UNIVERSITAT POLITÈCNICA DE VALÈNCIA

# Radiomics for diagnosing and assessing brain diseases: an approach based on texture analysis on magnetic resonance imaging

A dissertation submitted in partial fulfillment of
the requirements for the degree of

**Doctor of Philosophy**

Ph.D. in Technologies for Health and Well-being

*February 2019*


Author: **Rafael Ortiz Ramón**

Supervisor: **Prof. Dr. David Moratal Pérez**

Center for Biomaterials and Tissue Engineering
Universitat Politècnica de València
Valencia, Spain

| | |
|---|---|
| Supervisor: | **Prof. Dr. David Moratal Pérez**<br>*Universitat Politècnica de València*, Valencia, Spain |
| Reviewers Committee: | **Dr. Constantino Carlos Reyes-Aldasoro**<br>*City University of London*, London, United Kingdom |
| | **Prof. Dr. Kjersti Engan**<br>*Universitetet i Stavanger,* Stavanger, Norway |
| | **Dr. Víctor González Castro**<br>*Universidad de León*, León, Spain |
| Members of the Jury: | **Dr. Constantino Carlos Reyes Aldasoro**<br>*City University of London*, London, United Kingdom |
| | **Dr. Roser Sala Llonch**<br>*Universitat de Barcelona*, Barcelona, Spain |
| | **Prof. Dr. Enrique Domingo Guijarro Estelles**<br>*Universitat Politècnica de València*, Valencia, Spain |

---

The research described in this thesis was conducted in the Centre for Biomaterials and Tissue Engineering of the Universitat Politècnica de València in collaboration with:

- Servicio de Radiodiagnóstico, Fundación Instituto Valenciano de Oncología
- Servicio de Radiología, Hospital Universitario de La Ribera
- Centre for Clinical Brain Sciences, University of Edinburgh

This thesis was supported by grant ACIF/2015/078 and grant BEFPI/2017/004 from the Conselleria d'Educació, Investigació, Cultura i Esport of the Valencian Community (Spain).

# Acknowledgments

A mi director de tesis. Gracias David por haberme ofrecido esta oportunidad única y por haber confiado en mí en todo momento desde el primer día. Gracias por haber sabido aconsejarme, escucharme y aguantarme como solo un buen amigo sabe hacer. Para mí has sido mucho más que un director.

A todos los integrantes del Moratal's Team. En especial, gracias Darío, Antonio, Úrsula, Andrés y Silvia por haber sido mi familia durante esta etapa de mi vida. Gracias por todos los momentos vividos juntos y por seguir queriendo estar a mi lado a pesar de todos mis altibajos emocionales. Posiblemente no hubiese llegado hasta aquí si no hubiese sido por vosotros. Ojalá podamos procrastinar juntos muchos años más.

A todos los demás compañeros del CBIT. Gracias por haberme hecho un hueco en vuestro loco mundo de laboratorios, experimentos y siglas.

A todos los expertos con los que he tenido el placer de colaborar. En especial, gracias a Estanis, Enrique y María por haber confiado en mis aptitudes en algún momento de esta etapa y por haber ayudado al desarrollo de esta tesis doctoral.

A mi familia. Gracias a mis padres por no dudar de mí nunca y por ofrecerme el apoyo necesario para convertirme en la persona que soy hoy. Gracias a mi hermana por ser siempre un ejemplo a seguir para mí. Gracias a mi abuela por tanto cariño. Y gracias al que está por venir por haberme hecho tan feliz este último año, aunque aún no lo sepa.

A todos los amigos que se han preocupado e interesado por mí durante esta etapa. En especial, gracias por todas esas cervezas que me han dado fuerzas cuando pensaba que ya no me quedaban.

A Julia. Gracias por ser mi compañera de viaje. Gracias por tu paciencia conmigo y por creer en mí siempre. Gracias porque nada de esto tendría sentido si no estuvieses a mi lado. Chanandler Bong.

# Abstract

Over the last 20 years, researchers have attempted to exploit the information provided by medical images through the computation and evaluation of numerous imaging quantitative parameters in order to help clinicians with the diagnosis and assessment of many lesions and diseases. This practice has been recently named as *radiomics*, and its success relies mainly on the quality and informativeness of the medical images and the subsequent parameters. *Texture analysis* supply a wide range of features that allow quantifying the distinctive heterogeneity of different tissues, especially when obtained from *magnetic resonance imaging* (MRI). With this in mind, we decided to study the possibilities of texture features from MRI in order to characterize and categorize several disorders that affect the human brain. The potential of texture features was analyzed with various *machine learning* approaches, involving different classifiers and feature selection methods so as to find the optimal model to accomplish reliably each specific task. In this thesis, the implemented radiomics methodology was used to perform four independent projects related to four different clinical challenges.

In the first project, we studied the differentiation between glioblastomas (GBMs) and brain metastases (BMs) in conventional MRI. Sometimes these types of brain tumors can be misdiagnosed since they may present a similar radiological profile and the clinical data may be inconclusive. In these cases, a definitive diagnosis can only be made by means of histopathologic analyses. With the aim of avoiding exhaustive and invasive procedures, we studied the discriminatory power of a large amount of 2D texture features extracted from baseline original and filtered T1-weighted images. The results suggest that 2D texture features provide some heterogeneity information of GBMs and BMs that can help in their accurate discernment when using the proper machine learning approach.

In the second project, we analyzed the classification of BMs by their primary site of origin in baseline MRI. A percentage of cancer patients are diagnosed with BM as the first manifestation of an unknown primary tumor. These patients are subjected to exhaustive imaging evaluations and invasive procedures in order to detect the primary tumor, and sometimes the origin remains undiagnosed at the time of death. In order to detect the primary tumor in a faster non-invasive way, we examined the capability of 2D and 3D texture analysis to differentiate BMs derived from the most common primary tumors (lung cancer, breast cancer and melanoma) in T1-weighted images. The results showed that high accuracy was achieved when using a reduced set of 3D descriptors to differentiate lung cancer BMs from breast cancer and melanoma BMs, so volumetric MRI texture features can be useful to differentiate BMs from different primary cancers.

In the third project, we evaluated the hippocampus MRI profile of Alzheimer's disease (AD) patients to identify the different stages of the disease. The current criteria for diagnosing AD require the presence of relevant cognitive deficits, so the underlying neuropathological damage is important by the time the diagnosis is made. With the purpose of establishing new biomarkers to detect AD in its early stages, we evaluated a set of 2D and 3D texture features extracted from MRI scans of the hippocampus of patients with advanced AD, early mild cognitive impairment and cognitive normality. Many 3D texture parameters resulted to be statistically significant to differentiate between AD patients and subjects from the other two populations. When combining these 3D parameters with machine learning techniques, high accuracy was obtained, thus suggesting that texture analysis could at least help identify the presence of AD.

In the fourth project, we attempted to characterize the heterogeneity patterns of ischemic stroke in structural MRI. In brain MRI of older individuals, some pathological processes present similar imaging characteristics, like in the case of stroke lesions and white matter hyperintensities (WMH) of diverse natures, thus hindering the study of cerebrovascular diseases by means of imaging. Given that stroke effects are present not only in the affected region, but also in unaffected tissue, we investigated the feasibility of 3D texture features from WMH, normal-appearing white matter and subcortical structures to differentiate individuals who had a lacunar or cortical stroke visible on conventional brain MRI (T1-weighted, T2-weighted and FLAIR images) from subjects who did not. Texture features were not useful to differentiate between post-acute cortical and lacunar strokes, but promising results were achieved for discerning between patients presenting an old stroke and normal-ageing patients who never had a stroke. These results suggest that texture features may help in the detection of stroke lesions.

This thesis presents four novel feasibility studies to help clinicians in the evaluation of different brain disorders by means of a radiomics approach based on texture analysis in conventional MRI. The results achieved highlight the potential of this practice for defining and characterizing brain lesions in a fast, reliable and non-invasive way.

**Keywords**: *radiomics, magnetic resonance imaging, texture analysis, machine learning, glioblastoma, brain metastasis, Alzheimer's disease, stroke*

# Resumen

En los últimos 20 años, los investigadores han intentado explotar la información proporcionada por las imágenes médicas a través del cálculo y evaluación de numerosos parámetros cuantitativos para ayudar a los clínicos con el diagnóstico y la valoración de muchas enfermedades. Esta práctica ha sido bautizada recientemente como *radiomics*, y su éxito reside principalmente en la calidad e informatividad de las imágenes y los correspondientes parámetros. El *análisis de texturas* proporciona una gran variedad de parámetros que permiten cuantificar la heterogeneidad característica de diferentes tejidos, especialmente cuando se obtienen a partir de *imagen por resonancia magnética* (IRM). Basándonos en estos hechos, decidimos estudiar las posibilidades de los parámetros texturales extraídos de IRM para caracterizar varios trastornos que afectan el cerebro humano. El potencial de las características de textura se analizó con varios enfoques de *aprendizaje automático*, usando diferentes clasificadores y métodos de selección de características para hallar el modelo óptimo para cada tarea específica de forma eficaz. En esta tesis, la metodología radiomics implementada se usó para realizar cuatro proyectos independientes relacionados con cuatro desafíos clínicos distintos.

En el primer proyecto, estudiamos la diferenciación entre glioblastomas multiformes (GBMs) y metástasis cerebrales (MCs) en IRM convencional. En ocasiones, estos tipos de tumores cerebrales pueden confundirse al diagnosticarse, ya que pueden presentar un perfil radiológico similar y los datos clínicos pueden no ser concluyentes. En estos casos, el diagnóstico definitivo se debe realizar mediante un análisis histopatológico. Con el fin de evitar procedimientos exhaustivos e invasivos, estudiamos el poder discriminatorio de una gran cantidad de características de textura 2D extraídas de imágenes de referencia ponderadas en T1 filtradas y sin filtrar. Los resultados sugieren que las características de textura proporcionan información sobre la heterogeneidad de los GBMs y las MCs que puede ser de ayuda para distinguir con precisión ambas lesiones cuando se utiliza un enfoque de aprendizaje automático adecuado.

En el segundo proyecto, analizamos la clasificación de las MCs según su origen primario en IRM de referencia. En un porcentaje de pacientes con cáncer, las MCs son diagnosticadas como la primera manifestación de un tumor primario desconocido. Estos pacientes son sometidos a evaluaciones exhaustivas y procedimientos invasivos para detectar el tumor primario, y algunas veces el origen permanece sin diagnosticar en el momento de la muerte. Con el fin de detectar el tumor primario de una forma no invasiva y más rápida, examinamos la capacidad del análisis de texturas 2D y 3D para diferenciar las MCs derivadas de los tumores primarios más propensos a metastatizar (cáncer de pulmón, cáncer de mama

y melanoma) en imágenes ponderadas en T1. Los resultados mostraron que se logra una alta precisión al usar un conjunto reducido de parámetros 3D para diferenciar MCs de cáncer de pulmón de MCs de cáncer de mama y melanoma, por lo que los parámetros texturales 3D sacados de IRM pueden ser útiles para diferenciar las MCs de diferentes cánceres primarios.

En el tercer proyecto, evaluamos las propiedades del hipocampo en la IRM para identificar las diferentes etapas de la enfermedad de Alzheimer (EA). Los criterios actuales para diagnosticar la EA requieren la presencia de déficits cognitivos severos, por lo que el daño neuropatológico ya es grave en el momento de su diagnóstico. Con la idea de establecer nuevos biomarcadores para detectar la EA en sus primeras etapas, evaluamos un conjunto de parámetros texturales 2D y 3D extraídos de IRM del hipocampo de pacientes con EA avanzada, deterioro cognitivo leve temprano y normalidad cognitiva. Muchos parámetros de textura 3D resultaron ser estadísticamente significativos para diferenciar entre pacientes con EA y sujetos de las otras dos poblaciones. Al combinar estos parámetros 3D con técnicas de aprendizaje automático, se obtuvo una alta precisión, lo que sugiere que el análisis de textura podría al menos ayudar a identificar la presencia de AD.

En el cuarto proyecto, intentamos caracterizar los patrones de heterogeneidad del ictus cerebral isquémico en la IRM estructural. En la IRM cerebral de individuos de edad avanzada, algunos procesos patológicos presentan características de imagen similares, como en el caso de las lesiones por ictus y las hiperintensidades de la sustancia blanca (HSBs) de diversos orígenes, lo que dificulta el estudio de estos procesos por medio de técnicas de imagen. Dado que los ictus tienen efecto no solo en la región afectada, sino también en tejido adyacente, decidimos estudiar la viabilidad de los parámetros de textura 3D extraídos de las HSBs, la sustancia blanca no afectada y las estructuras subcorticales para diferenciar individuos afectados por ictus lacunares o corticales visibles en IRM convencional (imágenes ponderadas en T1, ponderadas en T2 y FLAIR) de sujetos de avanzada edad sin ictus. Las características de textura no fueron útiles para diferenciar entre ictus corticales y lacunares, pero se lograron resultados prometedores para discernir entre pacientes que han sufrido un ictus y pacientes que nunca lo han sufrido. Estos resultados preliminares sugieren que las características de textura pueden ayudar en la detección de lesiones por ictus.

Esta tesis presenta cuatro estudios de viabilidad originales para ayudar en la evaluación de diferentes trastornos cerebrales mediante un enfoque radiomics basado en el análisis de texturas sobre IRM convencional. Los resultados logrados resaltan el potencial de esta práctica para caracterizar lesiones cerebrales de manera rápida, eficiente y no invasiva.

**Palabras clave:** *radiomics, imagen por resonancia magnética, análisis de texturas, aprendizaje automático, glioblastoma, metástasis cerebral, enfermedad de Alzheimer, ictus cerebral.*

# Resum

En els últims 20 anys, els investigadors han intentat explotar la informació proporcionada per les imatges mèdiques a través del càlcul i avaluació de nombrosos paràmetres quantitatius per ajudar els clínics amb el diagnòstic i la valoració de moltes malalties. Aquesta pràctica ha sigut batejada recentment com *radiomics*, i el seu èxit resideix principalment en la qualitat i informativitat de les imatges i els corresponents paràmetres. L'*anàlisi de textures* proporciona una gran varietat de paràmetres que permeten quantificar l'heterogeneïtat característica de diferents teixits, especialment quan s'obtenen a partir d'*imatge per ressonància magnètica* (IRM). Basant-nos en aquests fets, vam decidir estudiar les possibilitats dels paràmetres texturals extrets d'IRM per caracteritzar diversos trastorns que afecten el cervell humà. El potencial de les característiques de textura es va analitzar amb diversos mètodes d'*aprenentatge automàtic*, usant diferents classificadors i mètodes de selecció de característiques per trobar el model òptim per a cada tasca específica de forma eficaç. En aquesta tesi, la metodologia radiomics implementada es va emprar per realitzar quatre projectes independents relacionats amb quatre desafiaments clínics diferents.

En el primer projecte, vam estudiar la diferenciació entre glioblastomes multiformes (GBMs) i metàstasis cerebrals (MCs) en IRM convencional. En ocasions, aquests tipus de tumors cerebrals poden confondre's al diagnosticar-se ja que solen presentar un perfil radiològic similar i les dades clíniques poden no ser concloents. En aquests casos, el diagnòstic definitiu s'ha de realitzar mitjançant una anàlisi histopatològic. Per tal d'evitar procediments exhaustius i invasius, vam estudiar el poder discriminatori d'una gran quantitat de característiques de textura 2D extretes d'imatges de referència ponderades en T1 filtrades i sense filtrar. Els resultats suggereixen que els paràmetres texturals proporcionen informació sobre l'heterogeneïtat dels GBMs i les MCs que pot ser d'ajuda per distingir amb precisió ambdues lesions quan s'utilitza una aproximació d'aprenentatge automàtic adequada.

En el segon projecte, vam analitzar la classificació de MCs segons el seu origen primari en IRM de referència. En un percentatge de pacients amb càncer, les MCs són diagnosticades com la primera manifestació d'un tumor primari desconegut. Aquests pacients són sotmesos a avaluacions exhaustives i procediments invasius per detectar el tumor primari, i algunes vegades l'origen roman sense diagnosticar en el moment de la mort. Per tal de detectar el tumor primari d'una forma no invasiva i més ràpida, vam examinar la capacitat de l'anàlisi de textures 2D i 3D per diferenciar les MCs derivades dels tumors primaris més propensos a metastatitzar (càncer de pulmó, càncer de mama i melanoma) en imatges ponderades en T1. Els resultats van mostrar que s'aconsegueix una alta precisió quan s'utilitza un conjunt reduït de paràmetres 3D per diferenciar les MCs de càncer de pulmó de

les MCs de càncer de mama i melanoma, de manera que els paràmetres texturals 3D obtinguts de la IRM poden ser útils per a diferenciar les MCs de diferents càncers primaris.

En el tercer projecte, vam avaluar les propietats de l'hipocamp en la IRM per identificar les diferents etapes de la malaltia d'Alzheimer (MA). Els criteris actuals per diagnosticar la MA requereixen la presència de dèficits cognitius severs, de manera que el dany neuropatològic ja és greu en el moment del seu diagnòstic. Amb la idea d'establir nous biomarcadors per detectar la MA en les seues primeres etapes, vam avaluar un conjunt de paràmetres texturals 2D i 3D extrets d'IRM de l'hipocamp de pacients amb MA avançada, deteriorament cognitiu lleu i normalitat cognitiva. Molts paràmetres de textura 3D van resultar ser estadísticament significatius per diferenciar entre pacients amb MA i individus de les altres dues poblacions. En combinar aquests paràmetres 3D amb tècniques d'aprenentatge automàtic, es va obtenir una alta precisió, el que suggereix que l'anàlisi de textura podria almenys ajudar a identificar la presència de la MA.

En el quart projecte, vam intentar caracteritzar els patrons d'heterogeneïtat de l'ictus cerebral isquèmic en la IRM estructural. En la IRM cerebral d'individus d'edat avançada, alguns processos patològics presenten característiques d'imatge similars, com en el cas de les lesions per ictus i les hiperintensitats de la substància blanca (HSBs) de diversos orígens, dificultant així l'estudi d'aquests processos per mitjà de tècniques d'imatge. Atès que els ictus tenen efecte no només a la regió afectada, sinó també en teixit adjacent, vam decidir estudiar la viabilitat dels paràmetres de textura 3D extrets de les HSBs, la substància blanca no afectada i les estructures subcorticals per diferenciar individus afectats per ictus llacunars o corticals visibles en IRM convencional (imatges ponderades en T1, ponderades en T2 i FLAIR) d'individus sense ictus. Els paràmetres de textura no van ser útils per diferenciar entre ictus corticals i llacunars, però es van aconseguir resultats prometedors per discernir entre pacients que han patit ictus i pacients que mai n'han patit. Aquests resultats preliminars suggereixen que les característiques de textura poden ajudar en la detecció de lesions per ictus.

Aquesta tesi presenta quatre estudis de viabilitat originals per ajudar els clínics en l'avaluació de diferents trastorns cerebrals mitjançant una aproximació radiomics basada en l'anàlisi de textures sobre IRM convencional. Els resultats obtinguts destaquen el potencial d'aquesta pràctica per definir i caracteritzar lesions cerebrals de manera ràpida, eficient i no invasiva.

**Paraules clau:** *radiomics, imatge per ressonància magnètica, anàlisi de textures, aprenentatge automàtic, glioblastoma, metàstasi cerebral, malaltia d'Alzheimer, ictus cerebral.*

# Abbreviations and Acronyms

***Brain and Diseases***

| | |
|---|---|
| **AD** | Alzheimer's disease |
| **ADNI** | Alzheimer's Disease Neuroimaging Initiative |
| **BM** | brain metastasis |
| **CN** | cognitively normal, cognitive normality |
| **CNS** | central nervous system |
| **CSF** | cerebrospinal fluid |
| **EMCI** | early mild cognitive impairment |
| **GBM** | glioblastoma multiforme |
| **MCI** | mild cognitive impairment |
| **NAWM** | normal appearing white matter |
| **PNS** | peripheral nervous system |
| **SRS** | stereotactic radiosurgery |
| **SS** | subcortical structures |
| **SVD** | small vessel disease |
| **WBRT** | whole-brain radiotherapy |
| **WHO** | World Health Organization |
| **WMH** | white matter hyperintensities |

*Medical Imaging*

| | |
|---|---|
| **CAD** | computer-aided detection and diagnosis |
| **CT** | computed tomography |
| **FID** | free induction decay |
| **FLAIR** | fluid-attenuated inversion recovery |
| **IR-SPGR** | inversion recovery spoiled gradient-echo |
| **MP-RAGE** | magnetization prepared rapid gradient echo |
| **MR** | magnetic resonance |
| **MRI** | magnetic resonance imaging |
| **PET** | positron emission tomography |
| **SENSE** | sensitivity encoding |
| **T1** | longitudinal relaxation time |
| **T1W** | T1-weighted |
| **T2** | transverse relaxation time |
| **T2W** | T2-weighted |
| **TE** | echo time or time-to-echo |
| **TR** | repetition time |
| **TSE** | turbo spin echo |

*Texture Analysis*

| | |
|---|---|
| **2D** | two dimensions, two-dimensional or bi-dimensional |
| **3D** | three dimensions or three-dimensional |
| **DWT** | discrete wavelet transform |
| **GLCM** | gray-level co-occurrence matrix |
| **GLRLM** | gray-level run-length matrix |

| | |
|---|---|
| **GLSZM** | gray-level size zone matrix |
| **HH** | high-high (diagonal details of the DWT decomposition) |
| **HL** | high-low (horizontal details of the DWT decomposition) |
| **ISBI** | Image Biomarker Standardisation Initiative |
| **LBP** | local binary patterns |
| **LH** | low-high (vertical details of the DWT decomposition) |
| **LL** | low-low (approximations of the DWT decomposition) |
| **NGL** | number of gray levels |
| **NGTDM** | neighborhood gray-tone difference matrix |
| **ROI** | region of interest |
| **SNR** | signal-to-noise ratio |
| **VOI** | volume of interest |
| **WCF** | wavelet co-occurrence features |
| **WSF** | wavelet statistical features |

***Machine Learning and Statistics***

| | |
|---|---|
| **ANN** | artificial neural network |
| **ANOVA** | analysis of variance |
| **AUC** | area under the receiver operating characteristics curve |
| **BH** | Benjamini-Hochberg procedure |
| **CM** | confusion matrix |
| **CV** | cross-validation |
| **FN** | false negatives |
| **FP** | false positives |
| **ICC** | intraclass correlation coefficient |

| | |
|---|---|
| **KNN** | k-nearest neighbors |
| **LDA** | linear discriminant analysis |
| **LGOCV** | leave-group-out cross-validation |
| **LOOCV** | leave-one-out cross-validation |
| **MDA** | mean decrease accuracy |
| **MDI** | mean decrease in impurity |
| **MIC** | maximal information coefficient |
| **MLP** | multilayer perceptron |
| **MWW** | Mann-Whitney-Wilcoxon test |
| **NB** | naive Bayes |
| **PCA** | principal component analysis |
| **RF** | random forest |
| **ROC** | receiver operating characteristics |
| **SD** | standard deviation |
| **SVM** | support vector machine |
| **TN** | true negatives |
| **TP** | true positives |

# Contents

*“There’s magic in fighting battles beyond endurance.*
*It’s the magic of risking everything for a dream that nobody sees but you”*

Million Dollar Baby, 2004

# Chapter 1.
# Introduction

## 1.1. The concept of "radiomics"

Medical images may possess abundant unexplored valuable information that could be used to assess some diseases since this information may reflect several pathophysiologic aspects of the tissue under examination. To analyze, process and take advantage of this information, a new promising field has arisen in the past decade: *radiomics* [1]. Radiomics is a new concept that comprises several independent research fields with the common aim of extracting from processed medical images as many quantitative features as possible, related to texture, color or shape. Hence, medical images are converted into mineable high-dimensional data with the purpose of increasing the power of decision support tools by combining these data with other clinical characteristics. Radiomics analysis has been proved to be a valuable source of information to improve the precision in diagnosis, to assess the prognosis or to predict treatment response, mainly in cancer research but also applicable to other diseases [1]–[5].

