomedical engineering: achievements and challenges,” *Journal of Biomedical Optics*, vol. 18, p. 100901, oct 2013.
- [34] G. Zonios, L. T. Perelman, V. Backman, R. Manoharan, M. Fitzmaurice, J. V. Dam, and M. S. Feld, “Diffuse reflectance spectroscopy of human adenomatous colon polyps in vivo,” *Applied Optics*, vol. 38, p. 6628, nov 1999.
- [35] L. V. Wang and H.-I. Wu, “Introduction,” in *Biomedical Optics*, pp. 1–15, John Wiley & Sons, Inc., oct 2012.
- [36] C. Balas, C. Pappas, and G. Eptropou, “Multi/hyper-spectral imaging,” in *Handbook of Biomedical Optics*, pp. 131–164, CRC Press, jun 2011.
- [37] V. V. Tuchin, *Tissue Optics: Light Scattering Methods and Instruments for Medical Diagnosis*. Society of Photo-Optical Instrumentation Engineers (SPIE), 2015.
- [38] D. G. Ferris, R. A. Lawhead, E. D. Dickman, N. Holtzapple, J. A. Miller, S. Grogan, S. Bambot, A. Agrawal, and M. L. Faupel, “Multimodal hyperspectral imaging for the noninvasive diagnosis of cervical neoplasia,” *Journal of Lower Genital Tract Disease*, vol. 5, pp. 65–72, apr 2001.

- [39] M. C. Pierce, R. A. Schwarz, V. S. Bhattar, S. Mondrik, M. D. Williams, J. J. Lee, R. Richards-Kortum, and A. M. Gillenwater, "Accuracy of in vivo multimodal optical imaging for detection of oral neoplasia," *Cancer Prevention Research*, vol. 5, pp. 801–809, may 2012.
- [40] P. Y. C. N. Mazzer, C. H. Barbieri, N. Mazzer, and V. P. S. Fazan, "Morphologic and morphometric evaluation of experimental acute crush injuries of the sciatic nerve of rats," *Journal of Neuroscience Methods*, vol. 173, pp. 249–258, aug 2008.
- [41] G. Bellisola and C. Sorio, "Infrared spectroscopy and microscopy in cancer research and diagnosis," *American Journal of Cancer Research*, vol. 2, pp. 1–21, January 2011.
- [42] A. Nouvong, B. Hoogwerf, E. Mohler, B. Davis, A. Tajaddini, and E. Medenilla, "Evaluation of diabetic foot ulcer healing with hyperspectral imaging of oxyhemoglobin and deoxyhemoglobin," *Diabetes Care*, vol. 32, pp. 2056–2061, jul 2009.
- [43] R. Salzer, H. H. Mantsch, J. Mansfield, E. N. Lewis, and G. Steiner, "Infrared and raman imaging of biological and biomimetic samples," *Fresenius Journal of Analytical Chemistry*, vol. 366, pp. 712–726, mar 2000.
- [44] A. M. Siddiqi, H. Li, F. Faruque, W. Williams, K. Lai, M. Hughson, S. Bigler, J. Beach, and W. Johnson, "Use of hyperspectral imaging to distinguish normal, precancerous, and cancerous cells," *Cancer*, vol. 114, pp. 13–21, jan 2008.
- [45] M. Á. T. Hernández, "Identificación de tumores cerebrales mediante el estudio de la forma y la composición de los tejidos usando imágenes hiperespectrales," PFC, Escuela de Ingeniería de Telecomunicación y Electrónica, Universidad de Las Palmas de Gran Canaria, Septiembre 2015.
- [46] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 12–16, 2002.
- [47] D. Stein, S. Beaven, L. Hoff, E. Winter, A. Schaum, and A. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 58–69, 2002.
- [48] A. Plaza, P. Gamba, K. Bakos, B. Waske, and J. B. Diaz, "Hyperinet_d4.1 - Processing Chain Definition Report," Enero 2008.
- [49] A. Plaza, P. Gamba, K. Bakos, and B. Waske, "Hyperinet_d4.2 - Processing Chain Implementation report," Enero 2009.
- [50] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, pp. 55–63, jan 1968.
- [51] K. Fukunaga, *Introduction to Statistical Pattern Recognition (2ª Ed.)*. San Diego, CA, USA: Academic Press Professional, Inc., 2 ed., 1990.
- [52] C.-I. Chang and A. Ifarraguerri, "Multispectral and hyperspectral image analysis with projection pursuit," *IEEE Trans, on Geoscience and Remote Sensing*, vol. 38, pp. 2529–2538, November 2000.