The motivation behind radiomics analysis is based on the hypothesis that medical images contain information at the tissue/organ level (macroscopic level) that reflects the underlying pathophysiology of the tissue (microscopic level) and that these relationships can be revealed by means of quantitative imaging features [1], [6]. Furthermore, these features can be statistically combined or correlated with other data of diverse nature such as clinical or genomic in order to define the disease more accurately.

Radiomics may appear to be a simple and straightforward practice. However, radiomics analyses involve several individual processes, each with its own challenges, that have been largely studied independently and that have been joined together to pursue a common goal. Radiomics practice include: (a) image acquisition, (b) identification,

segmentation and pre-processing of the regions (2D) or volumes (3D) of interest (i.e., those that contain tissue with possible valuable information), (c) extraction of descriptive features from the regions or volumes, and (d) mining of these data to develop classification models to predict outcomes either alone or in combination with additional information, such as demographic, clinical, histologic or genomic data [7]–[10]. Figure 1.1 shows a brief diagram of the steps included in the radiomics pipeline. Despite all the challenging processes included in the radiomics practice, the focus of interest of radiomics is the extraction of features that describe quantitatively the image region under analysis. For this purpose, *texture analysis* has been proved to be an excellent source of imaging biomarkers [11]. Texture analysis describes a wide range of techniques that enable the quantification of pixel interrelationships, gray level patterns, and spectral properties of an image. These techniques allow computing features that provide a measure of intralesional heterogeneity or standing out areas that exhibit different textural patterns, which are beyond human visual perception [12].

Radiomics can be applied to different/multiple imaging modalities, and the selection of the appropriate technique to investigate each disease or lesion depends on several factors, including the resolution of the images and the tissue under analysis. However, in the last years, *magnetic resonance imaging* (MRI) has become popular in radiomics studies due to its growing availability in the clinical routine and the resulting high-quality images that offer excellent anatomic details thanks to new advances in technology [13].

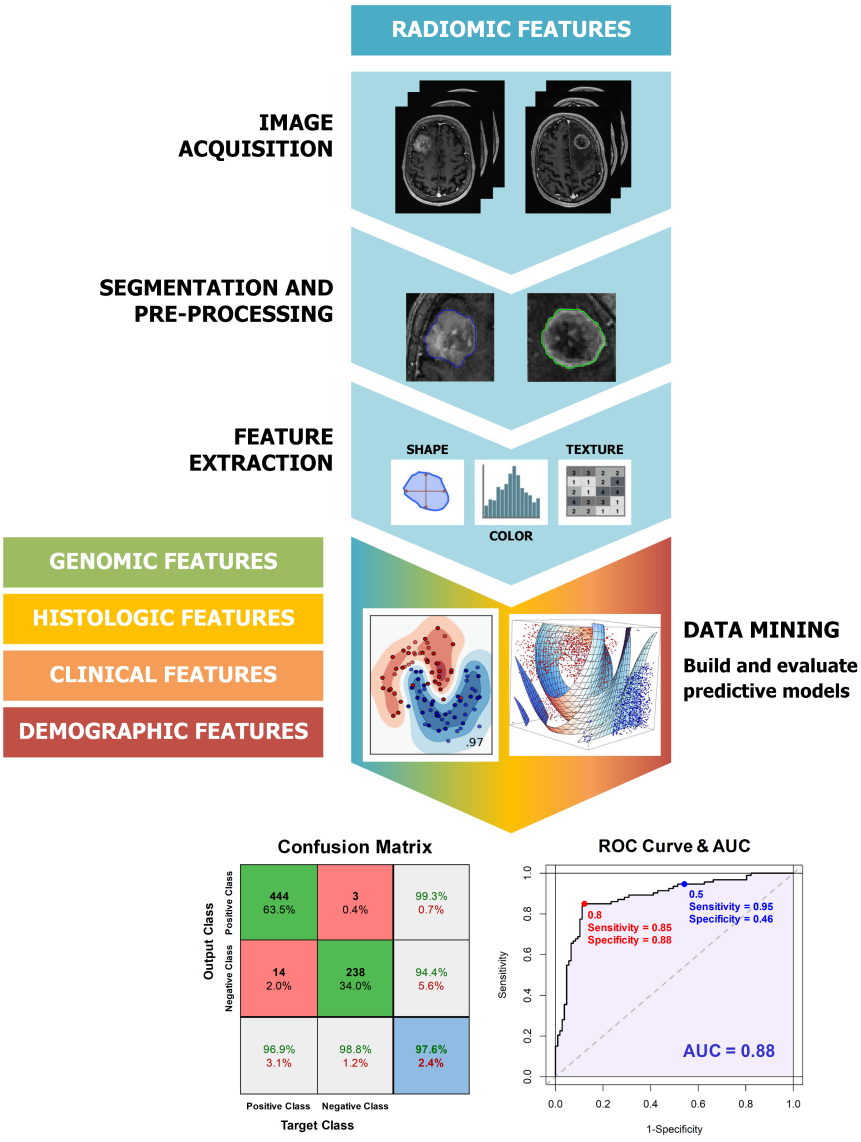



**Figure 1.1.** Radiomics practice overview showing the major steps: image acquisition, segmentation and pre-processing of the images, feature extraction, and data mining. The derived radiomics models and results are evaluated and interpreted with statistical metrics and graphical representations. It is important to mention that radiomics data may be combined or correlated with genomic, histologic, clinical or demographic data to improve precision medicine.

## 1.2. Objectives

The general objective of the present thesis consists on analyzing several common neurological diseases and lesions with a radiomics pipeline focused on texture analysis applied to magnetic resonance imaging. With this radiomics approach, we intend to provide clinicians with a complementary decision support tool to help in giving a prompt response to some medical challenges related to brain pathologies that usually are solved too late (in some cases in the advanced stages of the disease or even after the patient's death) by means of invasive procedures, exhaustive neuroimaging or neuropsychological tests.

In particular, four specific objectives related to four different current medical challenges are established in the context of this project:

1. Differentiation between glioblastomas multiforme (GBM) and brain metastases (BM) in conventional structural MRI.
2. Identification of the primary site of origin in patients with brain metastases (BM) from an unknown primary cancer in initial detection structural MRI.
3. Detection of the presence of Alzheimer's disease (AD) and its early diagnosis in structural MRI of the hippocampal region.
4. Characterization of different types of lesions related to ischemic stroke in structural MRI of elderly patients

In addition to trying to give a solution to these current diagnostic problems, we evaluate several aspects of the radiomics practice that may affect the final results:

- The benefits and detriments of examining texture features extracted from three-dimensional regions instead of bi-dimensional regions.
- The influence of the size of the regions of interest and the quantization of the image regions (i.e., reducing the number of gray levels of the image).
- The ability of several texture analysis methods to offer effective parameters.
- The performance of different machine learning approaches.

## 1.3. Contributions to Knowledge

This thesis offers four novel contributions for the assessment of patients with different neurological diseases or lesions by the study of texture analysis in conventional structural MRI combined with machine learning techniques, in the context of the radiomics practice.

The first contribution is that 2D texture features extracted from structural MRI are useful for classifying GBM and BM with a good level of accuracy when applying a machine learning scheme. These preliminary results indicate that texture analysis could be used by radiologists in the near future to help in the correct diagnosis of each type of brain tumor in its early stage, in particular in those cases when the radiological or clinical basic information is not conclusive. Moreover, with further research, texture analysis on MRI could be used as a faster alternative to biopsies in order to achieve a reliable definitive diagnostic without resorting to invasive procedures.

The second contribution is that texture analysis on MRI showed a promising capacity for identifying the primary site of origin in those patients with BM where the primary cancer is still unknown. Our results show that BM derived from lung cancer can be clearly distinguished from BM from breast cancer and melanoma. These results should be further investigated so as to validate texture features as new biomarkers of BM since BM from breast cancer and melanoma could not be differentiated and other primary sites of origins should be included. Patients presenting BM as a first manifestation of a primary tumor could avoid additional imaging or invasive procedures if texture analysis applied to MRI were confirmed as a reliable definitive method to identify the origin of the BM.

The third contribution is that AD presence can be detected with texture analysis on MRI of the hippocampal region, but an early diagnosis of the disease cannot be achieved with the texture features employed in our project, since control subjects could not be distinguished from patients suffering from early mild cognitive impairment (EMCI). These preliminary results should encourage researchers to further investigate this disease by means of texture analysis, since it is a fast and non-subjective way for assessing the AD that could help in the near future in the early detection of the disease.

The fourth and final contribution is that texture features can capture differences between structural MRI scans of normal-appearing tissue of older patients presenting different pathological brain processes related to stroke. Cortical and lacunar strokes were not clearly differentiated with the proposed machine learning approach. On the contrary, the results obtained for classifying elderly patients with an old stroke and without evidences of a stroke were very promising, thus suggesting that stroke effects can be detected in terms of heterogeneity patterns in those cases where other pathologies may be considered.

Apart from the above-mentioned contributions to the medical field, in this thesis other relevant technical conclusions are reached in the context of the radiomics practice that may be beneficial for future radiomics analysis and that confirm the conclusions presented in other related studies. In first place, 3D texture analysis improves the classification performance achieved by 2D texture analysis, but when volumetric analyses are not possible due to the resolution of the images or the complexity of the volume definition, texture analysis applied on single MRI scans can also offer remarkable results in terms of accuracy. In second place, quantization of image regions is a necessary process to reduce the computational cost and improve the signal-to-noise (SNR) ratio, but the proper number of gray levels used to quantize the images should be analyzed since different levels of gray may produce different classification results with statistical significance. Finally, the size of the regions of interest plays an important role in texture analysis because too small regions may not capture sufficient heterogeneity information.

## 1.4. Thesis Structure

This thesis is structured in 11 chapters. Chapter 1 presents a summary of the general and specific objectives of the thesis and the novel contributions to knowledge. Chapters 2 to 4 present the theoretical background that is essential for understanding the experimental studies. Chapters 5 to 8 present the experimental projects performed. Chapters 9 to 11 present the final overall conclusions of this thesis, the bibliography and the publications issued in different journals and conferences in the context of this thesis.

A summary of the important chapters of this thesis, which are self-contained and can be read independently, is introduced below:

**Chapter 2: Brain Magnetic Resonance Imaging**

This chapter gives a background on the principles of conventional MRI, with a focus on the techniques used to assess patients with neurological pathologies. The chapter begins with a summary of the physiological principles of the brain and the common diseases or lesions that can affect it, followed by an introduction of the general principles of MRI physics, and finished by an overview of MRI applied to the brain.

**Chapter 3: Texture Analysis**

This chapter gives a description of texture analysis and explains the process to follow in order to perform texture analysis on medical images. The factors to consider before performing texture analysis are also presented, since the texture outcome can be considerably affected depending on the processing of the regions of interest. The texture analysis methods that were used in the experimental studies are also described. Finally, a brief review of texture analysis applied to MRI and some additional important considerations are also provided.

**Chapter 4: Data Analysis with Machine Learning**

This chapter presents a brief overview of data exploration and machine learning, focusing on the predictive models used in this thesis, the importance of feature selection and the application of resampling techniques to enhance the classification results. It also provides a summary regarding the evaluation of model performance.

**Chapter 5: Differentiation between Brain Metastases and Glioblastomas**

This chapter shows the experimental study that aimed to classify glioblastomas multiforme (GBM) from brain metastases (BM) with 2D texture analysis in contrast-enhanced structural MRI. Texture features extracted from the original images and from filtered images were analyzed and compared within a nested cross-validation scheme that included five predictive models and three feature selection methods to compare. The influence of the quantization of the images was also evaluated.

**Chapter 6: Classification of Brain Metastases by their Primary Site of Origin**

This chapter presents the project developed to identify the primary site of origin in patients with BM from an unknown primary cancer. In this work, 2D and 3D texture analyses extracted from structural MRI were compared, the influence of the quantization process was assessed and the classification performance of several predictive models and feature selection techniques was evaluated with a multiclass and a one-versus-one approaches within a nested cross-validation scheme.

**Chapter 7: Evaluation of New Biomarkers for Alzheimer's disease**

This chapter shows the study carried out to identify the presence of Alzheimer's disease and to assess its early diagnosis in conventional MRI. Global, regional and local texture parameters extracted from 2D and 3D regions situated in the hippocampal region were first evaluated individually with statistical analyses and then they were analyzed with a nested cross-validation structure. Three predictive models were evaluated and the influence of the image quantization in the classification results was also studied.

**Chapter 8: Characterization of Ischemic Stroke**

This chapter presents the project elaborated to characterize the heterogeneity properties that describe the structural MRI of different stroke patients. In this project five groups of 3D textures extracted from different brain tissues were evaluated statistically and with a cross-validation machine learning approach based on two known classifiers. The effect of the feature selection and the influence of age were also analyzed.

**Chapter 9: Final Conclusions**

This chapter presents the final overall conclusions reached throughout the thesis and a deep reflection of the applicability and suitability of the work carried out.

# Chapter 2.
# Brain magnetic resonance imaging

## 2.1. The Human Brain

The nervous system is considered the most complex product of biological evolution. The constantly changing activity patterns of its billions of interactive units represent the fundamental physical basis of each of the aspects of human behavior. The nervous system is divided in two major subsystems: the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS is the structural and functional center of the nervous system and it mainly comprises the *brain* and the *spinal cord*, surrounded by the cerebrospinal fluid (CSF). On the contrary, the PNS consists of the cranial and spinal nerves that connect the CNS with the other parts of the body [14].

The brain is the command center for the human nervous system. It receives input from the sensory organs and sends output to the muscles. In other words, the human brain interprets the information that we receive from the outside world and is in charge of the control of all body functions. The human brain has the same basic structure as other mammal brains but is larger in relation to body size than any other brains. In particular, it weighs about 1400 grams and constitutes about 2% of the total body weight of an average adult [15]. The brain is usually divided in four structures, as shown in Figure 2.1 [16]:

- ***Brain stem:*** connects the brain with the spinal cord and is composed of the *medulla oblongata*, the *pons* and the *midbrain*. This structure works as a bidirectional conduction path where the sensory fibers transmit the impulses between the spinal cord and the rest of the brain, among other functions.

- ***Cerebellum:*** is the second largest part of the brain (often referred as "little brain") and controls several subconscious activities such as balance and muscular coordination or motor learning.
- ***Diencephalon:*** lies above the brain stem and includes the *thalamus*, *hypothalamus* and *pineal gland*. These three substructures are responsible for interpreting and regulating emotions and sensations of sound, smell, taste, touch, pain or temperature and for controlling some body functions like heartbeat or digestion, among other functions.
- ***Cerebrum:*** is the largest part of the human brain, accounting for the 85% of the brain's weight and controls higher brain functions such as language, logic, decision-making and creativity. It is divided into two almost symmetrical *hemispheres* (left and right) connected by a bundle of nerve fibers called corpus callosum. The left brain controls all the muscles on the right-hand side of the body and the right brain controls the left side. In the following paragraphs we will focus on this structure to explain the specific anatomy and functions of the human brain.

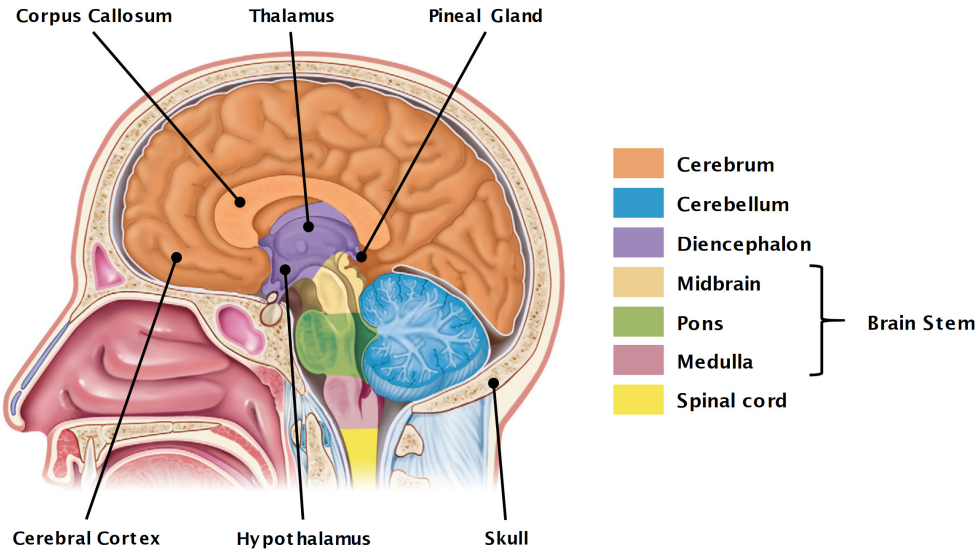



**Figure 2.1.** Major regions of the central nervous system (CNS) viewed in a sagittal plane. Adapted from [16].

Each of the two hemispheres that form the cerebrum can be divided in three basic regions (Figure 2.2). The outer layer of the cerebrum is the *cerebral cortex*, which is a thin layer of *gray matter* mainly composed of a large number of neuronal and glial cell bodies. The cerebral cortex is where our conscious mind is found and plays a key role in memory, perception, cognition, attention, awareness, thought and language. The ridges in the surface of cortex are called *gyri* while the indentations are called *sulci* (or *fissures* if they are deeper). The inner part of the cerebrum is known as *white matter* and is mainly composed of bundles of nerve fibers (tracts). The white matter is responsible for coordinating communication between different brain regions. Finally, within this white matter, there are a few islands of gray matter known as the *basal ganglia* (or basal nuclei), that is in charge of producing automatic movements and postures [17].

Focusing on the cerebral cortex, this region can be divided into sections called *lobes*. This division is useful from an anatomical, functional, and pathophysiological perspective. The most common division consists of four different lobes separated by three deep sulci (*central*, *parieto-occipital* and *lateral sulcus*) and named according to the four skull bones protecting them, as shown in Figure 2.2 [18]:

- ***Frontal lobe:*** is located at the front of the brain and is involved in reasoning, planning, organizing, selective attention and a variety of higher cognitive functions. At the back of this lobe lies the motor cortex, which receives information from various lobes of the brain and utilizes them to generate voluntary body movements.
- ***Parietal lobe:*** is located in the middle section of the brain and integrates sensory information from various parts of the body. It contains the somatosensory cortex, which is essential for processing body's senses like tactile sensory information (touch, pressure, pain, etc.) as well as for spatial orientation and navigation.
- ***Occipital lobe:*** is located at the back of the brain and includes the primary visual cortex, so it is the main region for interpreting visual stimuli and information.
- ***Temporal lobe:*** is located on the bottom section of the brain and contains the primary auditory cortex, which is important for assigning meaning to the sounds we hear. The hippocampus and the amygdala are also located in this lobe, so this lobe is also associated to the formation of memories and emotions, respectively.

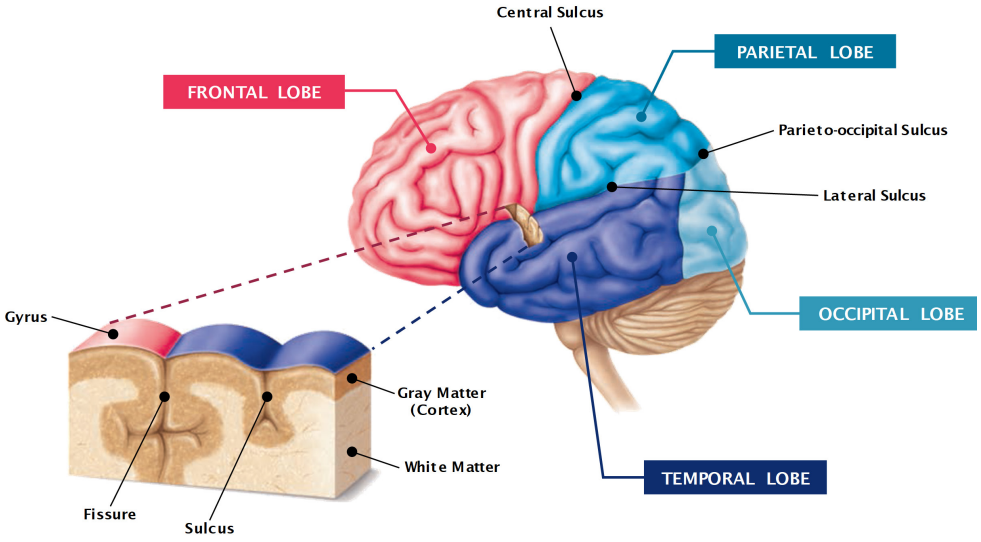



**Figure 2.2.** Lateral view of the left hemisphere with the cerebrum divided into four lobes separated by the three main sulci. The zoomed region shows the two types of tissue found in the cerebrum (gray and white matter). Adapted from [17].

It is important to know that each lobe of the brain does not function alone. There are very complex relationships between the lobes of the brain and between the right and left hemispheres. Additionally, the symptoms associated to some brain disorders may be related with the functional areas described by the lobes, thus allowing clinicians to detect the approximate location of the lesion or disease. For example, damage to the frontal lobe can lead to changes in socialization and attention, damage to the parietal lobe may result in problems for interpreting sensory stimuli, damage to the occipital lobe can cause visual problems and damage to the temporal lobe can produce problems with memory or speech perception.

## 2.2. Brain Disorders

A brain lesion or disease defines any damage caused to any part of the brain, producing inflammation, malfunction or destruction of brain cells or brain tissue. Classifying all the disorders that can affect the human brain may somewhat be a complex task since they can be categorized according to the primary location affected, the primary type of dysfunction involved or the primary type of cause. According to the classification

of brain, spinal cord, and nerve disorders included in *The Merck manual of diagnosis and therapy* [19], the following groups of disorders affecting the brain can be highlighted:

- *Brain dysfunctions*: such as agnosia, amnesia or apraxia
- *Brain infections*: such as abscesses, encephalitis or meningitis
- *Headaches*: such as migraines
- *Sleep disorders*: such as narcolepsy or parasomnias
- *Dementia and delirium*: such as Alzheimer's disease or Lewy body disorder
- *Movement disorders*: such as Huntington's disease or Parkinson's disease
- *Demyelinating diseases*: such as multiple sclerosis
- *Seizure disorders*: such as epilepsy
- *Cerebrovascular conditions*: such as ischemic or hemorrhagic stroke
- *Brain tumors*: such as primary and secondary brain tumors

This organization may vary among different literature, and several disorders that were separated in this classification may be included in the same group according to other patterns of classification. For example, Alzheimer's, Parkinson's and Huntington diseases (among others) are usually considered together as neurodegenerative diseases since they cause a deterioration to the brain and nerves over time. In the following subsections, we will briefly discuss the brain lesions and diseases treated in this thesis.

### *2.2.1. Brain Tumors*

In the human body, cells are constantly growing and dividing as a consequence of the body development or in order to replace dead cells of damaged tissues. The process of cell division is regulated by a sequence of control mechanisms that tells the cells when to start the division and when to remain static. However, some cells stop responding appropriately to these control signals and they start to grow and reproduce in an uncontrolled manner, invading normal tissues and organs and eventually spreading throughout the body (i.e., metastasis). This unregulated proliferation of cells generates abnormal tissues known as *tumor* or *neoplasm*. A brain tumor refers to the formation of abnormal tissues within the brain. According to the origin of the tumor, brain tumors can be categorized in primary and secondary brain tumors. Whereas primary brain tumors are originated in the brain, secondary brain tumors (also known as metastatic brain tumors or *brain metastases)* are originated due to a primary cancer located in other part of the body that spreads to the brain via the circulatory or lymphatic systems [19].

**Primary brain tumors**

Primary brain tumors can be benign or malignant (i.e., cancerous). Benign tumors have clearly defined borders, have a slow growth and remain confined to its original location neither invading surrounding normal tissue nor spreading to distant body sites. These brain tumors may be surgically removed since they tend to be well delimited and usually are not deeply rooted in the brain. Therefore, benign brain tumors usually do not represent a major life-threatening problem, but depending on their size and location, their management may be more or less complicated and dangerous. On the contrary, malignant tumors have usually irregular borders, typically grow faster than benign tumors and invade surrounding tissues in an aggressive way. Although brain malignant tumors rarely spread to other organs, they may metastasize to other parts of the brain or CNS, and additionally, they are more likely to recur after surgical resection than benign brain tumors.

According to the 2016 World Health Organization (WHO) classification of the tumors of the central nervous system, there exist over 120 histological types of primary brain tumors [20]. The WHO classifies the tumors in four grades depending on the way that tumor cells look under the microscope. This classification is indicative of the aggressiveness and severity of the brain tumor. For example, WHO grade I brain tumors encompasses least malignant tumors with slow cell growth, while WHO grade IV tumors comprises tumors with a very abnormal appearance where cells reproduce rapidly and where necrosis areas (dead cells) are often present.

Benign and malignant brain tumors are commonly named according to the cells that proliferate or the tissue in which they originate. For example, tumors that originate due to an uncontrolled reproduction of glial cells are called *gliomas*, and tumors initiated in the meninges are called *meningiomas*. In fact, according to recent reports of the incidence of brain and CNS tumors [21], meningiomas are the most common primary tumors, accounting for the 36.8% of all primary tumors, and gliomas are the most frequent malignant primary tumors, representing about the 80% of malignant tumors.

**Secondary brain tumors**

Secondary or metastatic brain tumors are always malignant tumors (i.e., cancerous) since they appear as a result of a proliferation of a cancer generated in other part of the body. Their exact incidence is certainly unknown, but they are estimated to be about ten times more common than primary tumors, they may occur in 9–17% of patients with cancer, and 35–50% of patients with brain metastases presents at least three

metastases. The primary tumors that metastasize more frequently to the brain are those originated in lung (≥ 50%), breast (15–25%) and skin (melanoma) (5–20%) [22]–[24].

### *2.2.2. Dementia and Alzheimer's disease*

The term *dementia* describes a set of neurodegenerative symptoms that mainly affects memory, attention, intellectual capacity and personality, among other human mental functions. Dementia englobes a variety of diseases that occur due to physical changes in the brain based on a degeneration or destruction of brain tissue and nerves. Diseases associated to dementia are characterized by a progressive brain functions deterioration over time that begins gradually with an uncertain beginning point and is usually irreversible. The rapid or slow progression of dementia depends greatly on the underlying cause of the dementia [19].

Dementia primarily affects elderly people, especially people older than 65 years old, and constitutes a major health problem worldwide. In 2011, it was estimated that 35.6 million people around the world suffered from dementia, whereas, in 2018, this figure already amounted to 50 million. At this rate, it is expected that this figure will almost double by 2030 and more than treble by 2050 [25], [26]. *Alzheimer's Disease* (AD) represents the most common type of dementia, accounting for an estimated 60 to 80 percent of cases [27], [28]. Other common types of dementia are vascular dementia, Lewy body dementia or frontotemporal dementia.

**Alzheimer's disease**

Alzheimer's disease is a neurodegenerative disease defined by the presence of an intellectual and behavioral deterioration of sudden onset and progressive course that appears during adulthood. It is estimated that one out of ten people aged 65 and older suffers from AD. This disease involves a continuous brain degradation that is broadly characterized by a preclinical stage, followed by a phase of mild cognitive impairment (MCI), and an final phase of dementia in the strict sense [27], [29]. The differences between typical age-related cognitive changes and signs of AD can be subtle. The most common initial symptom is a gradually worsening ability to remember new information. This occurs because the first neurons to be damaged and destroyed are usually located in brain regions involved in forming new memories. As neurons in other parts of the brain are damaged and destroyed, individuals experience other difficulties, including neurobehavioral symptoms such as agitation, sleeplessness and delusions [27].

The etiology of AD is certainly unknown although it is thought to probably have a multifactorial cause. The main neuropathological changes of AD are centered on the loss of neuronal synapses and neuronal death due to the abnormal aggregation of the proteins involved in the disease: intra-neuronal hyperphosphorylated tau in the form of neurofibrillary tangles (tau tangles) and extra-neuronal beta-amyloid in the form of senile plaques (beta-amyloid plaques). Beta-amyloid plaques are believed to contribute to neuronal death by interfering with neuron-to-neuron communication at synapses, while tau tangles block the transport of nutrients and other essential molecules inside neurons [27].

The diagnosis of AD remains nowadays fundamentally clinical, which means that it cannot be diagnosed until the first symptoms appear, or even later, because, as mentioned before, these early symptoms are usually associated with consequences due to normal aging [25]. Definitive diagnosis can only be made with histopathological confirmation of beta-amyloid plaques and tau tangles, usually at autopsy [30]. However, experimental evidence indicates that pathophysiological alterations take place in the brain more than a decade before clinical decline, in the pre-symptomatic phase known as preclinical stage [31], [32].

### *2.2.3. Cerebrovascular conditions*

Cerebrovascular diseases involve a variety of medical conditions that affect the blood vessels of the brain and the associated cerebral circulation. The most common representation of a cerebrovascular disease is the acute *stroke*, which occurs when part of the blood flow that circulates through the network of cerebral arteries and veins decreases drastically or stops, thus resulting in the death of an area of brain tissue. Strokes can be grouped into two types: *ischemic* and *hemorrhagic* strokes. Ischemic strokes occur when there is a lack of blood supply to the brain mainly due to a blood clot blocking a blood vessel, while hemorrhagic strokes occur when blood vessels are abnormal or weak and suddenly rupture, causing blood to leak into the brain and generating congestion and pressure on brain tissue [19]. Ischemic strokes account for roughly the 85% of all strokes and are a leading cause of mortality and disability, being the second most common cause of death worldwide [33], [34]. Likewise, ischemic strokes can be subtyped. The most common ischemic strokes are lacunar strokes, which are small infarcts (2-20 mm in diameter) resulting from the occlusion of one of the small perforating arteries that provide blood to the deep subcortical structures of the brain. They account for a quarter of all ischemic strokes and a fifth of all strokes [35].

The signs and symptoms of a stroke occur suddenly, and they may vary depending on the precise location of the blockage or bleeding in the brain and how much cerebral tissue is affected. For example, if an artery supplying the area of the brain that controls the left leg's muscle movements is blocked, the leg becomes weak or paralyzed and if the area of the brain that senses touch in the right arm is damaged, sensation in the right arm is lost [19]. However, one can have a stroke without noticing any outward stroke symptoms and the diagnose of this event is made when permanent damage is found on neuroimaging evaluations such as MRI. These cerebrovascular accidents are known as *silent strokes* and they are associated with an increased risk of subsequent stroke and cognitive dysfunction [36].

One of the brain conditions that directly relate to the causes of most of strokes is the cerebral small vessel disease (SVD). The SVD refers to a range of pathological processes affecting the small arteries, arterioles, capillaries and small veins of the brain, that can be described by clinical and imaging findings. This condition accounts for about 20% of all strokes worldwide and constitutes a major source of cognitive decline [37].

## 2.3. Overview of magnetic resonance imaging

The use of magnetic resonance imaging (MRI) for clinical and scientific purposes is relatively recent. In 1946, Felix Bloch [38] and Edward Purcell [39] established the principles of the nuclear magnetic resonance (MR). However, it was not until 1971 when Raymond Damadian reported that tumors and normal tissue could be distinguished in vivo by nuclear MR [40]. In 1973, Paul Lauterbur published the first true MR image [41] and thanks to this and further achievements, he was awarded with the 2003 Nobel Prize in Physiology or Medicine for his discoveries concerning MRI as a diagnostic tool, along with Sir Peter Mansfield, who developed the echo-planar imaging technique [42].

Currently, the role of MRI in the field of medicine is constantly expanding thanks to the rapid evolution and improvement of the functionality of MRI systems. Its clinical efficacy and its benefits over other ionizing imaging techniques such as computed tomography (CT) or positron emission tomography (PET) have been demonstrated in numerous studies. Moreover, the possibility of selecting the scan plane, acquiring true 3D images and generating excellent soft-tissue contrast, makes MRI the best choice for many clinical applications [43].

### 2.3.1. The Physics behind MRI

All matter is made up of atoms, and these in turn are composed of particles. These particles have intrinsic characteristics such as *spin*, which is defined as the property that describes the fact that a particle is rotating around an imaginary axis. The operation of an MR scanner is based on the analysis of the interaction between the particles that have this property and a certain magnetic field. Normally, MRI focuses on the measurement of hydrogen nuclei, mainly because this element is the most abundant in the human body (it is present in water and fat) and has one of nature's greatest response to the presence of a magnetic field. The hydrogen nucleus has a unique moving proton with positive charge that generates a magnetic field characterized by its magnetic moment. In addition, the hydrogen nucleus is characterized by an angular momentum because it has an odd number of protons [44].

In a certain volume of tissue formed by a set of equal protons, all the magnetic moments are oriented in arbitrary directions and the net magnetization of the material $M$ is null. When the protons of the hydrogen nuclei interact with an external magnetic field $B_0$, these particles acquire a gyroscopic movement, called *precession*, around the axis determined by $B_0$, whose frequency is proportional to the intensity of the external field $B_0$. In addition, the magnetic moments will tend to align with the direction of $B_0$, either parallel (high energy) or antiparallel (low energy). This process known as *polarization* will induce the appearance of a net magnetization $M$ parallel to the applied field, proportional to the difference between parallel and antiparallel moments [43], [44].

If we now disturb the created system by applying another field $B_1$ in the transverse plane, known as a radiofrequency pulse, the net magnetization $M$ will tend to align with the total field $B = B_0 + B_1$. This is the process of *excitation*. After the application of an RF pulse, the magnetization $M$ gradually returns to its equilibrium state, thus defining the process of *relaxation*. At a macroscopic level, relaxation is the process by which protons release the energy absorbed during the emission of radiofrequency pulses. There are two types of relaxation:

- *Longitudinal relaxation*: represents the exponential recovery of the longitudinal magnetization to its equilibrium value (in the direction of the main field $B_0$). The associated relaxation time is T1 (recovery of 63%).
- *Transverse relaxation*: defines the exponential decay of the net magnetization when leaving the transverse plane. The associated relaxation time is T2 (decay of 63%).

The electromagnetic radiation emitted by the spins during the relaxation process, in which the transverse magnetization decays and the longitudinal recovers, will induce a signal in a set of receiver antennas (coils) contained in the MRI system that cover the volume of the image. This MRI signal, called *free induction decay* (FID), is based on the release of energy suffered by the system at the end of the RF pulse and is represented by a sinusoid that decays with time. The FID signal represents the basis of the formation of the MR image [43].

### *2.3.2. Basics of MR images*

An MR image is a map that represents the spatial distribution of some properties of the spins contained in a sample of tissue volume. These properties may reflect the density of the spins, their mobility or the relaxation times T1 (longitudinal) or T2 (transverse) for the different tissues.

In clinical imaging, the contrast of an image is what allows the visualization of the different tissues or pathologies. The contrast is defined as the difference in the received magnitude of the MRI signal that comes from different spatial locations, which result in different levels of intensity or brightness. The difference in relaxation times T1 and T2 is a frequently used contrast mechanism since different biological tissues intrinsically exhibit different relaxation properties. Additionally, different RF pulse sequences (i.e., a temporal succession of RF pulses) are used to generate a predominant T1 or T2 enhancement of the contrast of the acquired image. This enhancement may be achieved by manipulating various parameters of these sequences such as the amplitude of the RF pulse, the repetition time (TR) or the echo time (TE) [45]. TR is the amount of time between successive pulse sequences applied to the same slice, while TE is the time between the delivery of the RF pulse and the receipt of the echo signal.

The appearance of some specific tissues in conventional structural MR images is well-known. Solid tissues, such as hard bone, or air areas provide low MRI signals, since water is practically immobilized or absent inside them. For this reason, these tissues appear dark in MR images compared to fluids or soft tissues. On the contrary, fluids and soft tissues can be represented with different contrasts according to the T1 or T2 enhancement chosen [45].

T1-weighted images are generated using shorter TR and TE and are the ones that best determine the anatomy as they show most clearly the boundaries between different tissues. This type of image is directly proportional to the release of energy in

the hydrogen nuclei. This means that, for example, liquids present longer T1 because the corresponding energy is released more slowly, and they appear darker in MR images (i.e., hypointense). If an auxiliary contrast agent (such as gadolinium) is used, MRI signal intensities change by shortening T1, thus making some lesions clearly visible in the images. In contrast-enhanced T1-weighted images, some pathological tissues, such as tumors or areas of inflammation/infection demonstrate accumulation of contrast (mostly due to leaky blood vessels) and therefore they appear brighter than surrounding tissue.

T2-weighted images are generated using longer TR and TE are the ones that best detect pathological areas since, in general, these areas have a higher water content. Therefore, liquids present longer T2 since their practically isolated nuclei perceive the same magnetic field around them and relax in a coherent way, and they appear brighter in MR images (i.e., hyperintense).

A third conventional type of MR weighted image commonly used in the clinical practice is the fluid attenuation inversion recovery (FLAIR) image. The FLAIR sequence is similar to a T2-weighted image but in this case, TE and TR times are very much longer in order to reduce the brightness of fluids while maintaining the brightness of the rest of the tissues in the image. This way, abnormalities near fluid areas can be better detected.

## 2.4. Conventional MRI of the Brain

Conventional MRI is the most commonly performed examination in neurology and neurosurgery at most institutions since this technique provides and reflects an enormous amount of information about such a complex organ as the brain. Brain MRI outperforms other brain imaging techniques such as CT and PET because it is a non-invasive technique (avoids ionizing radiation) that provides higher resolution detailed images of the brain in all three planes (axial, sagittal and coronal), shows an excellent contrast of soft tissues like gray and white matter, and allows controlling the visualization of different brain pathologies and injuries by varying the sequence parameters [43].

Broadly speaking, conventional brain MRI is useful to detect a variety of conditions of the brain such as cysts, tumors, bleeding, swelling, infections, inflammatory conditions, problems with the blood vessels, or damage caused by an injury or a stroke. Additionally, brain MRI may be appropriate for evaluating problems such as persistent headaches, dizziness, weakness, blurry vision or seizures, and it is the preferred choice for detecting certain chronic and neurodegenerative conditions of the

CNS, such as multiple sclerosis or Alzheimer's disease, due to its excellent contrast between gray and white matter [46].

As stated before, the appearance of some specific brain tissues according to the structural MRI sequence chosen is widely known. Focusing on the most common sequences used in almost every clinical MRI protocol (i.e., T1-weighted, T2-weighted and FLAIR sequences), the relative appearance of the structures that form the brain in each sequence is indicated in Table 2.1 and shown in Figure 2.3.

To conclude, it is important to emphasize the use of contrast-enhanced T1-weighted images for evaluating some brain disorders. Contrast injection improves the value of T1-weighted images by giving the greatest sensitivity for detecting pathological processes that break down the normal blood-brain barrier. Some brain pathologies such as tumors, infections and inflammations break down the blood-brain barrier and light up intensely on post-contrast T1-weighted images. Additionally, for patients with primary or secondary brain tumors, post-contrast imaging is important for detecting and characterizing metastatic brain tumors, determining the extent of primary tumor growth and defining the precise boundary between normal and neoplastic tissue [45], [47]. An example of the benefits of contrast-enhanced T1-weighted images for detecting brain lesions is shown in Figure 2.4.

**Table 2.1.** Relative appearance of some brain tissues and structures in conventional MR images. In T1-weighted images, darker appearance implies longer T1 while brighter appearance implies shorter T1. In T2-weighted images, darker appearance implies shorter T2 while brighter appearance implies longer T2. FLAIR images are analogous to T2-weighted images but with CSF set to dark.

| ***Tissue or Structure*** | **T1-weighted** | **T2-weighted** | **FLAIR** |
|---|---|---|---|
| *Cerebrospinal Fluid* | **Dark** | **Bright** | **Dark** |
| *White Matter* | **Light** | **Dark Gray** | **Dark Gray** |
| *Gray Matter (Cortex)* | **Gray** | **Light Gray** | **Light Gray** |
| *Fat (within bone marrow) and skin* | **Bright** | **Light** | **Light** |
| *Bone (skull)* | **Dark** | **Dark** | **Dark** |

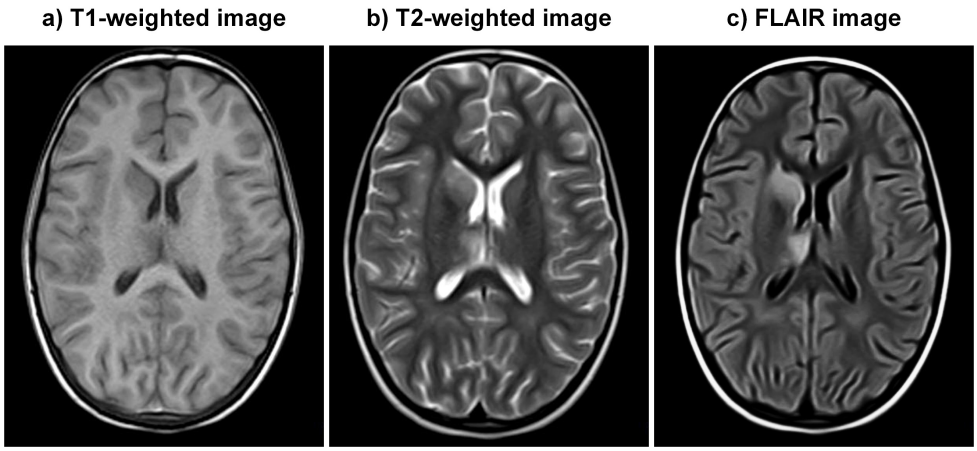



**Figure 2.3.** Comparison between axial views of T1-weighted, T2-weighted and FLAIR images of a patient with a demyelinating disease (multiple areas of abnormal signal intensity displaying bright signal in T2 and FLAIR). The difference in contrast between MRI sequences of different brain tissues can be observed in these images. Case courtesy of Dr Ahmed Abdrabou, Radiopaedia.org, rID: 22973.

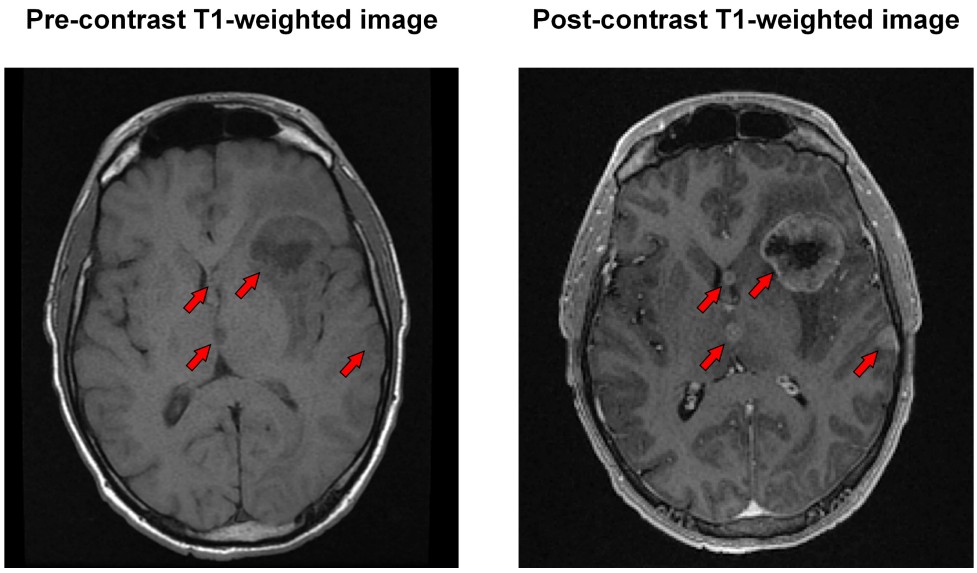



**Figure 2.4.** Axial T1-weighted scans of a patient with multiple brain metastases acquired before and after administration of a contrast agent. The red arrows indicate the metastatic lesions detected in this view. In the post-contrast image, these lesions are visible and present clearly delimited borders. In the pre-contrast image, only the biggest lesion can be slightly identified but without a clear boundary.

# Chapter 3.
# Texture analysis

## 3.1. Definition of Texture Analysis

There is no strict definition of what texture is. In general terms, every object around us presents texture and humans can easily perceive this texture as a visualization of complex patterns expressing the nature of a physical object, composed of randomly spatially organized subpatterns with uniform appearance which have characteristic brightness, color, slope or size. The spatial arrangement of these local subpatterns give rise to a set of perceived properties that describe the texture of an object: lightness, uniformity, density, roughness, regularity, linearity, frequency, phase, directionality, coarseness, randomness, fineness, smoothness or granulation, among others. In summary, the texture of an object describes the spatial arrangement of visual subpatterns in an object by means of a set of properties. However, humans usually assess these properties only qualitatively, while often a quantitative texture analysis is required to measure numerically the visual appearance on these properties [48]–[50].

In terms of image processing, texture analysis can be defined as the application of mathematical methods to extract parameters from the images that describe the pixel interrelationships, the gray level distributions and the spectral properties within these images. These texture parameters allow quantifying intrinsic heterogeneity properties from the images that are usually imperceptible to the human eye, thus characterizing and distinguishing different images accurately by their textural patterns [12]. In the past years, texture analysis has gained major attention in the medical imaging field because it is considered a reliable source of imaging biomarkers describing the internal structure of human tissues or organs that can be used to increase the precision in diagnosis or to predict treatment response, especially in cancer research [3], [51].

There exists a wide range of texture analysis methods that can be applied directly to raw medical images without processing them. However, most of the studies coincide that the best option is to process these images before applying texture analysis in order to make the most of texture analysis. In the following sections we will discuss the benefits and pitfalls of the most common processing strategies, we will introduce the texture analysis methods employed in the present thesis and we will review the state-of-the-art of texture analysis on magnetic resonance imaging (MRI).

## 3.2. Region of Interest Delineation

As previously stated, texture features can be extracted from the original entire images. However, this approach may present some important issues as unimportant heterogeneity information may blur the true texture profile of the specific tissue or organ under analysis. Therefore, texture features are usually extracted from predefined regions of interest (ROI), or volumes or interest (VOI) in the case of 3D texture analysis, containing the tissue or organ under analysis in order to delimit and characterize only certain structures present on the image. In the literature, these ROIs or VOIs are defined in different ways and there is not a definitive agreement about the best practice since different applications might require different approaches.

### *3.2.1. Segmentation of the Region*

Despite the advances in the field of automatic and semi-automatic segmentation methods and their advantages, manual delineation of the ROIs or VOIs by expert radiologists still remains as the preferred option in most of the studies since certain tissues or organs are difficult to segment properly without an expert supervision [10], [52], [53]. However, the manual segmentation approach has some disadvantages. Firstly, the manual delineation of regions may represent a time-consuming task not applicable in the clinical practice, especially when dealing with volumetric regions and large imaging datasets. Secondly, manual segmentation may be prone to high inter-observer variability. Consequently, when choosing a manual segmentation approach, it is important to assure the reproducibility and repeatability of the segmentation and the corresponding extracted features because different observers can segment certain regions differently, like in the case of brain tumors where the margins of the lesion are complex, thus leading to different texture values, especially in small regions [54], [55].

The state-of-the-art literature recommends that, when working with large datasets and volumetric regions, the segmentation method should be as automatic as possible to minimize the operator interaction and to be efficient in terms of time [1], [7], [56]. However, it is important to mention that, although automatic segmentation methods outperform manual segmentations in terms of repeatability, texture features may still be depending on the segmentation method [10].

### *3.2.2. Influence of the Region Shape and Size*

As mentioned in the previous subsection, texture analysis is usually applied to delimited image regions that only include the whole tissue or organ area (or volume) under analysis. This delimitation is usually performed with manual, semi-automatic or automatic methods, thus leading to possible inter-observer variability or time-consuming processes. However, many studies employ other simpler techniques for selecting the regions or volumes of interest based on predefined equal-sized geometric figures in 2D or 3D, like circles/spheres [57], [58] or squares/cubes [59], [60]. The benefits of using these geometric ROIs/VOIs are that the ROI definition process is easier and faster and that the size of the regions does not influence in the texture analysis performance since all the regions are sufficiently large and have the same size.