-
- [53] S. Subramanian, N. Gat, A. Ratcliff, and M. Eismann, "Hyperspectral data reduction using principal components transformation," *Proc. X -ASA/JPL Airborne Earth Science Workshop*, 2000.
 - [54] A. Green, M. Berman, P. Switzer, and M. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, no. 1, pp. 65–74, 1988.
 - [55] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
 - [56] C.-I. Chang and H. Ren, "An experiment-based quantitative and comparative analysis of target detection and image classification algorithms for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, pp. 1044–1063, mar 2000.
 - [57] J. M. Chen, "Spatial scaling of a remotely sensed surface parameter by contexture," *Remote Sensing of Environment*, vol. 69, pp. 30–42, jul 1999.
 - [58] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 62, pp. 77–89, oct 1997.
 - [59] S.-S. Chiang, C.-I. Chang, and I. Ginsberg, "Unsupervised target detection in hyperspectral images using projection pursuit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, pp. 1380–1391, jul 2001.
 - [60] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.
 - [61] F. V. Padilla, "Implementación eficiente de clasificadores prior-svm para matlab," PFC, Escuela Politécnica Superior, Universidad Carlos III de Madrid, Abril 2010.
 - [62] J. S. R. Fernández, "Clasificación de terrenos en imágenes multiespectrales mediante máquinas de vectores soporte. Los problemas de la clasificación, 29," proyecto fin de carrera, Universidad Carlos III de Madrid, 2004.
 - [63] J. P. Theiler and G. Gisler, "titlecontiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation/title," in *Algorithms, Devices, and Systems for Optical Information Processing* (B. Javidi and D. Psaltis, eds.), SPIE, oct 1997.
 - [64] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*. Springer Berlin Heidelberg, 1993.
 - [65] J. Sweet, J. Granahan, and M. Sharp, "An objective standard for hyperspectral image quality," *Proc. IX -ASA/JPL Airborne Earth Science Workshop*, 2000.
 - [66] X. Wen and X. Yang, "An unsupervised classification method for hyperspectral image using spectra clustering," in *2008 IEEE International Symposium on Knowledge Acquisition and Modeling Workshop*, pp. 1117 – 1120, IEEE, dec 2008.