When choosing geometric regions for defining the ROIs/VOIs, two approaches are usually considered: encompassing only part of the tissue under analysis, or defining the smallest enclosing area or volume containing the whole tissue or organ of interest. The first approach captures only heterogeneity information of the tissue under analysis, but some texture details may be lost since the geometric regions does not cover the entire area or volume of interest. The second approach overcome this last issue as it covers the entire tissue or organ, but it also comprises heterogeneity information of adjacent parts that may obscure the characteristic texture outcome of the specific tissue or organ. Despite the advantages of these approaches, the segmentation of the entire area or volume of interest is still considered the best approach for conducting texture analysis because analyzing the whole region tissue without including surrounding structures may offer better texture characterization of the specific tissue [61]. Figure 3.1 shows an example of the three approaches for delineating the area of interest commented in this section.

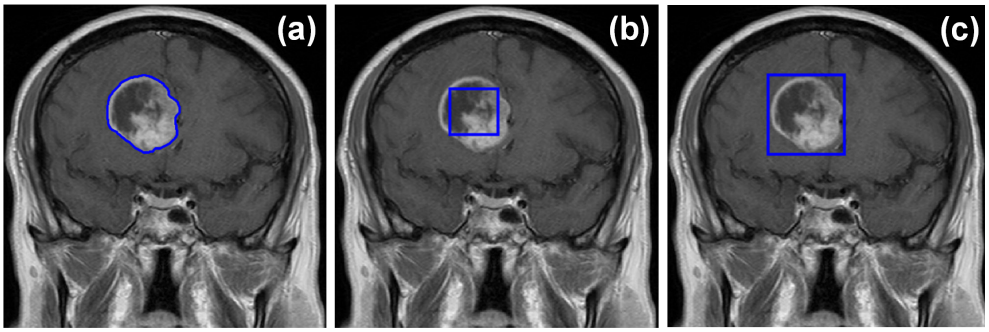



**Figure 3.1.** Approaches for defining a 2D region of interest (ROI) on a brain tumor. The preferred option is the segmentation of the entire area of interest (a) by manual or automatic strategies. However geometric regions like squares comprising only part of the tumor tissue (b) or covering the complete lesion and adjacent tissue (c) may be used to simplify the delineation process.

As briefly stated in the last paragraph, the size of the ROIs or VOIs may also influence in the texture outcome. The region size should be sufficiently large to capture the texture information that really represents the tissue. Additionally, many texture features may be dependent on the region size, thus probably leading to false results when classifying different tissues if the differences in the region sizes between groups become significant [50]. It is reported that the effect of the ROI size in the texture parameters becomes insignificant when using large areas in terms of pixels [62]. However, the region size depends directly on the imaging acquisition parameters and, although certain texture features may obey this rule, there is a wide range of texture features and not all of them have been proven to be unaffected by the ROI/VOI size [13]. Therefore, when performing texture analysis, it is important to assure that the regions are large enough to capture texture heterogeneity and that the difference between ROI/VOI sizes among groups is not statistically significant. To avoid possible effects of the region size, equal-sized geometric regions can be used, but, as mentioned before, the complete delineation of the ROI might offer better results. In conclusion, the use of a segmentation approach is recommended when the range of ROI/VOI sizes among samples is not significantly different or when the selected texture features are not affected by this difference; otherwise, geometric regions of the same size might be a better approach [13].

### *3.2.3. Dimensionality*

Traditionally, texture analysis has been conducted in 2D but in the past years, the biomedical community has made an effort to extend the 2D texture analysis

techniques to the 3D space. This is a consequence of the improvement in the 3D data acquisition and the reaching of high spatial resolutions. Hence, many studies state that promising 3D texture analysis should be considered instead of traditional 2D texture analysis because texture features extracted from volumetric regions capture more information about the tissue heterogeneity than features extracted from a single scan, thus increasing the discrimination between different tissues [51], [63].

However, 3D inter-slice texture analysis (i.e., true volumetric texture analysis) is not always applicable since not all the medical images present sufficient resolution in the third dimension. When the slice thickness of the image (the inter-slice distance) is very large in comparison to the in-plane resolution (the inter-pixel distance), 3D inter-slice texture analysis methods are not recommended [51]. This is because extending 2D texture analysis to 3D requires isotropic image resolution to ensure the conservation of scales and directions in all three dimensions, thus meaning that image interpolation should be applied in the third dimension and then the 3D image would be highly transformed due to the big difference between inter-slice and inter-pixel distances [51], [64]. In these cases, 3D intra-slice texture analysis approaches to capture the volumetric information of each image by evaluating 2D texture features may be proposed. One approach consists on extracting the 2D texture features from each scan, and then, averaging the texture values of all the scans to obtain the 3D texture features of the whole image region. Another approach consists on averaging the texture-based matrices or histograms obtained in each scan and then extracting the corresponding 3D texture features from these averaged matrices or histograms. Using these approaches, the gray-level distributions in the third dimension are not considered, but some studies demonstrated that features computed with these methods are more discriminative than features extracted from a single scan [65], [66]. Figure 3.2 shows a graphic explanation of all the approaches for computing 2D and 3D texture analysis

Despite the potential of 3D texture analysis, 2D texture analysis still remains as the selected option in many studies. Volumetric texture analysis presents some disadvantages that should be considered before discarding 2D texture analysis. First, the 3D segmentation of the VOI is more complex and time-consuming than the segmentation of a single scan, so for clinicians it is easier and faster to delineate 2D ROIs. Additionally, in some cases, only single scans present tissue of interest, thus making 3D texture analysis unviable. Finally, the clinical evaluation still remains mostly based on 2D scans, so 2D texture analysis is easier to combine with this procedure [64].

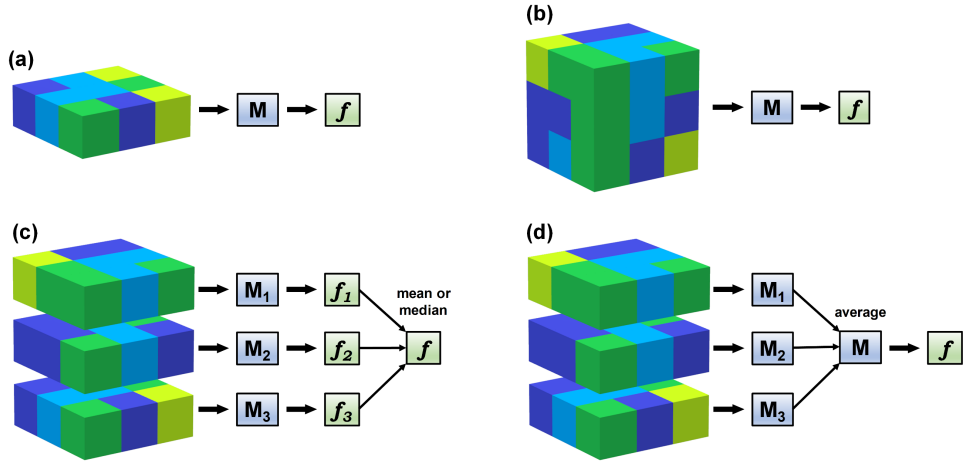



**Figure 3.2.** Approaches for computing texture analysis in 2D (a) and 3D (b, c and d), where M represents texture matrices and *f* represents texture features. Volumetric features can be obtained through 3D inter-slice analysis (extending directly 2D approaches to 3D) as in (b), or through 3D intra-slice analysis, by averaging 2D texture features (c) or 2D texture matrices (d).

## 3.3. Image Preprocessing

Once the area or volume of interest is selected, texture analysis can now be executed. Nevertheless, it is highly recommended to pre-process these image regions in order to enhance the characteristics of each tissue or organ and to minimize the effects of imaging acquisition protocols. The following pre-processing techniques are optional, but one should consider implementing them depending on the quality of the images and the purpose of each specific application. It is important to know that the texture outcome can be considerably affected depending on the methodology used throughout the process.

### *3.3.1. Image Interpolation*

Image spatial resolution is one of the most influential factors in texture analysis. Although it has been demonstrated that higher resolutions tend to improve texture discrimination power due to the increased level of detail of the captured tissue, high-resolution images are not always available in clinical routine because their acquisition time is high and they are prone to motion artifacts [67]–[70].

Image interpolation is a conventional option to enhance medical images with a low spatial resolution. The effect of image interpolation (Figure 3.3) on the texture outcome has been previously analyzed. For example, Mayerhoefer *et al.* [71] compared three interpolation methods applied on T2-weighted MRI images acquired at five different resolutions. They concluded that MR image interpolation does not improve classification rates of images at a very low resolution, but for clinical images with higher resolutions, image interpolation has the potential to improve the results of texture-based classification, recommending a maximum interpolation factor of four. Despite the promising results of this last study, the benefits of interpolation should be further investigated for each specific clinical task and for all texture analysis methods, since this method transforms the image and may not be appropriate for certain applications.

Image interpolation is of special interest when dealing with 3D texture analysis. As previously discussed, in most imaging sequences the slice thickness is larger than the in-plane resolution, so re-slicing all images to obtain isotropic image resolution is required in texture analysis to guarantee the conservation of scales and directions in the three dimensions. However, this approach is not suitable when difference between the slice thickness of the image and the in-plane resolution is substantially big [51].

### *3.3.2. Image Normalization*

Texture analysis is sensitive to all the imaging acquisition settings, including protocols, scanners and/or adjustments. Even if the same scanner and protocol are employed among subjects, the resulting images can show substantial intensity variations due to other acquisition conditions such as room temperature and hygrometry, calibration adjustment or slice location [72]. When performing texture analysis, this phenomenon may obscure true image texture since some higher-order texture parameters show dependency on global image characteristics defining the overall brightness or contrast, like mean intensity and variance [48].

Image normalization removes this dependency of texture parameters on the first order gray-level distribution, thus suppressing the effect of the imaging acquisition settings and enhancing the true texture profile of the image. This dependency was demonstrated on texture features extracted from T2-weighted images by Collewet *et al.* [72]. In this work, they concluded that a method consisting on adjusting the histogram to the $\mu \pm 3\sigma$ range, being $\mu$ is the mean value of gray levels inside the ROI and $\sigma$ is the SD, yielded the best classification results. Specifically, this simple method just ignores the gray levels located outside the range $[\mu - 3\sigma, \mu + 3\sigma]$ for further analyses. As a result

of this study, normalization has become an essential step prior to texture analysis, and this $\mu \pm 3\sigma$ normalization method has become a popular choice in most of the studies since many texture analysis software packages implement this method [49], [73].

### *3.3.3. Quantization of Gray levels*

Common images have a bit depth of 8 bits per pixel, but medical images are generally represented by bit depths ranging from 10 to 16 bits per pixels (bpp), thus resulting in contrast resolutions of 1024 to 65536 levels of gray. Texture analysis methods based on matrix computation are dependent on this gray-level range since they quantify the relationship between levels of gray. Therefore, the computation of this matrices and the corresponding texture features in images with high bit depths may become a tedious process [74], [75].

Quantization of gray levels is usually applied image regions prior to texture analysis in order to shorten the computational time of the matrix-based features by reducing the number of gray levels used to represent the image. Additionally, quantizing the images improves the signal-to-noise ratio (SNR) of the texture outcome [74]. Typical numbers of gray levels (i.e., dynamic range) tested in texture analysis studies are 8, 16, 32, 64, 128 or 256 gray levels. Figure 3.4 shows the effect of quantizing an image with different number of gray levels.

The influence of this quantization process in the discriminative power of the matrix-based texture features has been previously analyzed with diverse results. Several studies reported that no substantial difference was found when comparing the texture analysis results derived from quantizing the images with several number of gray levels [73], [76]. However, other studies showed that the discriminative power of texture features changed depending on the dynamic range chosen to quantize the images. For example, Mahmoud-Ghoneim *et al.* [75] concluded that better results were obtained for characterizing brain white matter regions when quantizing with 128 gray levels; Chen *et al.* [61] found that the optimal results for characterizing breast lesions were achieved when quantizing with 32 gray levels; and Leite *et al.* [77] observed that quantizing with 16 gray levels allowed identifying the etiology of brain white matter lesions more accurately. In the light of these mixed conclusions, it is possible to affirm that larger dynamic ranges do not necessarily give better texture results and that there is no optimal number of gray levels vary among studies. Consequently, the dynamic range should be optimized for each specific application because it can lead to better classification results [13], [75], [78].

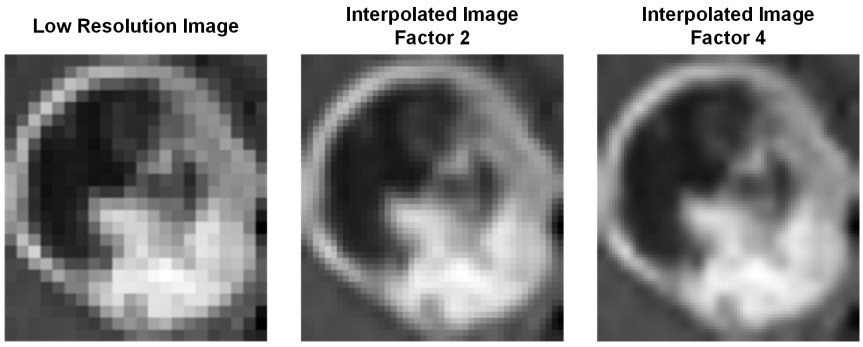



**Figure 3.3.** Effect of interpolating a bi-dimensional low-resolution image of a brain tumor by factors 2 and 4. The interpolated images show better definition and may characterize the texture profile of the tumor more appropriately.

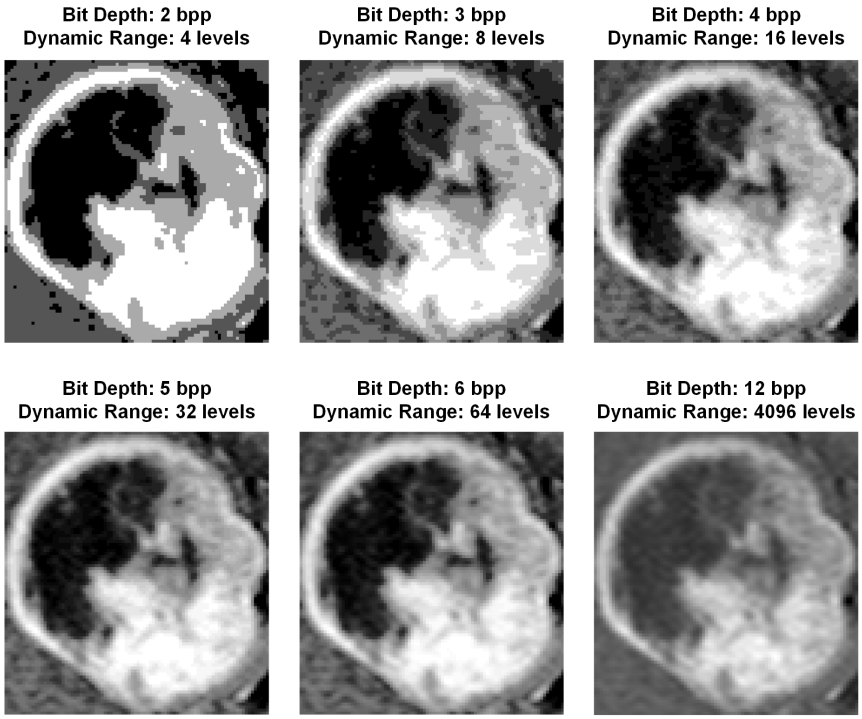



**Figure 3.4.** Effect of quantization in an image with a dynamic range of 4096 levels of gray. The contrast appearance of the image region is highly transformed, especially when using low bit depths. However, reducing the number of gray levels shorten the computational time and may improve the texture analysis performance.

## 3.4. Texture Analysis Methods

There exists a wide range of techniques to extract parameters that quantify the texture of an image. The first texture analysis methods date from the 70s but, nowadays, new methods are still being proposed and the number of possible texture features increases year after year. Testing all these texture features in a single specific study may result in a demanding and unfeasible task, so a reduced set of the best approaches is usually chosen depending on the images under analysis or the specific application, since some texture analysis methods may work better than others based on these conditions.

Extensive helpful reviews of existing feature extraction methods have been published during the past ten years [79]–[81]. However, in 2016, a large group of researchers from around the world decided to create the Image Biomarker Standardisation Initiative (IBSI), an independent international collaboration that works towards the standardization of imaging radiomics biomarkers for the purpose of high-throughput quantitative image analysis. As part of this initiative, a reference manual was published to provide a common nomenclature and definition for radiomics parameters, as well as reporting guidelines for the correct performance of radiomics analyses [82]. Therefore, this continuously updated review should be used as the reference manual for consulting the state-of-the-art of texture analysis methods and features since one of the main purposes of the radiomics practice is to establish a common framework for using reproducible and validated texture biomarkers.

The IBSI reference manual provides the equations and formulas to compute all the texture features included in the manual, so one may implement in-house functions to perform texture analysis. However, several useful software packages are currently available to conduct texture analysis. Open-source software packages mostly use MATLAB (The MathWorks Inc., Natick, MA, USA) as the preferred coding environment. Examples of open-source packages implemented in MATLAB are:

- *Radiomics* package, implemented by Vallieres *et al.* [73] (available from https://github.com/mvallieres/radiomics)
- Local binary patterns software, implemented by Ojala *et al.* [83] (available from http://www.cse.oulu.fi/CMV/Downloads/LBPSoftware)
- IBEX (Imaging Biomarker EXplorer), implemented by Zhang *et al.* [84] (available from http://bit.ly/IBEX_MDAnderson)

Other open-source software tools do not require MATLAB, like the famous MaZda package:

- MaZda (Institute of Electronics, Technical University of Lodz, Poland) [49] (available from http://www.eletel.p.lodz.pl/programy/mazda/)

As opposed to freely available packages, several companies decided to offer commercial tools following the success of texture analysis, for example:

- RadiomiX (OncoRadiomics, Maastricht, The Netherlands)
- TexRAD (Feedback plc, Cambridge, UK)

In the following subsections, we will give details of the texture analysis methods and the corresponding features used in the projects developed in the context of this thesis. All these features are included in the IBSI reference manual [82] and they were computed with open-source tools coded in MATLAB, using mainly a combination of the *Radiomics* package, the MATLAB implementation of the local binary patterns and our own self-implemented scripts.

### *3.4.1. Classification of Texture Analysis Methods*

Texture analysis methods have been traditionally categorized in four big groups according to how relationships between pixels/voxels are mathematically computed: structural, model-based, transform and statistical methods [12], [48], [79], [85].

**Structural methods**

These methods assume that textures are represented by well-defined primitives (microtexture) and a hierarchy of spatial arrangements (macrotexture) of these primitives dependent on the chosen spatial location rules. Microtextures can be as simple as individual pixels/voxels, a region with uniform gray levels or line segments. Macrotextures can be obtained by modeling geometric relationships between microtextures or by learning their statistical properties. These methods offer a good symbolic description of the image but, however, they are more useful for synthesis than for analysis tasks.

**Model-based methods**

These methods attempt to represent the texture by using sophisticated mathematical models, like fractals or autoregressive models. The parameters derived from each model are used to characterize the image. The main disadvantage of these methods resides in the computational complexity for estimating the descriptors.

### Transform methods

These methods transform the images to represent them in a different space, such as the spatial or the frequency domain, thus providing quantitative information about the texture that is not possible to obtain from the original image. In other words, transform methods consist on applying filter banks to the image and then extracting statistical properties from the filter responses. The most popular filters for computing this type of textures are the Fourier, Gabor or Wavelet transforms.

### Statistical methods

These methods represent the texture by quantifying directly the distributions and relationships between gray levels of an image in different ways, without considering any model behind the image generation and without interpreting reasons for pattern generation. Statistical methods are the most popular texture analysis methods due to their simplicity and their proved efficiency for achieving higher discrimination accuracies. Additionally, these methods provide more interpretable information that may be correlated with the pathological properties of the tissues. Therefore, most of the texture analysis methods used in the projects presented in this thesis belong to this group.

Statistical methods can be in turn classified in different subgroups according to several criteria. On the one hand, depending on the number of pixels/voxels involved in the computation, statistical methods can be categorized in:

- *First-order statistics*: describe the frequency distribution of the gray levels within the region of interest and are dependent on a single pixel/voxel value rather than its interaction with neighboring pixels/voxels.
- *Second-order statistics*: examine the relationship between a pair of pixels/voxels across the image domain by measuring the probability of finding a pair of gray levels at random distances and orientations over the image region.
- *Higher order statistics*: explore the spatial relationship among three or more pixels/voxels and are thought to closely resemble the human experience of the image.

On the other hand, according to the type of spatial relationship between pixels/voxels quantified, statistical methods can be classified in:

- *Global metrics*: quantify the whole gray-level distribution of the image region without spatial information.
- *Local metrics*: describe the spatial interrelationship between neighboring pixels/voxels with different or equal gray levels.
- *Regional metrics*: measure the distribution of groups of connected pixels/voxels with the same gray-level values (i.e., gray-level areas or volumes).

Table 3.1 presents a list of the texture analysis methods used in the context of this thesis. They are also classified according to the criteria explained above. Further details of these methods can be found in the following subsections. These methods have been selected mainly due to their popularity and efficiency in related studies, their availability in the software implementations used and their interpretability. However, as mentioned before, there exists a wide range of texture methods and, when possible, all these methods should be tested because the efficiency of one method may depend on the specific application.

**Table 3.1.** List of the texture analysis methods used in this thesis and how they are classified.

| Texture Analysis Method | Group | Order | Scale |
|---|---|---|---|
| Intensity histogram | Statistical | First-order | Global |
| Gray-level co-occurrence matrix | Statistical | Second-order | Local |
| Gray-level run-length matrix | Statistical | Higher-order | Regional |
| Gray-level size zone matrix | Statistical | Higher-order | Regional |
| Neighborhood gray-tone difference matrix | Statistical | Higher-order | Local |
| Local binary patterns | Statistical / Structural | Higher-order | Local |
| Discrete wavelet transform | Transform | - | - |

### 3.4.2. Intensity Histogram

Image intensity histogram is a first-order statistical texture analysis method that quantifies the global heterogeneity of an image region by measuring the frequency of appearance of each gray level, without taking into account correlations between pixels. This method is not only computationally simple, but also rotation and translation invariant [79]. However, histogram-derived parameters are not considered as texture descriptors by many researchers since they do not reflect spatial statistical relationships between image pixels and only refer to pixels or voxels treated as random variables. Despite everything, these parameters do characterize image regions and are usually combined with texture features in order to improve classification studies [50].

The shape of the histogram provides many clues as to the nature of the image, as shown in Figure 3.5. For example, a narrowly distributed histogram is associated to a low-contrast image and, on the contrary, a wide histogram is associated to an image with a higher range of gray levels. Bimodal histograms often suggest that the image contains an object with a narrow intensity range against a background of differing intensity [86].

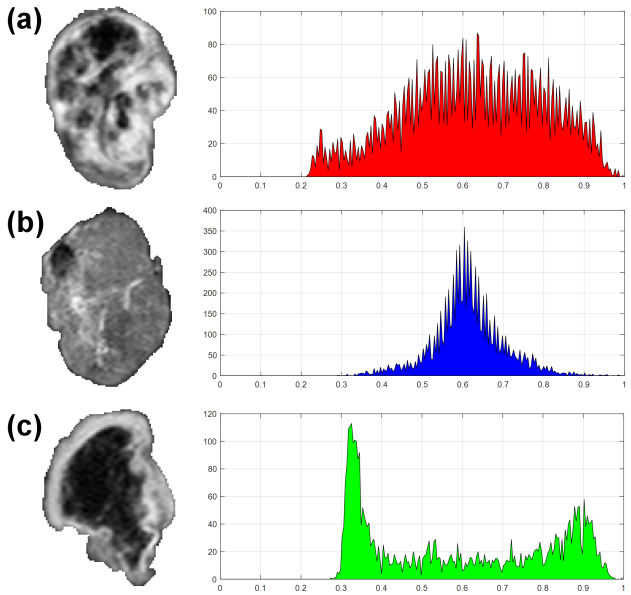



**Figure 3.5.** Histograms associated to different brain tumors. The first lesion (a) presents a wide histogram due to its heterogeneity. The second lesion is more homogeneous, so its histogram is narrower. The third lesion (c) is a ring-enhancing tumor mainly dominated by a central necrosis area and a brighter rim, which relates to a bimodal histogram with two peaks (necrosis and margin areas).

Different useful parameters can be derived from the histogram to quantitatively describe the heterogeneity properties of the image region. Assuming that $N_g$ is the number of distinct gray levels, and $p(i)$ is the normalized histogram vector (histogram whose entries are divided by the total number of pixels in the ROI) of intensity levels $i$, the most common features derived from the histogram are:

- ***Mean***: indicates the average level of intensity or brightness of the image region.

$$Mean = \mu = \sum_{i=1}^{N_g} ip(i) \qquad \textbf{Equation 3.1}$$

- ***Variance***: describes the variation of intensity around the mean, thus suggesting the level of heterogeneity of the region.

$$Variance = \sigma^2 = \sum_{i=1}^{N_g} (i-\mu)^2 p(i) \qquad \textbf{Equation 3.2}$$

- ***Skewness***: measures the asymmetry of the histogram. This feature is zero if the histogram is symmetrical about the mean and is either positive or negative depending whether it has been skewed above or below the mean.

$$Skewness = \sigma^{-3} \sum_{i=1}^{N_g} (i-\mu)^3 p(i) \qquad \textbf{Equation 3.3}$$

- ***Kurtosis***: measures the flatness (positive) or peakedness (negative) of the histogram in comparison to the normal distribution.

$$Kurtosis = \sigma^{-4} \sum_{i=1}^{N_g} [(i-\mu)^4 p(i)] - 3 \qquad \textbf{Equation 3.4}$$

- ***Energy***: describes the uniformity of the histogram. Assumes its maximum value only if the image has a very narrow histogram, dominated by a single value. Broader intensity variations cause the energy to decrease.

$$Energy = \sum_{i=1}^{N_g} [p(i)]^2 \qquad \textbf{Equation 3.5}$$

- ***Entropy***: quantifies the irregularity of the histogram. A predominantly random distribution has a high entropy. Highly correlated or uniform distributions have a low entropy.

$$Entropy = -\sum_{i=1}^{N_g} p(i) \log_2 p(i) \qquad \textbf{Equation 3.6}$$

### *3.4.3. Gray-Level Co-occurrence Matrix*

The gray-level co-occurrence matrix (GLCM) is a second-order statistical texture analysis method that was first proposed by Haralick *et al.* in 1973 [87] to describe local heterogeneity information. This method quantifies the relationship between gray levels in an image region by counting the pairs of pixels separated by a predefined distance ($d$) and direction (θ) that have the same distribution of gray-level values. Each pixel of the resulting matrix represents the number of times that the gray level of a reference pixel and the gray level of the neighbor pixel in the predefined distance and direction are seen in the image region under analysis. Consequently, the size of the GLCM will be $N_g \times N_g$, being $N_g$ the number of gray levels of the image. Figure 3.6 shows an example of computation of this GLCM.

When computing GLCMs, it is important to know that, according to the original definition, GLCMs are symmetric, thus meaning that the co-occurring pairs obtained, for example, for θ = 0° would be equal to those obtained for θ = 180° (i.e., when calculating the number of times the gray-level value $i$ is adjacent to the value $j$, counts both $(i, j)$ and $(j, i)$ pairings). Non-symmetric GLCMs, as the one showed in Figure 3.6, can be also computed, but they are less common than the original symmetric GLCMs.

As previously mentioned, GLCMs are dependent on the distance and the direction. The dependence on the distance is not a problem and the pixel distance may be chosen according to the application, although typically distances of $d$ = 1 pixel are selected in order to maintain the texture analysis as local as possible. However, the GLCMs present one major concern based on their dependence on the direction. If this

dependence is not suppressed in some way, different texture values may be obtained if the image is rotated, thus affecting the results when images from different patients have different orientations [13]. To solve this problem, rotation invariance may be achieved by computing symmetric GLCMs in the four directions (horizontal, vertical, 45°, 135°) of the 2D space (13 directions for 3D approaches) and then averaging or summing these matrices to obtain a single matrix from which rotation invariant features are extracted. Another approach consists on extracting firstly the features from the direction-dependent GLCMs and then averaging the features over all directions.

Several statistics can be mathematically computed from the GLCM in order to quantify the homogeneity (smoothness) or heterogeneity (coarseness) of the image region. Let us assume that $p(i,j)$ is the $(i,j)$-th entry in a normalized GLCM; $p_x(i) = \sum_{j=1}^{N_{rows}} p(i,j)$ and $p_y(j) = \sum_{i=1}^{N_{column}} p(i,j)$ are the $i$-th and $j$-th entries in the marginal probability matrix obtained by summing the rows and columns of $p(i,j)$ respectively; and $\mu_x, \mu_y, \sigma_x, \sigma_y$ represent the mean and SD of $p_x$ and $p_y$. Let us also define the diagonal and cross-diagonal probabilities $p_{x+y}(k)$ and $p_{x-y}(k)$, which represent the gray level sum $(k = i + j)$ and difference $(k = |i - j|)$ distributions respectively. Assuming the latter, the features used in this thesis are:

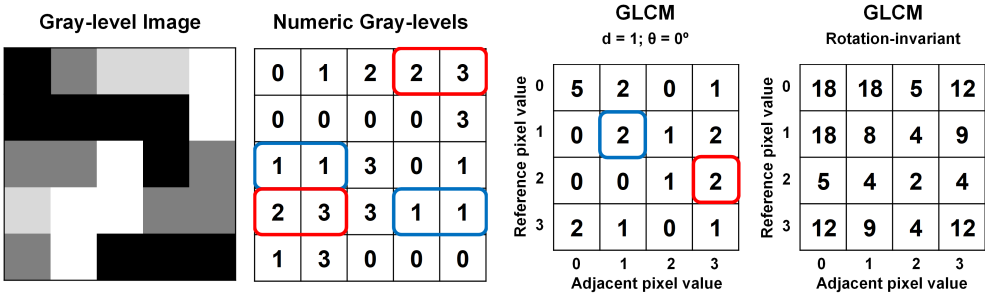



**Figure 3.6.** Computation of GLCMs for a given 5×5 pixel image with a dynamic range of 4 gray levels. The first GLCM is computed non-symmetrically for the horizontal direction (θ = 0°) and for one-pixel separation ($d$ = 1). The values in blue and red in this GLCM indicate the number of transitions of each of the gray-level pairings marked. The second GLCM is the result of summing the symmetric GLCMs over the four directions of the 2D space to achieve rotation invariance.

- ***Energy***: also known as *Angular second moment* or *Uniformity*, measures the local uniformity (or orderliness) of an image. High values are related to homogeneous regions, thus indicating that the intensities within the region are very similar.

$$Energy = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}[p(i,j)]^2 \qquad \textbf{Equation 3.7}$$

- ***Contrast***: describes the local intensity variations between different structures present in the image region. High values are related to high heterogeneity.

$$Contrast = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2 p(i,j) \qquad \textbf{Equation 3.8}$$

- ***Correlation***: measures the gray-level linear dependency between intensities. It presents how a reference pixel/voxel is locally correlated to its neighbor. High values indicate high level of correlation between pixels/voxels, that is, homogeneity.

$$Correlation = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{ijp(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \qquad \textbf{Equation 3.9}$$

- ***Homogeneity***: also known as *Inverse difference moment*, describes the local homogeneity of the image. High values are associated to smooth regions, in which most of the gray levels are the same.

$$Homogeneity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{p(i,j)}{1+(i-j)^2} \qquad \textbf{Equation 3.10}$$

- ***Variance***: also known as *Sum of squares*, measures the dispersion of the values around the mean of the gray-level distribution, that is, the randomness of the pixel/voxel distribution in the image. Higher values refer to heterogeneous regions.

$$Variance = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(1-\mu)^2 p(i,j) \qquad \textbf{Equation 3.11}$$

- ***Entropy***: expresses the degree of chaos or disorder (i.e., randomness) within an image. High values imply random distributions, that is, heterogeneous regions.

$$Entropy = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)log\big(p(i,j)\big) \qquad \textbf{Equation 3.12}$$

- ***Dissimilarity***: also known as *Difference Average*, measures the level dissimilarity between pairs of pixels/voxels in an image.

$$Dissimilarity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} |i-j|p(i,j) \qquad \textbf{Equation 3.13}$$

- ***Autocorrelation***: measures how pixel/voxel pairs are correlated.

$$Autocorrelation = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ijp(i,j) \qquad \textbf{Equation 3.14}$$

- ***Sum Average***: estimates the overall image brightness. Originally, this feature measures the mean of the cross-diagonal probabilities, but the following adapted version allows computing the mean of all gray-level combinations.

$$Sum\ Average = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} [ip(i,j) + jp(i,j)] \qquad \textbf{Equation 3.15}$$

- ***Sum Variance***: measures the dispersion (with regard to the mean) of the gray level sum distribution of the image.

$$Sum\ Variance = \sum_{k=2}^{2N_g}\left(k - \left[\sum_{k=2}^{2N_g} kp_{x+y}(k)\right]\right)^2 p_{x+y}(k) \qquad \textbf{Equation 3.16}$$

- ***Difference Variance***: measures the dispersion (with regard to the mean) of the gray level difference distribution of the image.

$$Difference\ Variance = \sum_{k=0}^{N_g-1} \left( k - \left[ \sum_{k=0}^{N_g-1} k p_{x-y}(k) \right] \right)^2 p_{x-y}(k) \quad \textbf{Equation 3.17}$$

- ***Sum Entropy***: estimates the disorder related to the gray level sum distribution of the image.

$$Sum\ Entropy = -\sum_{k=2}^{2N_g} p_{x+y}(k) \log(p_{x+y}(k)) \quad \textbf{Equation 3.18}$$

- ***Difference Entropy***: estimates the disorder related to the gray level difference distribution of the image.

$$Difference\ Entropy = -\sum_{k=0}^{N_g-1} p_{x-y}(k) \log(p_{x-y}(k)) \quad \textbf{Equation 3.19}$$

- ***Information Measure of Correlation***: quantifies the linear dependency or correlation between intensities but adding some desirable properties that are not represented by the original correlation descriptor. This measure is subdivided in two features: the first and the second information measures of correlation (FIMC and SIMC respectively).

$$FIMC = \frac{HXY - HXY_1}{\max\{HX, HY\}} \quad \textbf{Equation 3.20}$$

$$SIMC = \sqrt{(1 - \exp[-2(HXY_2 - HXY)])} \quad \textbf{Equation 3.21}$$

where $HXY$ is equivalent to the entropy descriptor, $HX$ and $HY$ are the entropies of $p_x(i)$ and $p_y(j)$ respectively, $HXY_1$ and $HXY_2$ are types of entropy defined as:

$$HXY_1 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) log\left(p_x(i) p_y(j)\right)$$

$$HXY_2 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_x(i) p_y(j) log\left(p_x(i) p_y(j)\right)$$

### 3.4.4. Gray-Level Run-Length Matrix

The gray-level run-length matrix (GLRLM) is a statistical matrix-based texture analysis method of a higher-order that describes regional heterogeneity information. This method, first proposed by Galloway in 1975 [88] and extended by Chu *et al.* [89] and Dasarathy and Holder [90], examines the times that each gray level value is seen consecutively in an image in a predefined direction (θ). The GLRLM is constructed by detecting and counting the runs (sequences of consecutive pixels with the same gray level) of different gray levels and their lengths in the image. Each row of the GLRLM represents a gray level and each column a specific length, so each element of the matrix indicates the number of runs of a specific gray level and length in the image (Figure 3.7).

The GLRLMs are originally symmetric, and runs are only counted once per pairing of directions (θ = 0º and 180º, θ = 45º and 225º, etc.). Additionally, features derived from the GLRLM are originally dependent on direction, as in the case of GLCM. To achieve rotation invariance, the method consisting on averaging the matrices computed over all directions is usually applied, as previously mentioned for the GLCMs.

In general, the features extracted from the GLRLM can be used to define fine textures (dominated by short runs) or coarse textures (dominated by longer runs) [80]. Assuming that $N_g$ is the number of gray levels, $N_r$ is the maximal possible run, and $r(i,j)$ is the $r(i,j)^{th}$ entry in the normalized GLRLM (i.e., the number of times there is a run of length $j$ having gray level $i$), the most common features derived from the histogram are:

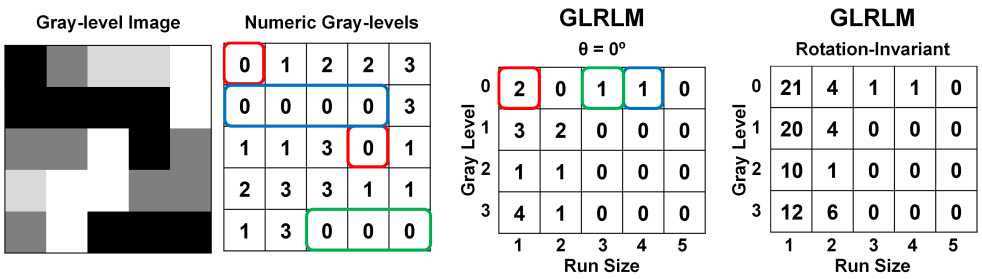



**Figure 3.7.** Computation of GLRLMs for a given 5×5 pixel image with a dynamic range of 4 gray levels. The first GLRLM is computed for the horizontal direction (θ = 0°). The values in blue, red and green in this GLRLM indicate the number of runs of different sizes (lengths) counted in the horizontal direction for the 0 gray-level. The second GLRLM is the result of summing the GLRLMs over the four directions of the 2D space to achieve rotation invariance.

- ***Short Run Emphasis (SRE)***: emphasizes short run lengths.

$$SRE = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{r(i,j)}{j^2} \quad \textbf{Equation 3.22}$$

- ***Long Run Emphasis (LRE)***: emphasizes long run lengths.

$$LRE = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} j^2 r(i,j) \quad \textbf{Equation 3.23}$$

- ***Gray-level Non-uniformity (GLN)***: assesses the distribution of runs over the gray values. Low values are related to runs equally distributed along gray levels.

$$GLN = \sum_{i=1}^{N_g} \left( \sum_{j=1}^{N_r} r(i,j) \right)^2 \quad \textbf{Equation 3.24}$$

- ***Run-Length Non-uniformity (RLN)***: assesses the distribution of runs over the run lengths. Low values are related to runs equally distributed along run lengths.

$$RLN = \sum_{j=1}^{N_r} \left( \sum_{i=1}^{N_g} r(i,j) \right)^2 \quad \textbf{Equation 3.25}$$

- ***Run Percentage (RP)***: assesses the fraction of the number of realized runs and the maximum number of potential runs. Low values are associated to strongly linear or highly uniform regions.

$$RP = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} r(i,j)}{\sum_{i=1}^{N_g} j \sum_{j=1}^{N_r} r(i,j)} \quad \textbf{Equation 3.26}$$

- ***Low Gray-level Run Emphasis (LGRE)***: is analogous to SRE but, instead of short run lengths, low gray levels are emphasized.

$$LGRE = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{r(i,j)}{i^2}$$ **Equation 3.27**

- ***High Gray-level Run Emphasis (HGRE)***: is analogous to LRE but, instead of short run lengths, high gray levels are emphasized.

$$HGRE = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i^2 r(i,j)$$ **Equation 3.28**

- ***Short Run Low Gray-level Emphasis (SRLGE)***: emphasizes runs in the upper left quadrant of the GLRLM, where short run lengths and low gray levels are located.

$$SRLGE = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{r(i,j)}{i^2 j^2}$$ **Equation 3.29**

- ***Short Run High Gray-level Emphasis (SRHGE)***: emphasizes runs in the lower left quadrant of the GLRLM, where short run lengths and high gray levels are located.

$$SRHGE = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{i^2 r(i,j)}{j^2}$$ **Equation 3.30**

- ***Long Run Low Gray-level Emphasis (LRLGE)***: emphasizes runs in the upper right quadrant of the GLRLM, where long run lengths and low gray levels are located.

$$LRLGE = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{j^2 r(i,j)}{i^2}$$ **Equation 3.31**

- ***Long Run High Gray-level Emphasis (LRHGE)***: emphasizes runs in the lower right quadrant of the GLRLM, where long run lengths and high gray levels are located.

$$LRHGE = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i^2 j^2 r(i,j)$$ **Equation 3.32**

- ***Gray-level Variance (GLV)***: estimates the variance in runs for the gray levels.

$$GLV = \frac{1}{N_g \times N_r} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \left( ir(i,j) - \sum_{i=1}^{N_g} i \sum_{j=1}^{N_r} r(i,j) \right)^2 \quad \textbf{Equation 3.33}$$

- ***Run-Length Variance (RLV)***: estimates the variance in runs for run lengths.

$$RLV = \frac{1}{N_g \times N_r} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \left( jr(i,j) - \sum_{j=1}^{N_r} j \sum_{i=1}^{N_g} r(i,j) \right)^2 \quad \textbf{Equation 3.34}$$

### *3.4.5. Gray-Level Size Zone Matrix*

The gray-level size zone matrix (GLSZM) is a higher-order statistical matrix-based texture analysis method introduced by Thibault *et al.* in 2009 [91] that describes regional heterogeneity information in a similar way to the GLRLM. The concept of GLSZMs is based on extending the GLRLM runs to areas or volumes. This way, a GLSZM counts the number of groups of connected pixels or voxels with the same gray levels (i.e., zones) along across the entire image. An example of computation of this matrix is shown in Figure 3.8.

An advantage of this method is that GLSZMs are originally independent on direction and distance, so they are rotation invariant by default and only one matrix is extracted per image region. The features extracted from the GLSZM are analogous to those defined for the GLRLM (Equation 3.22 to Equation 3.34) but replacing the number of runs ($N_r$) with the number of zones ($N_z$). Therefore, the features derived from the GLSZM are: Small Zone Emphasis (SZE), Large Zone Emphasis (LZE), Gray-level Non-uniformity (GLN), Zone-Size Non-uniformity (ZSN), Zone Percentage (ZP), Low Gray-level Zone Emphasis (LGZE), High Gray-level Zone Emphasis (HGZE), Small Zone Low Gray-level Emphasis (SZLGE), Small Zone High Gray-level Emphasis (SZHGE), Large Zone Low Gray-level Emphasis (LZLGE), Large Zone High Gray-level Emphasis (LZHGE), Gray-level Variance (GLV) and Zone-Size Variance (ZSV).

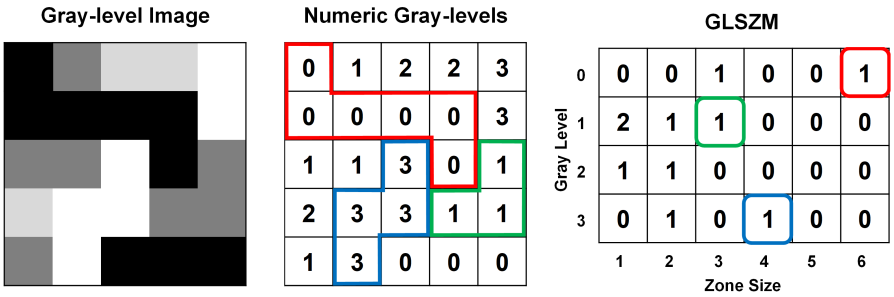



**Figure 3.8.** Computation of the GLSZM associated to a given 5×5 pixel image with a dynamic range of 4 gray levels. The GLSZM is not dependent on distance or direction, so it is invariant to rotation and only one GLSZM can be computed per image. The values in blue, red and green indicate the number of zones (linked pixels with the same gray level) of a particular size and gray level found in the image.

### *3.4.6. Neighborhood Gray-Tone Difference Matrix*

The neighborhood gray-tone difference matrix (NGTDM) is a higher-order statistical texture analysis method that describes the local heterogeneity of the image under analysis. This method, proposed by Amadasun and King in 1989 [92], is an alternative to GLCM in order to quantify the local properties of the image by counting the sum of gray level differences of pixels/voxels with discretized gray level $i$ and the average discretized gray level of neighboring pixels/voxels within a distance $d$. Therefore, the $i$-th entry of the NGTDM is a summation of the differences between all pixels with gray-tone $i$ and the average value of their surrounding neighbors at a distance $d$, as shown in Figure 3.9.

The NGTDM is not really a matrix but an array, since its size is $N_g \times 1$, being $N_g$ the highest gray level present in the image. In addition, as in the case of GLCM, this method is dependent on the distance, usually being $d = 1$ the preferred option, but, on the contrary, it is not dependent on the direction, so it is rotation invariant by default.

The mathematical formulation to compute the NGTDM and its features is somewhat complex. Assuming that $s(i)$ represents the NGTDM, the $i$-th entry of the NGTDM is defined as:

$$s(i) = \begin{cases} \sum |i - \bar{A}_i| & \text{for } i \in N_i \text{ if } N_i \neq 0, \\ 0 & \text{otherwise.} \end{cases} \qquad \textbf{Equation 3.35}$$

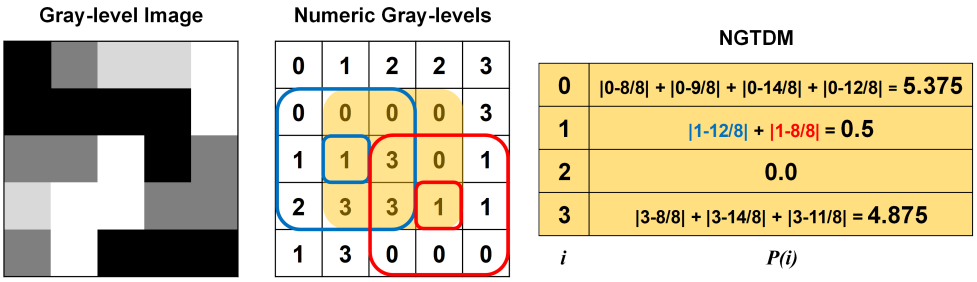



**Figure 3.9.** Computation of the NGTDM of $d = 1$ associated to a given 5×5 pixel image with a dynamic range of 4 gray levels. By selecting a distance $d = 1$, the neighborhood of centered pixels is set to a size of 3×3 (orange square). The pixels outside the orange square are in the periphery of the image. Therefore, the NGTDM elements are computed by summing the differences between all centered pixels (inside the orange square) with gray level ***i*** and the average value of their neighbors. For example, the NGTDM entry corresponding to the gray level $i = 1$ is marked in blue and red.

being $N_i$ the set of all pixels with gray-level $i$ in the image excluding the peripheral region of width $d$, and $\overline{A}_i$ is the average gray-level summation of the neighbors around a center pixel with gray-level $i$ located at position $(k, l)$. Therefore, $\overline{A}_i$ is defined as follows:

$$\overline{A_i} = \overline{A}(k,l) = \frac{1}{W-1}\left[\sum_{m=-d}^{d}\sum_{n=-d}^{d} f(k+m, l+n)\right], (m,n) \neq (0,0) \qquad \textbf{Equation 3.36}$$

being $d$ the chosen distance that coincide with the neighborhood size, and $W = (2d+1)^2$. Figure 3.9 also shows an example of how to compute this matrix.

In general, five textures are extracted from the NGTDM. Assuming that for an $N \times N$ image, the probability of occurrence of the gray-level value $i$ is defined as $n_i = \frac{N_i}{n^2}$, where $n^2 = N - 2d$, the NGTDM features are defined as:

- ***Contrast***: represents the level of contrast of the image. High values indicate that the intensity difference between neighboring regions is large.

$$Contrast = \left[\frac{1}{{N_g}^{eff}[{N_g}^{eff}-1]}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} n_i n_j (i-j)^2\right]\left[\frac{1}{n^2}\sum_{i=1}^{N_g} s(i)\right] \qquad \textbf{Equation 3.37}$$

where ${N_g}^{eff}$ is the effective number of different gray levels in the image (some gray levels may not appear in the image region due to the quantization process).