- [67] G. Girouard, A. Bannari, A. E. Harti, and A. Desrochers, "Validated spectral angle mapper algorithm for geological mapping: Comparative study between quickbird and landsat-tm," *GeoImagery Bridging Continents*, pp. 599–604, 2004.
- [68] G. Allende, P. Beatriz, C. Portilla, O. María, M. Serrano, J. María, C. de Cos, A. María, L. Higuera, and J. Miguel, "Clasificación en tiempo real de materias primas mediante un sensor espectroscópico de imagen y técnicas lineales de procesamiento de datos," (Bilbao), pp. 543–548, 5ª Reunión Española de Optoelectrónica, OPTOEL07, 2007.
- [69] Wikipedia, "Regresión logística multinomial — wikipedia, la enciclopedia libre." Disponible en: es.wikipedia.org/w/index.php?title=Regresi%C3%B3n_log%C3%ADstica_multinomial&oldid=98709276, 2017. [En línea; visitado 28-abril-2017].
- [70] I. Dopido, J. Li, P. R. Marpu, A. Plaza, J. M. B. Dias, and J. A. Benediktsson, "Semisupervised self-learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 4032–4044, jul 2013.
- [71] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. A. T. Figueiredo, "On semi-supervised classification," pp. 721–728, MIT Press, 2005.
- [72] E. M. F. de Velasco, "Herramientas de optimización convexa y aplicaciones en telecomunicaciones," PFC, Escuela Politécnica Superior, Universidad Carlos III de Madrid, Leganés, Septiembre 2011.
- [73] S. Raschka, "Linear Discriminant Analysis. Bit by Bit." Disponible en: sebastianraschka.com/Articles/2014_python_lda.html, 2014. [En línea; visitado el 29 de Abril de 2017].
- [74] R. Gutierrez-Osuna, "Lecture 10: Linear discriminant analysis." vision.eecs.ucf.edu/courses/cap5415/fall2011/Lecture-14.5-LDA.pdf. [En línea; visitado el 30 de Abril de 2017].
- [75] P. Gong, "Conifer species recognition: An exploratory analysis of in situ hyperspectral data," *Remote Sensing of Environment*, vol. 62, pp. 189–200, nov 1997.
- [76] M. CLARK, D. ROBERTS, and D. CLARK, "Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales," *Remote Sensing of Environment*, vol. 96, pp. 375–398, jun 2005.
- [77] O. Meshi, "Learning symmetric relational markov random fields," PhD, School of Computer Science and Engineering, The Hebrew University of Jerusalem, December 2007.
- [78] M. Cuadra, L. Cammoun, T. Butz, O. Cuisenaire, and J.-P. Thiran, "Comparison and validation of tissue modelization and statistical classification methods in t1-weighted MR brain images," *IEEE Transactions on Medical Imaging*, vol. 24, pp. 1548–1565, dec 2005.
- [79] M. Pujol, R. Rizo, P. Arques, P. Compañy, F. Escolano, and R. Molina, "Aplicación de los modelos de campos aleatorios de markov en visión artificial," *Revista Electrónica de Visión por Computador, Universidad de Alicante*, octubre 2000.

-
- [80] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York, 2009.
- [81] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [82] Wikipedia, “Random forest — wikipedia, la enciclopedia libre.” Disponible en: es.wikipedia.org/w/index.php?title=Random_forest&oldid=91962254, 2016. [En línea; visitado 2-mayo-2017].
- [83] Y. Qi, “Random forest for bioinformatics,” in *Ensemble Machine Learning*, pp. 307–323, Springer US, 2012.
- [84] D. Shiffman, *The Nature of Code*. Daniel Shiffman, December 2012.
- [85] D. S. Bogatinov, M. Bogdanoski, and S. Angelevski, “AI-based cyber defense for more secure cyberspace,” in *Advances in Digital Crime, Forensics, and Cyber Terrorism*, pp. 220–237, IGI Global.
- [86] C. Stergiou and D. Siganos, “Neural networks.” Disponible en: www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html. [En línea; visitado el 3 de Mayo de 2017].
- [87] J. Burger, “A basic introduction to neural networks.” Disponible en: pages.cs.wisc.edu/~bolo/shipyard/neural/local.html. [En línea; visitado el 3 de Mayo de 2017].
- [88] V. N. Vapnik, *Statistical Learning Theory*. October 1998.
- [89] C. Cortes and V. Vapnik *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [90] B. Scholkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999.
- [91] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the fifth annual workshop on Computational learning theory - COLT 92*, ACM Press, 1992.
- [92] C. J. Burges *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [93] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, “An introduction to kernel-based learning algorithms,” *IEEE Transactions on Neural Networks*, vol. 12, pp. 181–201, mar 2001.
- [94] G. Mercier and M. Lennon, “Support vector machines for hyperspectral image classification with spectral-based kernels,” in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, IEEE.
- [95] G. M. Foody, “RVM-based multi-class classification of remotely sensed data,” *International Journal of Remote Sensing*, vol. 29, pp. 1817–1823, mar 2008.
- [96] C. J. Burges and B. Schölkopf, “Improving the accuracy and speed of support vector machines,” *Advances in Neural Information Processing Systems*, pp. 375–381, May 1997.