- ***Coarseness***: gives an indication of the level of spatial rate of change in intensity. High values represent coarse textures, where gray-level differences are small.

$$Coarseness = \left[\epsilon + \sum_{i=1}^{N_g} n_i s(i)\right]^{-1} \quad \textbf{Equation 3.38}$$

where the constant $\epsilon$ is a small number to prevent this parameter becoming infinite.

- ***Busyness***: represents the level of spatial frequency of intensity changes. High values indicate busy textures (i.e., there are rapid changes of intensity from one pixel to its neighbor).

$$Busyness = \frac{\sum_{i=1}^{N_g} n_i s(i)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(in_i - jn_j)}, n_i \neq 0, n_j \neq 0 \quad \textbf{Equation 3.39}$$

- ***Complexity***: refers to the visual information content of a texture. High values represent complex textures, that is, non-uniform textures with high information content, for example when many patches or primitives are present.

$$Complexity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{|i-j|\left[n_i s(i) + n_j s(j)\right]}{n^2(n_i + n_j)}, n_i \neq 0, n_j \neq 0 \quad \textbf{Equation 3.40}$$

- ***Strength***: gives an idea of the level of detail of the primitives. High values are related to strong textures, where the primitives are easily definable and clearly visible.

$$Strength = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(n_i + n_j)(i-j)^2}{\left[\epsilon + \sum_{i=1}^{N_g} s(i)\right]}, n_i \neq 0, n_j \neq 0 \quad \textbf{Equation 3.41}$$

where the constant $\epsilon$ is a small number to prevent infinite values.

### 3.4.7. Local Binary Patterns

The local binary patterns (LBP) is a texture analysis method introduced by Ojala *et al*. in 1996 [93] that has become very popular in the past years due to its high discrimination efficiency and its computational simplicity at the same time. This method is a combination of a higher-order statistical method and a structural approach that measures the local heterogeneity of the image. The LBP functioning consists on labelling each pixel of the region under analysis by comparing its gray level with the gray levels of the surrounding pixels and then assigning a specific binary number. This binary number for each pixel is obtained by allocating a value of 1 to those surrounding pixels with a greater or equal gray level value and a 0 to those surrounding pixels with a lower gray level value. Originally, LBP was defined for patches of 3 × 3 pixels, but it was later extended for blocks of $P$ surrounding pixels separated by a distance $R$. Taking this generalization into account and given a pixel $c$ with coordinates $(x_c, y_c)$, the LBP binary number assigned to each pixel of the image is calculated using Equation 3.42.

$$LBP_{R,P} = \sum_{p=0}^{P-1} sign(g_p - g_c) \times 2^p \qquad \textbf{Equation 3.42}$$

Where $g_p$ and $g_c$ are the gray level values of the central pixel $c$ and its neighbor pixel $p$, and the function $sign(g_p - g_c)$ is defined as:

$$sign(g_p - g_c) = \begin{cases} 1 & if \ \ g_p - g_c \geq 0 \\ 0 & if \ \ g_p - g_c < 0 \end{cases} \qquad \textbf{Equation 3.43}$$

Once the Equation 3.42 is applied to all the pixels in the image, an LBP image or map is obtained and all the bins of the histogram of this image are used as texture features. Figure 3.10 shows an example of computation of this LBP map. Other statistics can be extracted from the LBP image and used as texture features like the mean or the variance.

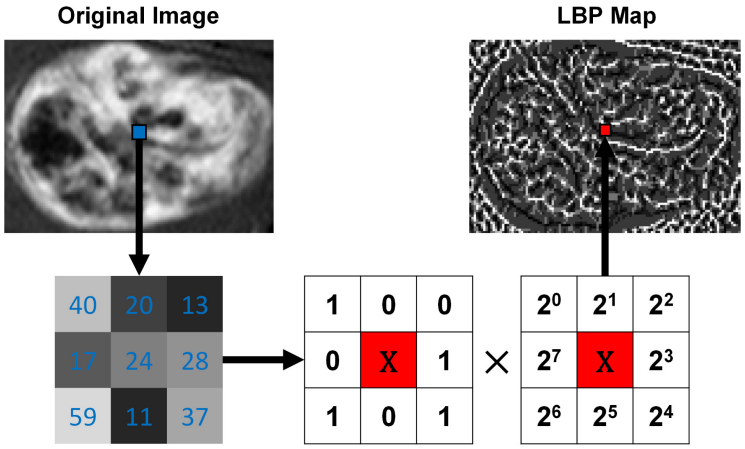
In most of the studies, the original LBP operator (patches of 3 × 3 pixels: $P = 8$, $R = 1$) is employed to preserve the texture analysis as local as possible since some regions may not be very large. Using this approach, 256 texture features are computed, corresponding to the $2^P = 256$ bins of the associated histogram. However, the original version of the LBP is not invariant to rotation, so an improved approach was proposed by Ojala *et al.* in 2002 [83] to solve this concern. The authors aimed to achieve rotation invariance by performing a circular bit-wise right shift operation (rotating the neighbor

pixel set clockwise) and assigning the smallest LBP binary number. Using this approach with the original LBP parameters ($P = 8$, $R = 1$), 36 unique rotation invariant histogram-based LBP features are obtained, as only 36 LBP binary numbers can occur for $P = 8$. Another advantage of this rotation invariant LBP approach is that a previous image quantization step is not required since it is robust to intensity variations [94].

### *3.4.8. Wavelet Transform for Texture Analysis*

The discrete Wavelet transform (DWT) is a technique that examines the spatial frequency patterns of an image within different scales and frequency directions, considering that frequency is directly proportional to gray level variations in an image. The DWT applied to an image produces four matrices of coefficients (subimages) that represent the approximations or low frequencies (LL: low-low) and the details or high frequencies in the vertical (LH: low-high), horizontal (HL: high-low) and diagonal (HH: high-high) directions, as shown in Figure 3.11. The DWT can be repeated consecutively to achieve a major image decomposition: the first level of decomposition ($LL_1$, $LH_1$, $HL_1$ and $HH_1$) is applied to the original image as mentioned before and the subsequent levels are applied to the matrix of approximations of the previous level ($LL_i$, $LH_i$, $HL_i$ and $HH_i$, where $i$ is the level of decomposition).

The DWT can be used as a transform texture analysis method by processing these subimages to obtain statistical parameters that describe the spatial frequency information of the image. This method has received much attention because it presents some advantages. Firstly, textures may be represented at the most appropriate scale by varying the spatial resolution, and moreover, a wide range of choices for the DWT function are available in order to enhance texture analysis according to each specific application [48]. Additionally, DWT-based texture features were demonstrated to be less sensitive to changes in the acquisition protocols [95]. Statistical texture descriptors (like the ones presented in the previous subsections) extracted from the DWT subimages have been previously used in some studies with successful results [96]–[98].




$$X = 2^0 + 2^3 + 2^4 + 2^6 = 1 + 8 + 16 + 64 = 89$$

**Figure 3.10.** Computation of an LBP map using the original LBP operator (patches of 3×3 pixels: P = 8, R = 1). The gray level of each pixel in the original image is compared to the gray level of the 8 connected pixels. A value of 1 is assigned to the pixels with gray level greater or equal than the central pixel, and a value of 0 otherwise. Then, a binary number is calculated and this value is allocated in the same position of the original pixel. This process is repeated for all the pixels of the original image.

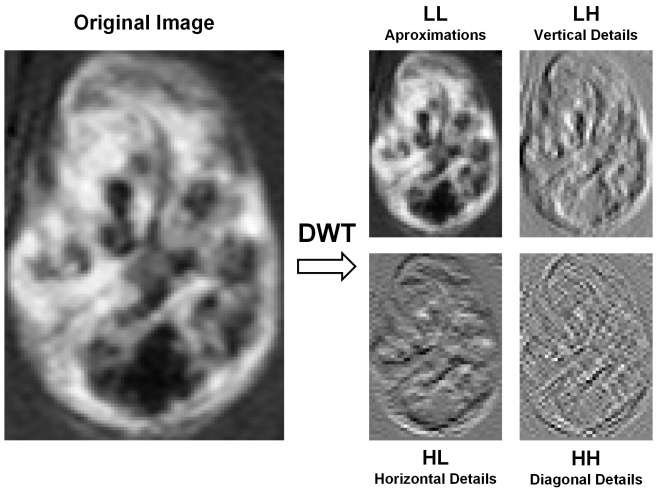



**Figure 3.11.** First DWT decomposition of MRI T1-weighted image of a brain tumor. The high-high (HH) subimage represents diagonal high frequencies, high-low (HL) extracts the horizontal high frequencies, low-high (LH) vertical high frequencies and the subimage low-low (LL) represents the lowest frequencies.

## 3.5. Review of Texture Analysis in MRI

Texture analysis may be applied to different or multiple imaging modalities and selecting the one (or ones) to investigate depends on each specific application. However, the most popular techniques in the context of the radiomics practice are Computed Tomography (CT) and MRI, since they are believed to contain valuable undiscovered information that may have a potential impact in routine care [11].

Of all the imaging modalities, CT appears to be the most appropriate technique to conduct reproducible radiomics studies because imaging performance tends to be standardized across institutions and vendors, thus facilitating the comparisons between results. Furthermore, texture data extraction and interpretation from CT is relatively easy because their units of measurement, Hounsfield units (HUs), represent tissue radiodensity. This means that CT images may contain information directly associated to structural properties of the tissue that can be translated into clinically meaningful data [7]. Nevertheless, radiomics studies in CT present some disadvantages such as the implicit patient exposure to ionizing radiation or the fact that CT acquisition conditions like the reconstruction algorithms or the image noise might influence the appearance of texture features [99], [100].

Despite the success of CT radiomics studies in texture analysis' earliest years, the majority of literature over the past years has been directed toward tissue classification and characterization in MRI, especially for brain studies [51], [85]. The current success of MRI in radiomics studies is mainly based on its growing availability in the clinical routine and the resulting high-quality images that offer excellent anatomic details thanks to new advances in technology. Additionally, existing MRI techniques provide different contrasts or modalities that have the ability to sense not only macroscopic alterations, but also microscopic-level organization of the tissues from different perspectives. Texture analysis on MRI is believed to capture these microscopic-level patterns and to extract and quantify the heterogeneous characteristics that may define each pathology. However, MRI present some specific issues that have to be considered and assessed before performing texture analysis in order to obtain reliable and reproducible results.

In the following subsections we will discuss the disadvantages of MRI that difficult radiomics analyses and we will introduce the most studied applications of texture analysis on MRI.

### *3.5.1. The Issues of Texture Analysis in MRI*

Despite the potential of MRI for providing quantitative biomarkers, the acquisition conditions associated to this imaging technique presents some important characteristic drawbacks that may obscure texture analysis if they are not addressed properly.

**Influence of the field strength**

The field strength of the MRI scanners influence in the texture outcome since scanners with higher field strength produce images with higher spatial and temporal resolutions. Not all MRI scanners have the same field strength but, in clinical routine, only scanners of 1.5T and 3T are commonly employed. Based on the resulting image resolution, 3T scanners are expected to be better for texture characterization because of the higher definition of the images. However, texture features extracted from images obtained with 3T scanners may be more sensitive to changes in the MRI acquisition parameters [71]. Additionally, scanners with higher field strength are prone to enhance artefacts derived from the patient movement, thus possibly affecting the true texture profile of the image region under analysis [101].

**Selection of the best MRI sequence**

There is no general agreement on which the best MRI sequence for texture analysis is since it depends on the application. Therefore, it is recommended to check previous works before applying texture analysis in order to select the MRI sequence that better suits each specific task. For example, contrast-enhanced T1-weighted MRI is the most popular MRI protocol for to assessing brain tumor characterization by means of texture analysis as it is employed for initial brain tumor detection and contains abundant diagnostic information [102]–[107]. However, the most desirable approach would be to compare the performance of different modalities, but these imaging data is not always available.

**Inhomogeneity correction in MRI**

Intensity inhomogeneity (also known as also intensity non-uniformity or spatial bias) is a residual MRI artifact that produces subtle smooth variations of intensity in the resulting images. This issue is mainly caused by static magnetic field inhomogeneity and imperfections of the radiofrequency coils [108], [109]. As mentioned before, texture features quantify the intensity profile of the image in different ways, so this intensity inhomogeneity may affect the texture outcome. Normalization of the image region may

solve partially this problem but, in some cases, this residual effect is not completely eliminated with normalization and correction of these inhomogeneity artifacts in MRI is recommended as a preprocessing step prior to region normalization [110]. However, it is important to remark that this issue principally affects large regions containing entire organs or even the whole brain, but for uncommon brain structures/masses like brain tumors, the heterogeneity of the tissue may change considerably when applying inhomogeneity correction, thus affecting the true lesion texture profile.

**Multicenter validation and reproducibility**

As previously introduced, variations in acquisition parameters may result in differences in the texture outcome that are not due to underlying biologic effects. This issue is not very important in studies using CT images because acquisition settings tend to be standardized across centers. On the contrary, in studies using MRI, this is a major problem of reproducibility as different centers and manufacturers may apply different acquisition parameters. Therefore, when reporting results of texture analysis, one should explain in detail the acquisition protocol to allow reproducibility of the work [108], [111]. To reduce the possible differences among multicenter MRI scans, images can be normalized into a common spatial and intensity space. However, the best approach would be to standardize the MRI protocols and parameters to ensure the utility of texture descriptors as reliable biomarkers regardless of acquisition conditions.

**Interpretation of the information**

As opposed to CT images, where signal intensity can be directly correlated with the density of the tissue, the signal intensities in MR images are difficult to associate to physical properties of tissue since they are generated from a complex interaction of parameters intrinsic to the technology, such as pulse sequences, relaxation times and acquisition conditions [7], [112]. However, in the past years many studies have focused on studying the correlation between the heterogeneity properties derived from MRI and the histopathology of the tissue under analysis [113]–[115]. The results are promising but still uncertain and dependent on the application, so substantial efforts are still required to integrate radiomic features on MRI in the clinical routine for functioning as general biomarkers.

### *3.5.2. The Present of Texture Analysis in MRI*

As previously mentioned, texture analysis has been applied to different imaging modalities since it was first proposed in the 70s. Although CT has been the preferred imaging modality to conduct texture analysis over the years and it is still useful for some medical applications, major attention has been paid to MRI texture analysis in the last years, since this technique offers excellent anatomic details and enables the enhancement of different types of tissues by modifying the acquisition settings.

In 2014, Depeursinge *et al*. [51] conducted a systematic review of publications using 3D texture analysis on medical images and concluded that MRI was the preferred imaging modality for performing volumetric analysis (38.5% of the papers). Regarding the organs under study, Depeursinge *et al*. found that brain lesions and diseases were the most investigated organs in the context of texture analysis, being MRI the preferred option to image the brain. However, for studying other organs like liver or lung, CT images were still the favorite modality.

In a recent and specific review conducted by Larroza *et al*. [13], they revealed that a total of 140 original studies published before February 2016 dealt with texture analysis in clinical MRI. The distribution of these publications per organ revealed that nearly the 54% of the published works examined different diseases or lesions affecting the brain.

To know the current state-of-the-art of texture analysis in MRI, we performed a similar search in SciVerse Scopus (https://www.scopus.com as of January 29$^{th}$ 2019) of original articles and conference proceedings containing one of the following combinations of keywords in the title:

- "texture" and "MRI"
- "texture" and "magnetic resonance"
- "radiomics" and "MRI"
- "radiomics" and "magnetic resonance"

We tested these combinations of keywords because in the past years the concept of "texture analysis" has been interchanged in many studies with "radiomics analysis", as it also may comprise the analysis of other quantitative features apart from texture features. In this search we obtained a total of 429 documents between 1977 and 2018. From this quantity, 197 documents were published during the past 3 years (44 in 2016, 62 in 2017 and 91 in 2018), which represents a 45.92% of the total amount of documents presenting texture analyses on MRI. These figures reveal that there still exists a huge

interest in this research field since the number of publications per year seems to be increasing as the years go by. The distribution of papers per organ obtained in this search also confirms that texture analysis on MRI for studying brain disorders is still the most popular application with 75 publications in the past 3 years (representing the 38.07% of the total amount of papers published between 2016 and 2018), so it is clear that one should consider MRI as the best imaging modality for examining different brain pathologies with texture analysis.

### *3.5.3. Applications of Texture Analysis in MRI*

Texture analysis applications in MRI mainly encompass segmentation of specific anatomical structures or lesions and differentiation between pathological and healthy tissues. Concerning the specific objectives of these studies, texture analysis has been applied to examine a huge variety of diseases and lesions in different organs, but, as stated in the previous subsection, major attention has been paid to neurological applications [116]. Table 3.2 shows some relevant example of applications of texture analysis on MRI to evaluate lesions or diseases affecting different organs published during the past decade.

However, over the past few years, substantial efforts have been made towards understanding the biologic profile of cancer by means of radiomic features extracted from MRI, in order to improve the diagnostic, treatment and follow-up stages of the disease [4], [5], [8], [10]. Radiomics analyses using CT images have already demonstrated that imaging quantitative features may be useful for characterizing the tumor nature [52], [100], so based on these promising results, this seems to be the main goal of texture analysis studies on MRI for the following years.

**Table 3.2.** Examples of applications of texture analysis in MRI published during the past decade.

| Organ | Lesion / Disease | Objectives | References |
|---|---|---|---|
| Brain | Primary brain tumors | Classification of benign and malign tumors; Grading of gliomas | [103], [104], [117], [118] |
| Brain | Brain metastases | Differentiation from radiation necrosis; Identification of the primary cancer | [106], [107] |
| Brain | Dementia | Identification of Alzheimer's disease | [58], [119] |
| Brain | Multiple sclerosis | Early diagnosis | [120], [121] |
| Brain | Ischemic Stroke | Prediction of hemorrhagic transformation; Evaluation of small vessel disease | [122], [123] |
| Brain | Mild traumatic brain injury | Effect of trauma in cerebral tissue | [57] |
| Heart | Myocardial infarction | Differentiation between acute and chronic | [124] |
| Heart | Arrhythmias | Classification of low and high-risk patients | [125] |
| Breast | Breast cancer | Classification of benign and malign lesions; Classification of cancer molecular subtypes | [60], [61], [126] |
| Prostate | Prostate cancer | Detection of cancerous tissue | [127], [128] |
| Kidney | Autosomal dominant polycystic disease | Prediction of renal function decline | [129] |
| Liver | Liver fibrosis | Assessment of the disease | [130] |
| Knee | Knee osteoarthritis | Quantification of subchondral bone architecture; Identification of bone marrow lesions | [131], [132] |

# Chapter 4.
# Data analysis with machine learning

## 4.1. What is Machine Learning?

Machine learning refers to the branch of the artificial intelligence field that uses sophisticated statistical algorithms to give computer systems (i.e., the *machines*) the power of learning from mineable data with the purpose of recognizing patterns automatically and building predictive models capable of comprehending new unseen data and solve practical tasks. The concept of *learning* refers to the autonomous optimization of the algorithm parameters in order to improve progressively the performance of the specific task.

Thanks to its power and efficacy, machine learning has seen an increased interest over the years in many fields [133] such as speech recognition and translation between languages [134], business intelligence [135], autonomous navigation [136] or fraud and credit scoring [137]. In the medical imaging field, this discipline has been implemented more slowly than in other fields. However, in the past decades, machine learning has demonstrated its growing potential for complementing medical imaging thanks to the advances in computer technology and new applications have been developed mainly in the areas of computer-aided detection and diagnosis (CAD) and clinical decision support systems. The implementation of machine learning algorithms in these systems helps clinicians to interpret appropriately medical imaging findings and reduce interpretation times [138], [139]. Machine learning problems can be categorized mainly into two learning types according to the main purpose of the task to be solved [140], [141]:

- ***Supervised learning***: implies that each sample is formed by two elements: the set of input observations (*features*) and the output observation (*label* or *class*). The main goal is prediction, that is, to deduce a functional relationship between training data features that generalizes well to testing data in order to predict their class. Supervised learning can be subdivided in *classification* (when the labels are grouped into discrete categorical classes) and *regression* (when the labels are continuous numerical values) problems.
- ***Unsupervised learning***: implies that each sample is only represented by the input observation or features, and no output observation or label is provided. The main goal is description, that is, to discover associations or patterns between samples and reveal the latent classes behind the features. The most representative example of unsupervised learning is *clustering* (grouping data without prior information of the groups).

Machine learning is part of the *data mining* practice, which also encompasses data acquisition and exploration. In this thesis, the data acquisition process represents the extraction and processing of texture features from MR images, and the data exploration process implies every statistical test or data visualization technique (e.g., boxplots or scatterplots) conducted for detecting preliminary tendencies, incomplete data or outlier values. In this chapter, we are going to focus only on machine learning, concretely on supervised classification methods and the corresponding previous issues and challenges to be studied so as to make the most of them.

## 4.2. Machine Learning Algorithms for Classification

In practice, machine learning becomes an iterative process where not only one algorithm is applied. The development of the optimal predictive model implies the selection of the most appropriate technique. This selection is not direct and requires the implementation of several models so as to test them and choose the one that offers best results in terms of accuracy, generalization and interpretability [142].

In this thesis, we dealt with supervised classification problems in all the projects conducted, as all our samples were previously labelled and our main objective was to evaluate the categorization of these samples in various known groups or classes. Therefore, we had to choose the most appropriate classification models (also known as *classifiers*) from a wide range of available algorithms in order to perform our analyses and compare their performances. Concretely, for our classification analyses we used the

Caret package (short for Classification And REgression Training) [143] in R language (R Development Core Team, Vienna, Austria), which presents 189 different classification models, so we had to select a reduced diverse set of commonly used classifiers, including models from different families and nature, considering their proved efficacy in other applications [144]–[146]. Testing all the available classification models would be an exhaustive, unfeasible task that may not result in a remarkable improvement of the classification results.

Hence, in this section we only provide a brief description of the five different classification models implemented throughout the realization of the experimental studies involved in this thesis.

### *4.2.1. Naive Bayes Classifier*

The naive Bayes (NB) classifier is one of the oldest machine learning models and stands out for its computational simplicity, time efficiency and robustness. This classifier works well in many real-life applications. In particular, it is widely used in areas such as text classification [147] and spam filtering [148], but it has been also successfully applied in the study of medical data [141], [149].

The NB classifier is a probabilistic model that predicts the probability of a given case belonging to a certain class. It uses the Bayes rule, based on the Bayes theorem. This rule allows answering the following question: "based on the features that we have observed, what is the probability that the outcome belongs to class $C_i$?". Mathematically, let $Y$ represent the class variable and $X = (x_1, x_2, \dots, x_n)$ the collection of $N$ feature variables. The goal is to estimate the posterior probability $P(Y = C_i|X)$, which is defined as the probability that the outcome is the $i$-th class given $X$. The corresponding formula to compute this probability is:

$$P(Y = C_i|X) = \frac{P(X|Y = C_i)P(Y = C_i)}{P(X)} \qquad \textbf{Equation 4.1}$$

where $P(X|Y = C_i)$ is the conditional probability (the probability of observing the set of feature values $X$ for the data associated to class $C_i$), $P(Y)$ is the prior probability of the outcome (the probability of expecting the class $C_i$), and $P(X)$ is the probability of expecting the set feature values $X$. The final predicted class will be the one associated with the largest posterior probability.

The naive Bayes model simplifies the computation of these last probabilities by assuming that the effect feature value on a certain class is independent of the values of the other features. This assumption, known as class conditional independence, is extremely strong (naive) but, however, yields a significant reduction in the complexity of the calculations [142], [150].

As mentioned before, the NB model is very attractive because it is simple and quick even for large training datasets, and it is often surprisingly effective even in its simplest form. A large number of modifications have been introduced over the years in an attempt to make it more flexible, but such modifications may introduce complications that detract from its basic simplicity [151].