- [97] J. Dukart, "Support Vector Machine Classification - Basic Principles and Application." Disponible en: www.unil.ch/files/live/sites/lren/files/shared/-map_UNIL_biophore/-Support_Vector_Machine_Classification_Basic_Principles.pdf, 2012. [En línea; visitado el 4 de Mayo de 2017].
- [98] G. Rellier, X. Descombes, and J. Zerubia, "Local registration and deformation of a road cartographic database on a SPOT satellite image," *Pattern Recognition*, vol. 35, pp. 2213–2221, oct 2002.
- [99] V. Madhok and D. Landgrebe, *Spectral-Spatial Analysis of Remote Sensing Data: An Image Model and A Procedural Design*. Tesis doctoral, School of Electrical Engineering and Computer Science, Purdue University, January 1998.
- [100] S. Tadjudin and D. Landgrebe, *Classification of High Dimensional Data With Limited Training Samples*. Tesis doctoral, School of Electrical Engineering and Computer Science, Purdue University, 1998.
- [101] E. Næsset, "Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data," *Remote Sensing of Environment*, vol. 80, pp. 88–99, apr 2002.
- [102] R. Congalton, "Considerations and techniques for assessing the accuracy of remotely sensed data," in *12th Canadian Symposium on Remote Sensing Geoscience and Remote Sensing Symposium IGARSS*, vol. 3, pp. 1847–1850, IEEE, 1989.
- [103] S. V. Stehman, "Practical implications of design-based sampling inference for thematic map accuracy assessment," *Remote Sensing of Environment*, vol. 72, pp. 35–45, apr 2000.
- [104] B. M. Steele, J. Winne, and R. L. Redmond, "Estimation and mapping of misclassification probabilities for thematic land cover maps," *Remote Sensing of Environment*, vol. 66, pp. 192–202, nov 1998.
- [105] G. Jager and U. Benz, "Measures of classification accuracy based on fuzzy similarity," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, pp. 1462–1467, may 2000.
- [106] M. V. Becker, L. Carrera, and E. Fernández, "Pruebas diagnósticas: Algunas probabilidades importantes para su aplicación en medicina," in *X Congreso Latinoamericano de Sociedades de Estadística*, (Córdoba, Argentina), Octubre 2012.
- [107] Y. M. P. y Juana Robles González, "Una aplicación de la regla de bayes en ciencias de la salud," in *XXVI Simposio Internacional de Estadística*, (Sincelejo, Sucre, Colombia), Agosto 2016.
- [108] CobaltNiche, "Agilent 600 Series FTIR Spectrometer & Microscope." Disponible en: www.cobaltniche.com/portfolio/all/1/9-agilent-600-series-ftir-spectrometer-microscope, 2017. [En línea; visitado el 8 de Mayo de 2017].
- [109] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM – A Library for Support Vector Machines." Disponible en: www.csie.ntu.edu.tw/~cjlin/libsvm/, 2016. [En línea; visitado el 16 de Mayo de 2017].

-
- [110] R. Pourreza-Shahri, F. Saki, N. Kehtarnavaz, P. Leboulluec, and H. Liu, "Classification of ex-vivo breast cancer positive margins measured by hyperspectral imaging," in *2013 IEEE International Conference on Image Processing*, IEEE, sep 2013.
- [111] R. A. Española, "Diccionario de la lengua española. 23^a edición." [Consultado en: www.rae.es], 2014.
- [112] C. N. de Investigaciones Oncológicas, "Histopatología – introducción." Disponible en: www.cnio.es/es/servicios/histopatologia/default.aspx. [En línea; visitado el 28 de Abril de 2017].
- [113] M. González, "Cromóforo." Disponible en: quimica.laguia2000.com/enlaces-quimicos/cromoforo, 2011. [En línea; visitado el 27 de Abril de 2017].
- [114] B. S. Giraldo, *Espectroscopía óptica de fluorescencia aplicada al soporte de diagnóstico médico de precánceres de tejidos de cuello uterino*. PhD, Facultad de Ingeniería y Arquitectura, Universidad Nacional de Colombia, 2009.
- [115] V. A. T. Marcani, "Tejido conectivo." Disponible en: www.tejidoconectivo.galeon.com/index.html. [En línea; visitado el 28 de Abril de 2017].
- [116] I. G. Dópido, *New Techniques for Hyperspectral Image Classification*. PhD thesis, Universidad de Extremadura, 2013.
- [117] DBpedia, "Verdad terreno." Disponible en: es.dbpedia.org/page/Verdad_Terreno. [En línea; visitado el 28 de Abril de 2017].
- [118] G. de Topología Computacional y Matemática Aplicada, "Metodos estadísticos (co-ocurrencias)." Disponible en: <http://www.grupo.us.es/gtocom/pid/pid10/-texturaestad.htm>. [En línea; visitado el 28 de Abril de 2017].
- [119] I. Laradji, "Non-convex optimization." Disponible en: www.cs.ubc.ca/labs/lci/mlrg/slides/non_convex_optimization.pdf, 2017. [En línea; visitado el 4 de Mayo de 2017].
- [120] Neural Information Processing Systems Foundation, "Nips 2016 workshop on nonconvex optimization for machine learning: Theory and practice." Disponible en: www.cs.ubc.ca/labs/lci/mlrg/slides/non_convex_optimization.pdf, 2016. [En línea; visitado el 4 de Mayo de 2017].

PRESUPUESTO



P.1 Introducción

El presupuesto del presente Proyecto Fin de Carrera se ha valorado en función de la última lista de Honorarios Orientativos publicada por el Colegio Oficial de Ingenieros de Telecomunicación, denominada “Costes Estimados de Trabajos Profesionales” correspondiente al año 2008. Así, el presupuesto se ha estructurado en siete secciones:

- 1 Recursos humanos
- 2 Recursos hardware
- 3 Recursos software
- 4 Material fungible
- 5 Redacción del proyecto
- 6 Aplicación de impuestos
- 7 Presupuesto total

Hay que indicar que el motivo por el que este presupuesto toma como referencia una lista publicada en 2008 se debe a las modificaciones introducidas en el ordenamiento jurídico y a la actuación de los colegios profesionales en la Ley 25/2009 del 22 de diciembre, por la cual no es posible seguir publicando este tipo de listas por parte del COIT.

P.2 Recursos humanos

El coste de los recursos humanos está asociado al tiempo empleado por un ingeniero en la realización del proyecto. Este coste se establece aplicando la fórmula propuesta por el COIT para trabajos por tiempo empleado.

$$\text{Honorarios (€)} = (74,88 \cdot H_n + 96,72 \cdot H_e) \cdot C_t \quad (\text{P. 1})$$

Donde H_n representa el número de horas normales dentro de la jornada laboral, mientras que H_e se considera el número de horas especiales, siendo 260,00 euros el honorario mínimo a cobrar independientemente del número de horas trabajadas. El coeficiente C_t es el factor de corrección que se debe aplicar al número de horas trabajadas, variando en función del número de horas empleadas según la Tabla P . 1.

Tabla P . 1: Factor de corrección en función de horas trabajadas.

Horas trabajadas	C_t
0 – 36	1,00
36 – 72	0,90
72 – 108	0,80
108 – 144	0,70
144 – 180	0,65
180 – 360	0,60
360 – 540	0,55
540 – 720	0,50
720 – 1080	0,45
+1080	0,40

Teniendo una jornada laboral de ocho horas diarias a razón de veinte días laborales cada mes durante 10 meses, el número total de horas empleadas es de 1600 horas normales, siendo cero el número de horas especiales trabajadas. Según la tabla anterior, el factor de corrección que corresponde al número de horas trabajadas es $C_t = 0,40$.

En la Tabla P . 2 se muestra el resultado de la aplicación de la ecuación (P. 1), recogiendo los costes asociados a los recursos humanos libres de impuestos, que ascienden a *cuarenta y siete mil novecientos veintitrés euros con veinte céntimos* (47.923,20 €).

Tabla P . 2: Coste total de recursos humanos.

Concepto	Tiempo trabajado (horas)	Factor de corrección	Importe
Ingeniero	1600	0,40	47.923,20 €
COSTE TOTAL			47.923,20 €

P.3 Recursos hardware

El coste de los recursos hardware viene determinado por los equipos informáticos empleados en la realización del presente Proyecto Fin de Carrera que se listan a continuación:

- Ordenador de sobremesa DELL IJ7670 con procesador Intel® Core™2 Duo 6400 @ 2.13 GHz, con 4 GB de memoria RAM y 250 GB de HDD.
- Impresora Láser Multifunción.