### *4.2.2. K-Nearest Neighbors*

The k-nearest neighbors (KNN) classifier is a non-parametric method that works easily with multiclass problems as well as with applications in which an object can have many class labels, and is usually tested due to its simplicity, flexibility and good performance [145], [151]. This classifier is a type of instance-based learning, or lazy learning, method where a generalized model is not built explicitly and all computation is deferred until classification, thus meaning that there is no explicit training phase and all (or most) of the training data is needed during the testing phase.

Typical supervised classification methods search for linear or nonlinear boundaries that optimally separate the data. Instead, KNN classifier takes a different approach based on the concept of similarity that uses the sample's geographic neighborhood to define these boundaries. The KNN classifier predicts a new sample using the *K* closest samples of the training set, so that only the distance to the *K* nearest neighbors and their associated classes are used to classify new samples. Class probability estimates for the new sample are calculated as the proportion of training set neighbors in each class [150]. Figure 4.1 shows an example of prediction of a new sample using KNN.

Basically, the accuracy of the KNN method relies on the number of neighbors taking into account and distance or similarity metric evaluated between samples. The choice of the number of neighbors *K* depends mainly on the characteristics of the data. Small *K* values may generate many small regions of each class, thus representing a highly localized fitting (i.e., overfitting), whereas high *K* values may lead to the formation of fewer, larger and less flexible regions that do not really represent the local structure of the data.

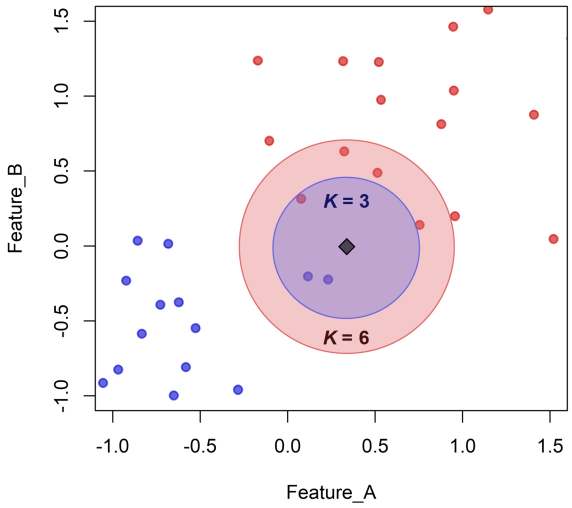



**Figure 4.1.** Example of KNN classification for a dataset with two features and two classes illustrated as blue and red points. A new sample (the black rhombus) must be assigned to one of the classes. The new sample will belong to the most commonly occurring class depending on the number of neighbors *K* selected. In this case, if *K* = 1, 2 or 3 (blue circle), the new sample will be predicted as blue since the two closest samples are blue. If K > 5 (for example, the red circle *K* = 6), it will be predicted as red.

Regarding the measurement of distance to assess the closeness of the training samples, several metrics can be considered. A common metric is the Minkowski distance, whose equation is:

$$d_{Minkowski}(x,y) = \left(\sum_{i=1}^{N}(|x_i - y_i|)^r\right)^{1/r} \qquad \textbf{Equation 4.2}$$

where $x$ and $y$ represent the $N$ values of the features of two cases and $r$ is the order. Minkowski distance is typically used with $r$ being 1 or 2, where the former is known as the Manhattan distance and the latter is the Euclidean distance.

It is important to recall that, for any distance metric, the scales of the features affect the resulting distance calculations because, if features are on widely different scales, the distance between samples will be biased towards features with larger scales. Hence, it is recommended to center and scale all features prior to KNN classification in order to allow each feature to contribute equally to the distance estimate [150].

### 4.2.3. Support Vector Machines

Support Vector Machine (SVM) is an advanced classification model with high popularity due to its robustness, flexibility and efficacy in a variety of applications, especially in computational biology and CAD systems in radiology [141], [151], [152]. Originally, this classifier was developed to solve binary classification problems involving linearly separable data, but it was extended over the years to deal with different classification tasks effectively.

In a binary classification task, the main function of the SVM algorithm is to maximize the margin distance between the classification boundary (i.e., *hyperplane*) and the closest samples of both classes. In the case of perfectly separable classes, there exist infinite hyperplanes that can separate the classes correctly but SVMs only select the one that allows the linear maximum margin classification, that is, the one that achieves the maximum space from the boundary to the closest training set samples from both classes, as shown in Figure 4.2. The associated margin is usually known as *hard margin* and the samples on this margin are called *support vectors*. The reason behind finding the hyperplanes that accomplish the maximum margin distance is that they deliver the best classification performance on the training data and they offer the best generalization ability by leaving much room for the correct classification of future data [150], [151].

When working with more complex data where the classes are not perfectly separable, it is not possible to obtain the unique hyperplane that satisfies the maximum margin classification. In those cases, the SVM algorithm constructs a hyperplane with a *soft margin*, which also separates both classes but allowing some misclassified points with the purpose of performing better when grouping the remaining data points. The SVM algorithm allows controlling and optimizing the trade-off between misclassification of the training data and the size of the soft margin by adjusting the internal parameter known as *cost* or *C* in the training process.

The use of hyperplanes allows classifying linearly separable data, but when the data are not linearly divisible, a kernel function may be used to map the data into a higher dimensional space where it is possible to separate the data linearly. Non-linear kernels like Gaussian or polynomial (Figure 4.3) allow more flexible classification boundaries at the expense of risk of overfitting and a time-consuming training process with more internal parameters to be adjusted.

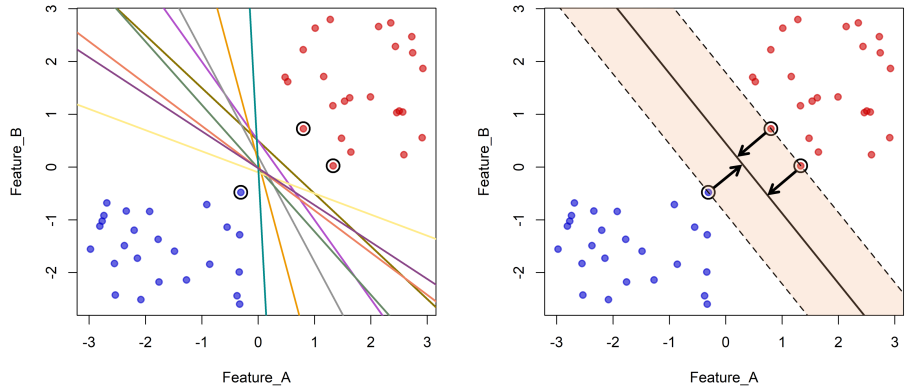



**Figure 4.2.** Dataset with linearly separable classes, illustrated in blue and red points, classified with SVM. The left plot shows several hyperplanes, out of many possible, that correctly separates both classes. The right plot illustrates the maximal margin hyperplane, shown as a solid black line. The margin is the distance from the solid line to either of the dashed lines (in light orange). The points circled in black that lie on the dashed lines are the support vectors.

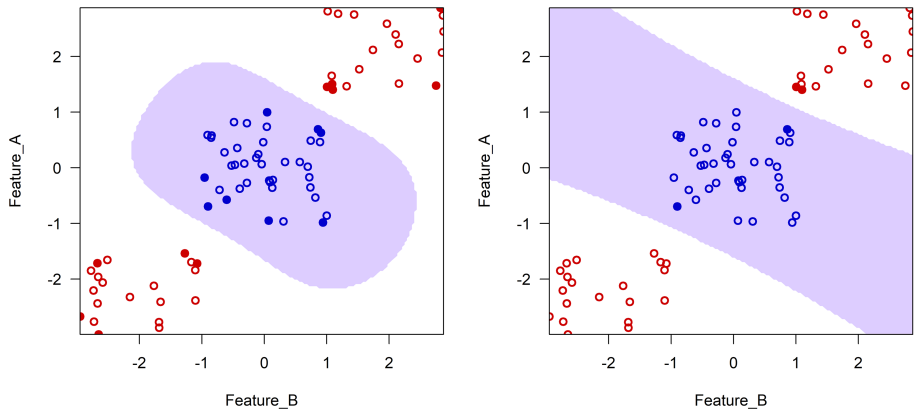



**Figure 4.3.** Data set with non-linearly separable classes, illustrated in blue and red points, classified with SVM with non-linear kernels. The plots show how a Gaussian kernel (left) and polynomial kernel of degree 2 (right) can separate both classes. The classification boundaries are shown in light purple while the support vectors are highlighted with solid points.

### 4.2.4. Decision Trees and Random Forests

The Random Forest (RF) classifier is an ensemble machine learning model that has been proved to provide better results than other models in many studies (including medical imaging studies) involving classification tasks [144], [153], especially when dealing with multiclass problems [146] and image segmentation challenges [154]. This classifier is an ensemble model because it combines the results of a multitude of independent and decorrelated decision trees with different structures in the training process.

Decision or classification trees are predictive models based on the accomplishment of a set of successive binary rules that results in the partition of data into smaller, more homogeneous groups. The construction of a decision tree is based on the principle of "divide and conquer": through a supervised learning algorithm, successive divisions of the data based on the feature values are carried out in order to maximize the distance between groups in each division. A basic tree is formed by a collection of *nodes* organized in a hierarchical structure and interconnected with *edges* or *branches* that represent the decisions made. Nodes are subdivided into internal (or *split*) nodes and terminal (or *leaf*) nodes. The function of the internal nodes is to store test functions based on the feature values that are used to sort sequentially each new incoming sample introduced in the decision tree through the root node, which corresponds to the feature that best divides the training data. Each terminal node stores the final answer. Therefore, a decision tree can be interpreted as a technique for splitting complex problems into a hierarchy of simpler ones. Additionally, decision trees present an important advantage: they produce human-readable, comprehensible rules regarding how to classify a given sample [142], [153]. Figure 4.4 shows an example of a simple decision tree trained for solving a binary task.

Decision trees are prone to overfitting if grown sufficiently deep and tend to ignore some variables that may be important in the case of small sample sizes and large number of features. The RF classifier allows improving the generalization of the model and the robustness against overfitting and small sample size problems by evaluating multiple decision trees, at the expense of interpretability [155], [156]. Specifically, the RF algorithm for constructing each tree of the forest and predict the class of new samples works as follows:

- A specific number of trees to be built is selected
- A subset of the training data is selected randomly to grow each tree to the largest extent possible.

- A subset of features is selected randomly out of all feature variables and the best candidates are used to split the nodes successively.
- The remaining subset of training data is used as a validation set to assess the accuracy of the model. The average error obtained for each training sample in all the trees (out-of-bag error estimate) is used to determine the overall performance of the classification in the training phase.
- For assigning a new sample to a class, each tree in the forest generates a vote for the classification of the new sample, and the class chosen will be the one with the highest proportion of votes, that is, the most represented among all the trees in the forest.

Another advantage of the advantages of RF model is the little parameter tuning required. The parameter *mtry*, which identifies the number of random features used in each tree, controls the strength (i.e., how accurate the individual trees are) and the correlation (i.e., the dependence between trees) of the RF model. Another tuning parameter may be the number of trees to be built, that has to be chosen according to each specific application.

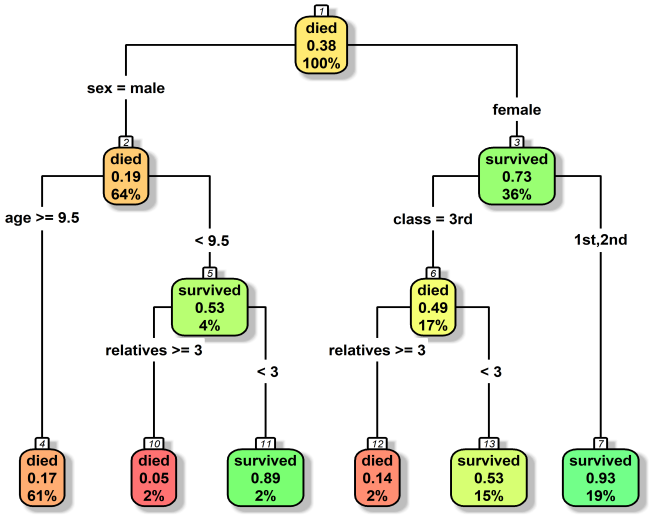



**Figure 4.4.** A simple classification tree trained for assessing the survival of the Titanic sinking by classifying passengers into dead and surviving. Each passenger of the training dataset is characterized by four features (sex, age, class and number of relatives). Each node shows the selected predicted class (died or survived), the predicted probability of survival and the percentage of observations in the node. Titanic passengers data acquired from the "Titanic: Machine Learning from Disaster" competition dataset provided by Kaggle (https://www.kaggle.com/c/titanic)

### 4.2.5. Artificial Neural Networks and Multilayer Perceptrons

An artificial neural network (ANN) is a learning technique inspired by the way the human brain works, learns and processes information. The original purpose of the ANN approach was to solve problems in the same way that a human brain would. However, over time, this technique was also applied to study other diverse classification tasks such as speech recognition, machine translation, social network filtering or forecasting. In the biomedical field, ANNs have been used extensively for clinical diagnosis, medical imaging analysis, histopathology assessment, drug evaluation and follow-up of a wide variety of diseases [141], [157], [158].

The ANNs simulate the way the brain sends a receives information by implementing neurons joined together in a pattern of inter-connections. The neurons in the ANN are known as *nodes*, and they process the information and transmit signals to other nodes through established connections with associated weights. Each node applies an activation or transfer function to the sum of inputs weighted according to the connection weights, thus producing an output value that is transmitted to the remaining nodes of the network. If the activation function chosen is a binary step function, each node will behave like a switch, that is, activating only under certain thresholds just as neurons are activated only when sufficient neurotransmitter is accumulated. By doing so, the ANN behaves like a linear classifier. However, non-linear activation functions are usually preferred because this way the combination of nodes and activation functions will be able to represent non-linear relationships [142].

The nodes in an ANN are organized in layers. Three types of layers can be found in an ANN: *input layers*, which receive the information to be processed; *output layers*, which show the results of the processing; and *hidden layers*, situated in between. Different layers may perform different kinds of transformations on their inputs.

There is a large variety of ANN models, classified depending mainly on the topology (number of neurons and hidden layers, and how they are connected) and the learning algorithm. In practice, the most commonly used ANN model is the multilayer perceptron (MLP). The MLP is a feedforward ANN (i.e., allow signals to travel only in one way, from input to output) that consists of at least three layers (input, hidden and output layers). Each node of the input layer corresponds to a feature and each node of the output layer represents a classification (e.g., one output layer corresponds to a binary prediction based on a certain cut off point). The MLP can be composed of different number of hidden layers, with a minimum of one, which transform and transfer the data from the input to the output layer. For training the MLP, a supervised learning technique

called backpropagation is used. It consists of distributing the error term computed at the output layer backwards throughout the layers, by modifying the weights at each node. Figure 4.5 shows an example of a simple trained MLP.

One of the main concerns of MLPs is determining the proper size of the hidden layer/layers, because an underestimate of the number of hidden neurons may lead to poor approximation and generalization abilities, while excessive nodes can result in computationally expensive, complex networks prone to overfitting. Additionally, MLPs are said to lack of interpretability since it is difficult to understand the nature of the underlying representations generated by the networks. However, MLPs have several advantages that make them one of the preferred machine learning methods for classification: modelling of any sort of functional relationship, adaptive and flexible learning through examples and robustness in handling redundant and inaccurate information [142], [159].

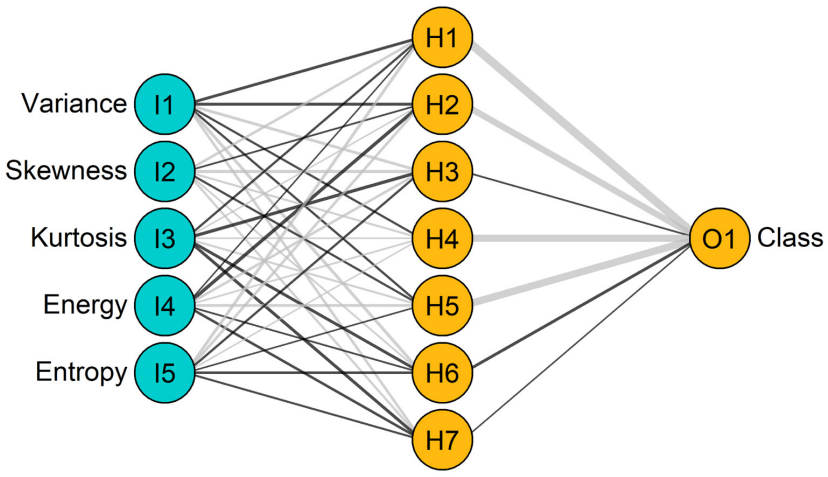



**Figure 4.5.** An example of an MLP with three layers (one input layer, one output layer and only one hidden layer). The input layer has five nodes (I1 to I5) corresponding to five features. The output layer has only one node (O1) since it is a binary prediction. The hidden layer has seven nodes (H1 to H7). The interconnections between nodes are indicated with lines, where thicker lines indicate larger weights and color indicates the sign of the weight value (positive weights in black and negative weights in gray).

## 4.3. Feature Selection

Most of the machine learning algorithms are strongly influenced by the quality of the features used to train the models. In many cases, the success of a classification task mainly depends on choosing the right features to build the right models. Consequently, it can be said that features, also commonly referred indistinctively as *predictors*, *parameters* or *variables*, are intimately connected to classification models because a model is only as good as its features are [160]. However, choosing the right features is not always an easy task that can be simply assessed with data exploration and statistical tests because features that may be completely useless by themselves can provide a significant performance improvement when combined with others within a machine learning approach [161].

One important aspect of studying a set of features is that they may interact in various ways, thus having different effects on the classification results. Sometimes such interaction can be exploited to enhance discrimination, sometimes it can be ignored, and sometimes it implies a challenge. The main challenge appears when dealing with a high-dimensional feature space, that is, when a huge number of features of different nature are processed and combined within a machine learning structure. In first place, as the number of features increases, the classification modelling becomes more complex. In addition, some of the features may be highly correlated and some may contain irrelevant information that can affect negatively the construction of the classification model [162].

Feature selection is the process that allows managing the choice of the proper subset of features that maximizes the classification accuracy. Most of the feature selection techniques make use of statistical metrics or search algorithms in order to generate a feature ranking that indicates which are the most meaningful features and which the redundant ones. This way, the proper combination of features for each specific task according to their relevance can be chosen. In contrast, some feature selection methods skip the ranking generation step and search and choose the proper subset of features automatically, without user interaction. Implementing a feature selection method in the classification process allows reducing the dimensionality of the feature space, speeding up the classification of new data and increasing the accuracy of the predictive model in most of the cases [161], [163].

It is important to mention that feature selection methods differ from other commonly used dimensionality reduction techniques such as principal component analysis (PCA) or linear discriminant analysis (LDA). The latter attempt to reduce the

dimension of the feature space by altering and transforming the original representation of all the features whereas feature selection methods select an optimal subset of features, thus preserving their meaning and avoiding the loss of interpretability [164].

Feature selection methods can be categorized in three main groups depending on how they combine the feature selection process with the construction of the predictive model: *filter*, *wrapper* and *embedded* methods. Each category has its own advantages and drawbacks, which are discussed hereafter.

### *4.3.1. Filter Methods*

Filter feature selection methods make use of a certain metric to rank and assess the discriminatory power of each feature individually, without taking the interaction with the classifier into consideration. This last fact may seem a disadvantage, but, on the contrary, it leads to important advantages: they are adaptable to all classification techniques, easily scalable to high-dimensional datasets and computationally simple and fast. Additionally, these methods can be used as preliminary analyses in order to measure the importance of the features for accomplishing certain classification task according to some criteria. However, most of the filter feature selection techniques are *univariate*, meaning that they evaluate the relevance of each feature individually and probably ignore some dependencies between features that might improve the accuracy of the classification task [164].

Typical statistical methods, such as the *t*-test or the Mann-Whitney-Wilcoxon (MWW) test, which is analog to the *t*-test but without the requirement of the normality assumption, can be used as univariate methods to rank and select features presenting statistical significance between groups, according to the corresponding *p*-value [162].

Another group of univariate metrics commonly used to assess the variable importance is the one based on mutual information and information gain measures, which give an overall estimation of the mutual dependence between the features and the classes [165]. One of the most recent metrics based on mutual information theory is the *maximal information coefficient* (MIC), which can be defined as a correlation coefficient computed by using binning as a means to find the largest mutual information value and that belongs to the maximal information-based nonparametric exploration (MINE) class of statistics [166]. The use of MIC for feature selection can be also seen as a correlation-based feature selection technique which measures the strength of the linear or non-linear association between each feature and the corresponding class distribution [167], [168].

This metric has a similar scale to the simple Pearson correlation statistic (which only measures linear associations), where a value of zero indicates no relationship between the feature and the class, whereas a value of one demonstrate an extremely strong functional relationship or correlation. Therefore, higher values of MIC suggest that the corresponding features can explain the distribution of classes better than other features.

To solve the main problem presented by univariate techniques, some *multivariate* filter methods have been proposed over the years, aiming at the incorporation of feature dependencies to some degree. One of the most recognized families of multivariate filter methods is the one based on the *Relief* algorithm [169]. The original Relief algorithm [170] uses the interaction between randomly selected samples and their closest neighbors in order to compute a score for each feature. In particular, to calculate this score, this technique performs iterations through a procedure of weight updating. First, this method randomly selects a sample from the training set and then identifies the nearest samples of both classes (by Euclidean distance), considering each of these samples as a *hit* when belonging to the same class or a *miss* otherwise. For each feature, the difference between the feature values at the random point and the hits and misses is calculated. This process is repeated for a predefined number of randomly selected samples and the weight of each feature is an accumulation of these differences, in such a way that the weight increases if the hit is close to the randomly selected sample and decreases if the miss is close. The idea is that a feature with a high discriminatory power between classes should have hits nearby and misses far away (i.e., higher Relief weight or score). Over the years, improved variants of the Relief algorithm were proposed, being the *ReliefF* algorithm the most successful version [171]. The ReliefF variant, unlike the original Relief algorithm, allows dealing with more than two classes and is more robust since it is able to work with noisy or even incomplete data by means of conditional probabilities. This method also uses a difference metric but, in this case, more than one neighbors can be selected to determine the weights or scores.

### *4.3.2. Wrapper Methods*

Wrapper feature selection methods employ search algorithms in combination with multiple classifiers to evaluate the predictive capacity of features in groups and determine the optimal feature subset. The main advantage of these methods is that they analyze the interaction between features, so the optimal feature subset should include those features that together provide the best classification accuracy. Nevertheless, the evaluation of the candidates for optimal feature subset in wrapper approaches is obtained

by training and testing several times a specific classification algorithm, thus tying the outcome to a specific model. Additionally, this classifier dependence makes wrapper methods very computationally expensive, especially as the feature space grows, and susceptible to overfitting if all possible feature combinations are not tested [164].

The quality of the selected subset of features in wrapper methods depends fundamentally on the search algorithm. The most straightforward approach to find the optimal combination of features is to perform an *exhaustive search*, consisting on testing all possible combinations of features with the selected classifier in order to find the one that yields the best classification accuracy. However, this method is extremely time-consuming, especially when the feature space is very high, so an algorithm that uses some type of smart search strategy is usually preferred. The most popular search strategies are *forward selection* and *backward elimination*. Forward selection starts with an empty testing subset and features are added one (or more) at time to this subset so as to form progressively larger subsets. On the contrary, backward elimination starts with the full set of features as the testing subset and the least promising features are iteratively removed one (or more) at time in order to form smaller subsets. Both strategies evaluate the candidates feature subsets by using an objective function, such as the classification accuracy of a predictive model. In general, forward selection is computationally more efficient but it may result in more unstable subsets since the interaction of all features is not completely assessed [161].