Para el cálculo de los costes de los recursos materiales, hardware y software, se utilizará un sistema de amortización lineal o constante, en el que se supone que el inmovilizado material se deprecia de forma constante a lo largo de su vida útil. La cuota de amortización anual se calcula usando la siguiente fórmula:

$$Cuota\ anual = \frac{Valor\ de\ adquisición - Valor\ residual}{Años\ de\ vida\ útil} \quad (P. 2)$$

El valor residual es el valor teórico que tendrá el elemento analizado después de su vida útil, que se considera nulo para los recursos materiales de este Proyecto Fin de Carrera.

El periodo de amortización de la estación de trabajo y de la impresora láser ha sido estimado en 36 meses. Los costes asociados a los recursos hardware libres de impuestos se recogen en la Tabla P . 3 y ascienden a *doscientos treinta y ocho euros con ochenta y ocho céntimos* (238,88 €).

Tabla P . 3: Coste total de recursos hardware.

Concepto	Coste unitario (€)	Amortización (meses)	Coste mensual (€)	Tiempo de uso (meses)	Importe
Ordenador	800,00	36	25,00	10	222,22 €
Impresora	60,00	36	1,66	10	16,66 €
COSTE TOTAL					238,88 €

P.4 Recursos software

Para los recursos software se utilizan los mismos criterios que en el apartado de recursos hardware, estimando para este caso un periodo de amortización de 24 meses. Las herramientas software utilizadas han sido las siguientes:

- Sistema operativo Microsoft Windows 7 Enterprise Edition 64 bits.
- Microsoft Office Suite 2013.
- Matlab R2016b.

En la Tabla P. 4 se recogen los costes asociados a los recursos software libres de impuestos. Los costes ascienden a *mil ciento ochenta y dos euros con noventa céntimos* (1.182,90 €).

Tabla P. 4: Coste total de recursos software.

Concepto	Coste unitario (€)	Amortización (meses)	Coste mensual (€)	Tiempo de uso (meses)	Importe
Windows 7	300,00	24	12,50	10	125,00 €
Office Suite 2013	539,00	24	22,46	10	224,60 €
Matlab R2016b	2000 €	24	83,33	10	833,30 €
COSTE TOTAL					1.182,90 €

P.5 Material fungible

En este apartado se recopilan los costes relacionados con los materiales utilizados en la realización del proyecto, como son: material de papelería, discos CD-R, tinta de impresora láser y encuadernación.

Los costes asociados al material fungible libres de impuestos ascienden a *doscientos euros* (200 €).

P.6 Redacción del proyecto

De acuerdo a los honorarios orientativos del COIT, el importe de la redacción del presente PFC se calcula mediante la siguiente ecuación:

$$R = 0,07 \cdot P_T \cdot C_r + 0,03 \cdot P_C \cdot C_r \quad (\text{P. 3})$$

Donde P_T es el presupuesto de ejecución material de telecomunicaciones, P_C es el presupuesto de obra civil y C_r es el coeficiente de ponderación por tramos en función del coste del presupuesto. Este Proyecto Fin de Carrera no tiene asociada ninguna obra civil, por lo que P_C es nulo. El presupuesto de ejecución material corresponde a la suma de los cuatro apartados anteriores:

$$P_T = 47.923,20 + 238,88 + 1.182,90 + 200 = 49.544,98 \text{ €}$$

Dado que el presupuesto se encuentra en el tramo, exceso de 30.050 hasta 60.101, se establece un coeficiente de ponderación: $C_r = 0,9$.

Aplicando la ecuación (P. 3):

$$R = 0,07 \cdot 49.544,98 \cdot 0,9 = 3.121,33 \text{ €}$$

Por lo tanto, los costes asociados a la redacción del proyecto libres de impuestos ascienden a *tres mil ciento veintiún euros con treinta y tres céntimos* (3.121,33 €).

P.7 Aplicación de impuestos

En los anteriores apartados se ha recogido cada uno de los costes que se han generado para el desarrollo del Proyecto Fin de Carrera. La realización del mismo está gravada con el Impuesto General Indirecto Canario (IGIC), en un siete por ciento (7%).

P.8 Presupuesto total

En la Tabla P . 5 se recogen todos los costes asociados al Proyecto Fin de Carrera y el importe total del mismo después de impuestos.

Tabla P . 5: Coste total del Proyecto Fin de Carrera.

Concepto	Importe
Recursos humanos	47.923,20 €
Recursos hardware	238,88 €
Recursos software	1.182,90 €
Material fungible	200 €
Redacción del PFC	3.121,33 €
Subtotal	52.666,31 €
IGIC (7%)	3.686,64 €
TOTAL	56.352,95 €

Así pues, D. Ruymán Hernández López declara que el presupuesto para el Proyecto Fin de Carrera “Identificación de Tumores Cerebrales Usando Datos Espectrográficos” asciende a un total de *cincuenta y seis mil trescientos cincuenta y dos euros con noventa y cinco céntimos* (56.352,95 €).

Fdo. Ruymán Hernández López

En Las Palmas de Gran Canaria, Junio de 2017

**PLIEGO
DE
CONDICIONES**



PC.1 Introducción

En el pliego de condiciones se describen las condiciones que se exigen en este Proyecto Fin de Carrera para su correcta reproducibilidad. A continuación, se describe de forma muy breve el conjunto de los componentes hardware y software empleados durante la realización del proyecto, además de las muestras biológicas que se han utilizado.

PC.2 Recursos hardware

- **Espectrómetro Varian FT-IR 600.** Espectrómetro utilizado por el Hospital Universitario de Southampton para la adquisición de muestras de la campaña 1. Este espectrómetro emplea el método de adquisición de imágenes hiperespectrales denominado Espectroscopia Infrarroja mediante Transformada de Fourier (*Fourier Transform Infrared Spectroscopy, FT-IR*) para obtener muestras espectrales de un único píxel, pero con una amplia resolución espectral.
- **Espectrómetro PerkinElmer FT-IR System, Spectrum BX.** Espectrómetro utilizado por el Hospital Universitario de Southampton para la adquisición de muestras de la campaña 2. Este espectrómetro también emplea el método Espectroscopia Infrarroja mediante Transformada de Fourier para obtener muestras espectrales de un único píxel pero con una amplia resolución espectral.
- **Ordenador personal DELL IJ7670.** Ordenador de sobremesa en el que se ha efectuado el procesado de los datos espectrográficos, el desarrollo de los clasificadores y la redacción de este documento. Las principales características de este PC son las siguientes:
 - Procesador Intel® Core™2 Duo 6400 @ 2.13 GHz.
 - 4 GB de memoria RAM.
 - 250 GB de HDD.

PC.3 Recursos software

- **Agilent Resolutions Pro V.5.** Software empleado por el Hospital Universitario de Southampton para la adquisición de muestras a través del espectrómetro PerkinElmer FT-IR System, Spectrum BX.
- **PerkinElmer Spectrum V5.3.1.** Software empleado por el Hospital Universitario de Southampton para la adquisición de muestras a través del espectrómetro Varian FT-IR 600.

- **Microsoft Windows 7.** Sistema operativo Windows 7 Enterprise Edition de 64 bits, instalado en la estación de trabajo en la que se lleva a cabo el este proyecto.
- **Microsoft Office Suite 2013.** Suite ofimática empleada para la elaboración del presente documento, así como para los diversos informes realizados en el transcurso de este PFC.
- **Matlab R2016b.** Herramienta de software matemático que ofrece un entorno de desarrollo integrado (IDE) con un lenguaje de programación propio (lenguaje M). Permite el análisis y la visualización de los datos.

PC.4 Recursos biológicos

Las muestras biológicas que se han empleado en este PFC se corresponden con muestras de tejido cerebral extraídas por médicos del Hospital Universitario de Southampton en diversas intervenciones quirúrgicas. Posteriormente, las muestras fueron etiquetadas por un patólogo que clasificó cada ejemplar con una diagnosis detallada, el tipo de tejido del que provenía y la fecha. Los pacientes de los que proceden las muestras biológicas han dado su consentimiento expreso para su recolección.

Por otra parte, la utilización y compartición anónima de estas muestras está contemplada en el documento Aspectos Éticos (*Ethical Issues*) del proyecto HELICoiD.

Fdo. Ruymán Hernández López

En Las Palmas de Gran Canaria, Junio de 2017