### *4.3.3. Embedded Methods*

Embedded feature selection methods are similar to wrapper approaches, but in this case, the process of selecting the optimal feature subset is implemented by default in the construction of certain classification models. Therefore, the advantages and disadvantages of embedded methods are similar to those of wrapper methods but, in this case, the computational complexity is better [164].

One of the most popular embedded techniques is the internal estimate of variable importance implemented in the RF training scheme [156], [172]. The RF classifier evaluates the importance of each feature automatically during the training phase of the multiple decision trees with several measures, and, at the end, all features are ranked according to these measures. One of these variable importance measures is the *Gini importance* or *mean decrease in impurity* (MDI), which is defined as the total decrease in impurity of the nodes representing certain feature, weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that

node) and averaged over all trees of the ensemble. Another more interesting measure is the *permutation importance* or *mean decrease in accuracy* (MDA), which directly measures the impact of each feature on the accuracy of the model during the out-of-bag error estimate. This measure is computed by permuting the values of each feature and measuring how much this permutation decreases the accuracy of the model. Permuting unimportant variables should have little effect on model accuracy, while permuting important variables should significantly decrease it. Therefore, features with larger MDA values are considered as more important and are ranked first. An advantage of these variable importance rankings provided by the RF modelling is that once they are computed, they can be extracted and tested with another model.

## 4.4. Resampling Techniques

Many complex non-linear models are capable of learning data so well that can correctly predict every sample used in for training the model. However, this is not the goal of building a classifier. The true interest is to accurately predict new samples that were not used when fitting the model. As more flexible is the model it will better fit the data but its generalization to new samples will be very poor. This problem has been previously introduced in some classification models such as ANN, MLP or SVM with non-linear kernels and is commonly known as *overfitting*. To avoid overestimated values, it is always recommended to split the data into non-overlapping training and testing sets so that results on new data can be reported (*hold-out validation*). The testing set will only be used to estimate the classification performance of the model developed with the training set. However, this simple approach may result in *selection bias*, that is, the model may produce overoptimistic results since only a random portion of the dataset that may not be representative of the population is used for testing. If the selection bias is not taken into account, then some conclusions of the study may not be accurate. This problem becomes more relevant when dealing with small sample sizes because in these cases the reliability of the results is highly dependent on the choice for the training/test split (the instances chosen for inclusion in the test set may be too easy or too difficult to classify, thus skewing the results). Furthermore, when training datasets with a considerably reduced number of samples, it is recommended that all available samples contribute to the modelling of the classifier in order to obtain a more generalized model. To address all these issues, *resampling techniques* are usually applied when building a classifier [173]–[175].

Resampling methods such as *cross-validation* (CV) approaches and *bootstrap* can be implemented within the model training process to get good estimates of the performance, generalization and stability of the model using only the training set, without the need to separate a set of samples for testing purposes. Therefore, the entire dataset can be used as the training set. In general, resampling techniques partitions multiple times this training set into a subset of samples for fitting the model and another subset for validating the model. The whole model performance accuracy is then evaluated by averaging the validation results obtained across repetitions. This process allows obtaining non-overfitted results and addresses the selection bias problem by testing diverse subsets of samples, thus giving an idea of the true generalization of the model [173], [174].

Resampling methods can be classified according to the way they partition the data. Hereafter, we present two of the most used resampling techniques and further objectives of these methods so as to improve the reliability of the model performance.

### *4.4.1. Further Applications of Resampling Techniques*

Apart from reporting good estimates of the model performance accuracy, resampling techniques can be used to assess other issues regarding model building so as to get a good sense of how the model works without setting apart a test set.

In first place, it is known that most of the models have one or more intrinsic parameters (or *hyperparameters*) that control their complexity. These parameters have to be properly selected so as to make the most of the models, but they cannot be straightforwardly estimated by only analyzing the input data externally. For example, as previously mentioned, choosing the number of neighbors in KNN is critical since few neighbors may result in overfitting, whereas many features may result in less flexible and sensitive classification. Another example commented before is related to the size of the hidden layers in MLP: few hidden nodes may lead to poor approximation and generalization, while excessive nodes can lead to complex networks prone to overfitting. Resampling techniques allow evaluating repeatedly a group of candidate parameter values within the model building process in order to choose the appropriate values, that is, the ones that maximize the classification accuracy through all the repetitions of the modelling process. This procedure is commonly known as *model tuning* and the hyperparameters to be adjusted are usually referred as *tuning parameters* [173].

Another use of resampling techniques is related to the correct application of the feature selection procedure. When using a feature selection method to select the optimal combination of features in a specific dataset, one could argue that if the same feature selection approach were repeated with a slightly different dataset, the resulting optimal subset of features might change, thus misinterpreting the results and resulting in a poor generalization. This fact can be considered as a methodological error that occurs when feature selection is applied to the entire dataset as a previous step to the model building process and as part of it. Therefore, it is recommended to include the feature selection method within the model building process and implement a resampling technique to validate it. This way, feature selection is repeated for each iteration of the resampling procedure in order to obtain a different optimal group of features for each subset of training samples and capture this variation of feature selection in the results. By using this approach, the chances of overfitting the features will be drastically reduced, especially when dealing with small training sets, at the expenses of a decrease in the computational efficiency of the feature selection process [162], [176].

To assess the two concerns presented above within the resampling procedure, it is necessary to implement two levels of resampling, that is, to apply two nested resampling techniques (i.e., one inside the other) so as to evaluate each issue within the same process but separately. This practice is commonly known as *nested cross-validation* and is organized in two resampling layers: the *inner loop* and the *outer loop*. Usually, the inner loop is used to assess the model tuning while the outer loop is employed for implementing the feature selection method and estimate and compare the generalization performance of several models with the selected features.

### *4.4.2. K-Fold Cross Validation*

One of the most common resampling techniques used due to its accurate performance estimation is the *k-fold cross-validation*. This approach randomly partitions each dataset into $K$ equally (or nearly equally) sized subsets of samples or *folds* accomplishing stratification, that is, maintaining a balanced amount of the involved classes in each fold to ensure each fold is a good representation of the whole. Then, $K$ models are trained and tested so that each of the $K$ folds is used once as the test set, while the $K$-1 remaining folds are used to train the model. A schematic example representing this resampling technique is shown in Figure 4.6. Common values of $K$ are 5 or 10, but there is no formal rule of which $K$ value is better. As $K$ gets larger, the difference in size

between the training set and the folds gets smaller, as well as the difference between the estimated and true values of the performance [173].

A common strategy for reducing the variance of the validation results produced by this resampling technique and avoiding possible selection bias in the random separation of the folds is to repeat several times the partition into $K$ folds. This resampling approach is known as *repeated k-fold cross-validation* and is said to increase the precision of the estimates at the expenses of an increase in computation time. This approach is the most recommended resampling technique in the literature [173].

A particular version of this resampling technique is the *leave-one-out cross-validation* (LOOCV), which is the special case where $K$ is the number of samples. In this case, since only one sample is held-out at a time, the final performance is computed from the $K$ individual held-out predictions. This variant is not recommended for large sample sizes because it requires as many model fits as data samples and each model fit uses a subset that is nearly the same size of the training set, thus becoming very computationally taxing. Additionally, the LOOCV strategy is said to have high variance, thus resulting in overfitting [174].

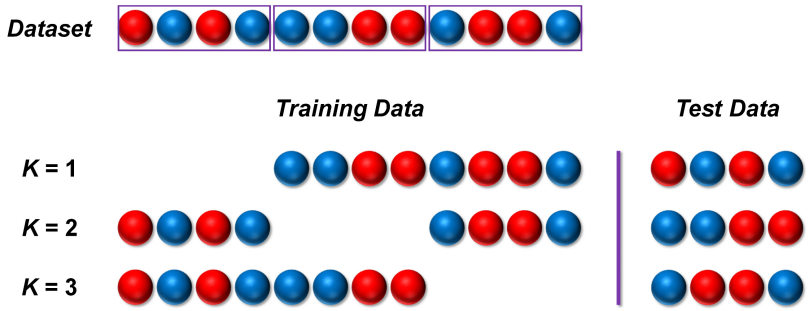



**Figure 4.6.** Scheme of stratified $k$-fold cross-validation with $K = 3$. A dataset with twelve samples belonging to two classes (blue and red spheres) are subdivided in three groups or folds. Each of these folds works as test set in each of the three iterations, thus ensuring that every sample tested once. The remaining two folds in each iteration are used as training set. Performance estimates are calculated from each held-out fold and the average of the three estimates would be the cross-validation estimate of the model performance.

### 4.4.3. Leave-Group-Out Cross-Validation

A conceptually simple resampling technique that also produces stable estimates of the model performance is the *leave-group-out cross-validation* (LGOCV), also known as *Montecarlo cross-validation.* This method randomly divides each dataset into training and test sets *N* times, forming *N* groups. Then, each group is examined independently: the samples of the training set of a group are used to build the model and then this model is evaluated using the samples of the test set of the same group. Finally, the classification results provided by the estimates of all groups are averaged. This approach is quite flexible since the proportion of the data going into each subset (training and test) and the number of repetitions can be controlled [173]. Additionally, as opposed to k-fold cross-validation, LGOCV does not allows controlling the times a specific sample is included for testing, thus increasing the randomness. A simple example of this resampling technique is represented in Figure 4.7.

The LGOCV approach is usually recommended among other techniques because it is a more randomized and generalized strategy and allows reducing the uncertainty (variance) and the bias by choosing a recommended number of repetitions (*N* between 50 and 200) and a recommended percentage of held-out samples (20-25%) [173]. This strategy is also recommended when dealing with datasets with a reduced number of samples [177].

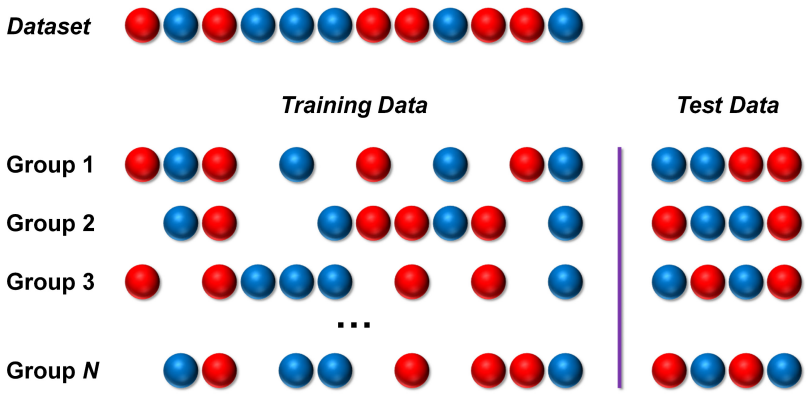



**Figure 4.7.** Scheme of stratified leave-group-out cross-validation with *N* repetitions and 1/3 of the samples used as test set in each repetition. A dataset with twelve samples belonging to two classes (blue and red spheres) is randomly subdivided *N* times, forming *N* groups. A model is built in each repetition using 2/3 of the samples and tested with the remaining 1/3. Samples can be represented in multiple held-out subsets and the times that each sample is included in the test set is different for all samples.

## 4.5. Measures for Evaluating Classification

With the purpose of evaluating the quality of a predictive model and discard models that are not optimal enough to carry out certain classification task, several methods such as metrics or graphical representations are used to facilitate the task of interpreting the results. These methods represent the classification performance of the model when applied to a test set and the corresponding measures can be based on the predicted classes (discrete values) or the class probabilities (continuous values) [178], [179].

### *4.5.1. Measures Based on Predicted Classes*

The most common and simplest method for evaluating and describing the performance of a classifier based on the predicted classes is the *confusion matrix* (CM). The CM is a cross-tabulation of the observed (or true) and predicted classes in a classification task. In essence, the diagonal cells of the matrix represent the number of samples that were correctly classified in each class, while the off-diagonal cells represent the number of samples that were erroneously classified in other class.

Considering a classification task with two classes (positive and negative), the elements of the CM are defined as follows. The first row of the matrix contains the *true positives* or TP on the left (i.e., number of positive samples that were correctly classified as positive) and the *false positives* or FP on the right (i.e., number of negative samples that were incorrectly classified as positive). The second row of the matrix contains the *false negatives* or FN on the left (i.e., number of positive samples that were incorrectly classified as negative) and the *true negatives* or TN on the right (i.e., number of negative samples that were correctly classified as negative). Figure 4.8 shows a CM for a binary classification task representing the latter concepts. The computation of the CM can be also extended to multiclass problems.

Several metrics can be extracted from the CM to assess numerically the performance of the classification. All these metrics values range from 0 to 1, where 1 is the optimal value. Although there are numerous metrics, the most cited in the literature are the following ones:

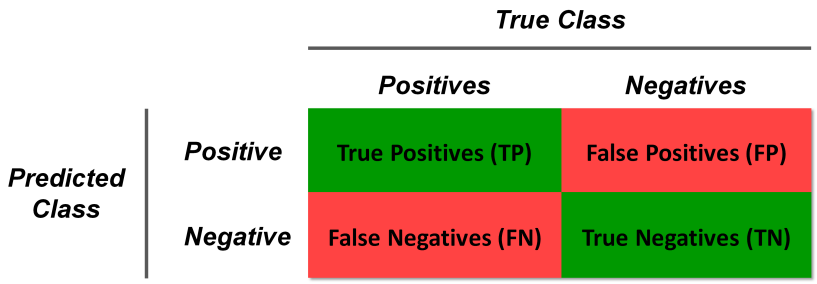



**Figure 4.8.** Confusion matrix for a two-class problem (positives and negatives). The matrix elements indicate the number of the true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN).

- ***Sensitivity***: represents the proportion of all samples belonging to positive class that are correctly predicted as positive. It is also known as the *true positive rate* (TPR).

$$Sensitivity = \frac{TP}{TP + FN} \qquad \textbf{Equation 4.3}$$

- ***Specificity***: represents the proportion of all samples belonging to negative class that are correctly predicted as negative. The *false positive rate* (FPR) is defined as one minus the specificity.

$$Specificity = \frac{TN}{FP + TN} \qquad \textbf{Equation 4.4}$$

- ***Overall Accuracy***: reflects the proportion of true results (both true positives and true negatives) among the total number of cases evaluated.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad \textbf{Equation 4.5}$$

Although the overall accuracy is probably the most used classification measure, it does not take into account class imbalance (i.e., disparities in the class proportions). In order to consider the class distribution of the training set samples, other metrics such as the *Kappa statistic*, also known as the *Cohen's Kappa*, can be used. This metric evaluates the accuracy that would be generated simply by chance and its equation is as follows:

$$Kappa = \frac{O - E}{1 - E}$$ **Equation 4.6**

where $O$ is the observed accuracy and $E$ is the expected accuracy based on the marginal totals of the CM. A Kappa value of 1 indicates perfect concordance of the model prediction and the observed class while a value of 0 indicates no concordance. However, in some cases Kappa values within 0.30 to 0.50 may indicate reasonable agreement [178].

### *4.5.2. Measures Based on Class Probabilities*

Apart from providing a predicted class for each sample, many classifiers also generate continuous predictions when evaluating the model with the test set that are in the form of a probability for each class. In some cases, these class probabilities obtained for each predicted value may offer more information about the model predictive capacity than the simple class value. In a binary classification task with positive and negative classes, a probability threshold value may be established for the vector of probabilities of the positive class to define a boundary between classes so that probabilities above this threshold will be classified as positive and probabilities below this threshold will be classified as negative. Many classification models usually take the probability of 0.5 as threshold but, however, in some applications it may be of interest to set a different value.

The *receiver operating characteristic* (ROC) curve is the most common method to evaluate graphically the confidence of the classification model with class probabilities. It is constructed by evaluating the class probabilities across different thresholds and recomputing the corresponding predicted classes. For each threshold, the TPR (sensitivity) and FPR (one minus specificity) are computed and plotted against each other, thus resulting in a curve that reflects the variations that sensitivity and specificity suffer depending on the chosen threshold. Therefore, one can choose a specific threshold (usually referred as *cutoff point*) so as to obtain better sensitivity or specificity depending on the purpose of the application, although the common approach is to choose the threshold that maximizes both measures equally. Figure 4.9 shows an example ROC curve representing two cutoff points that provide different sensitivity and specificity measure. A model that perfectly separates two classes would have unity sensitivity and specificity and this will be represented as a single step between (0, 0) and (0,1) and constant from (0, 1) to (1, 1) on the ROC curve.

The *area under the curve* (AUC) is a way to explain and evaluate the ROC curve with a single numerical value representing the overall expected performance of a classifier. Defining this measure in a simple and understandable way is somewhat difficult. In terms of probabilities, the AUC describes the probability that, the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [179]. A perfect classifier will have AUC = 1 whereas a completely ineffective classifier will have AUC = 0.5, which is plotted as a diagonal line. The ROC curve has the advantage to be insensitive to class imbalance since it is a function of sensitivity and specificity, so it is recommended to report classification results using the AUC instead of the overall accuracy, especially when dealing with imbalanced datasets, because the accuracy is mostly influenced by the majority class [178].

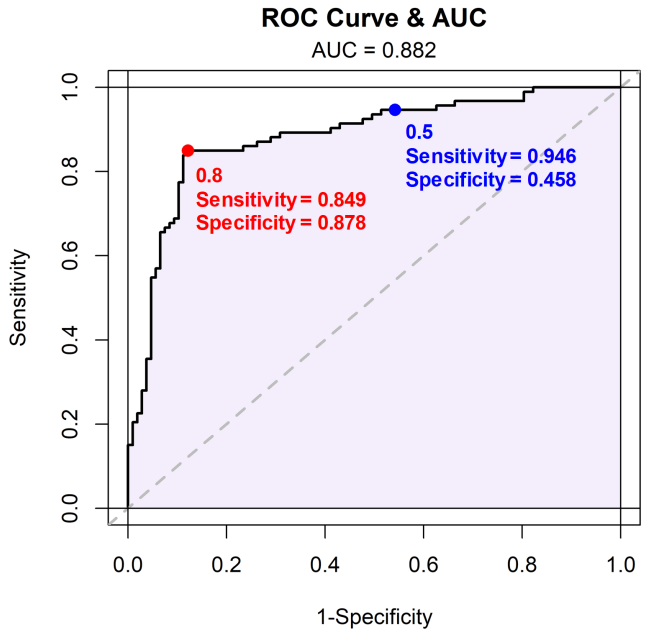



**Figure 4.9.** A receiver operating characteristic (ROC) curve for a given data. The blue point indicates the value corresponding to a probability threshold of 50%, which gives better sensitivity, while the red point corresponds to a threshold of 80%, which maximizes both sensitivity and specificity at the same time. The zone in light purple represents the area under the ROC curve (AUC).

# Chapter 5.
# Differentiation between brain metastases and glioblastomas

## 5.1. Introduction and Motivation

Primary glioblastoma multiforme (GBM) and intracranial brain metastasis (BM) are the two more common malignant brain tumors in adults.

Glioblastomas are the most aggressive diffuse gliomas and are defined as grade IV cancer according to the World Health Organization (WHO) classification of tumors of the central nervous system (CNS) [20], thus meaning that these types of brain tumors have a high level of malignancy and tend to grow and spread rapidly. GBM is the most common malignant primary brain and CNS tumor. According to recent studies, GBMs are estimated to account for 56% of all gliomas, 47% of malignant primary brain and CNS tumors and 15% of all primary and CNS tumors approximately [21]. GBM has a very poor prognosis, with less than 5% of patients surviving 5 years past diagnosis and a median survival of 15-18 months for patients with newly diagnosed GBM [180], [181].

Brain metastases are the most common tumors of the CNS in adults, even more than primary brain tumors [22], [24], [182]. The prognosis of patients diagnosed with metastatic brain tumors is poor: the median survival of these patients is estimated to be limited to months even for patients under treatment [183], [184]. The exact incidence of BM is certainly unknown, although some studies reported that these lesions occur in 9–17% of patients with cancer [23], [185]. However, these rates are currently increasing

due to improved imaging techniques for diagnosis and prolonged survival from primary cancers, among other reasons [23].

The clinical symptoms associated with GBM and BM may be similar, but each brain tumor type has a different biologic nature, so different treatment strategies are required for each tumor type. Therefore, differentiating these types of brain tumors rapidly is crucial [186]. The current standard for treating newly diagnosed GBMs implies a maximal safe surgical resection followed by a combination of radiotherapy and chemotherapy [187], [188]. Conversely, BMs are mainly treated with stereotactic radiosurgery (SRS) and/or whole-brain radiotherapy (WBRT) depending on the number of metastatic lesions, among other reasons [189], [190]

In some cases, a correct diagnosis can be achieved by combining the clinical information of the patient and the radiological information provided by conventional structural MRI. Generally, a GBM appears as a solitary big tumor mass in the brain. In contrast, a patient with BM usually tends to have multiple metastatic brain lesions and the primary cancer is commonly known. However, there are exceptions to these rules (e.g., multifocal GBM, solitary BM or unknown primary cancer) and additionally, GBM and BM may present similar appearance in conventional MR images [191]–[194], as shown in Figure 5.1. For example, GBMs frequently appear as ring-enhancing lesions with a predominant necrosis area in structural MRI, but this appearance is not specific and some BMs can also show a similar radiologic profile [195]–[197]. Therefore, a definitive diagnosis cannot be made exclusively with this information.

Nowadays, the histopathologic analysis of a sample of the tumor region (i.e., biopsy) is the only accepted way to make a definitive diagnosis [47], [198]. However, the use of non-invasive, fast methods to identify correctly the type of lesion would be preferable. Furthermore, invasive procedures are not allowed in those cases when there exists an important risk for the patient (e.g., when the mass is located near an eloquent area), so in these cases, non-invasive procedures are mandatory [191]. In this context, several studies tried to classify both types of tumors by using advanced MRI techniques (e.g., MR spectroscopy or diffusion and perfusion weighted imaging) and multi-parametric imaging analysis [192]–[194], [199]–[201]. Despite the promising results obtained, the MR techniques proposed in these studies are not commonly performed when screening the tumor for the first time and require long acquisition times, so these analyses may be quite demanding and time-consuming, among other disadvantages [202]. Other studies investigated other imaging techniques like positron emission tomography (PET) to this end, obtaining good results [203]. However, this technique in

particular exposes the patient to ionizing radiation as opposed to MRI, and for this reason, MRI would be the preferable imaging technique for screening the brain tumor. To this effect, it would be desirable to have an efficient approach for differentiating the type of brain lesion (GBM and BM) based on the conventional MR images obtained in the first screening of the tumor mass in order to accelerate the process of identification. With this approach, the patient would not be exposed to additional imaging, nor invasive or ionizing procedures.

To accomplish the objective of classifying BM from GBM using conventional MRI, morphometric analysis could be of help, like showed in [191]. However, morphometric features do not take advantage of all the information that MR images may provide about the tumor lesion. Texture analysis is a powerful methodology that processes the images to extract a wide variety of features that define the properties of the MR image region under analysis in different ways. In contrast to morphometric features that only define the shape of the tumor region, texture features describe the intrinsic heterogeneous characteristics of the lesions, thus making possible the characterization of different tumor lesions based on the information extracted from the MR images. Several studies attempted to examine the differentiation of several brain tumor types, including GBM and BM, by means of texture analysis [104], [105], [204]. The results obtained in these studies were promising, and we took them as a basis for our work. However, although these studies considered GBM and BM, they also included other brain tumor types like meningiomas or WHO grade II or III gliomas in their analyses, thus interfering in the results of classification between GBM and BM. In the present work, we only focused on the discrimination between GBM and BM since these two types of brain tumors are the most common aggressive tumors in adults and they can be radiologically confused, as explained before. We additionally increased the number of features, tested different texture analysis methods and used other classification approaches to reduce the possibilities of achieving overfitted results.

The main objective of this project was to evaluate the potential of 2D texture features extracted from T1-weighted MR images for differentiating between BM and GBM by using machine learning techniques. With this approach, we proposed a decision support tool to help in the diagnosis of the lesion in a fast, non-invasive way by analyzing the conventional MR images used for detecting the brain tumor for the first time. We additionally analyzed the influence of the number of gray levels used to quantize the images and the performance of different classifiers and feature selection methods.

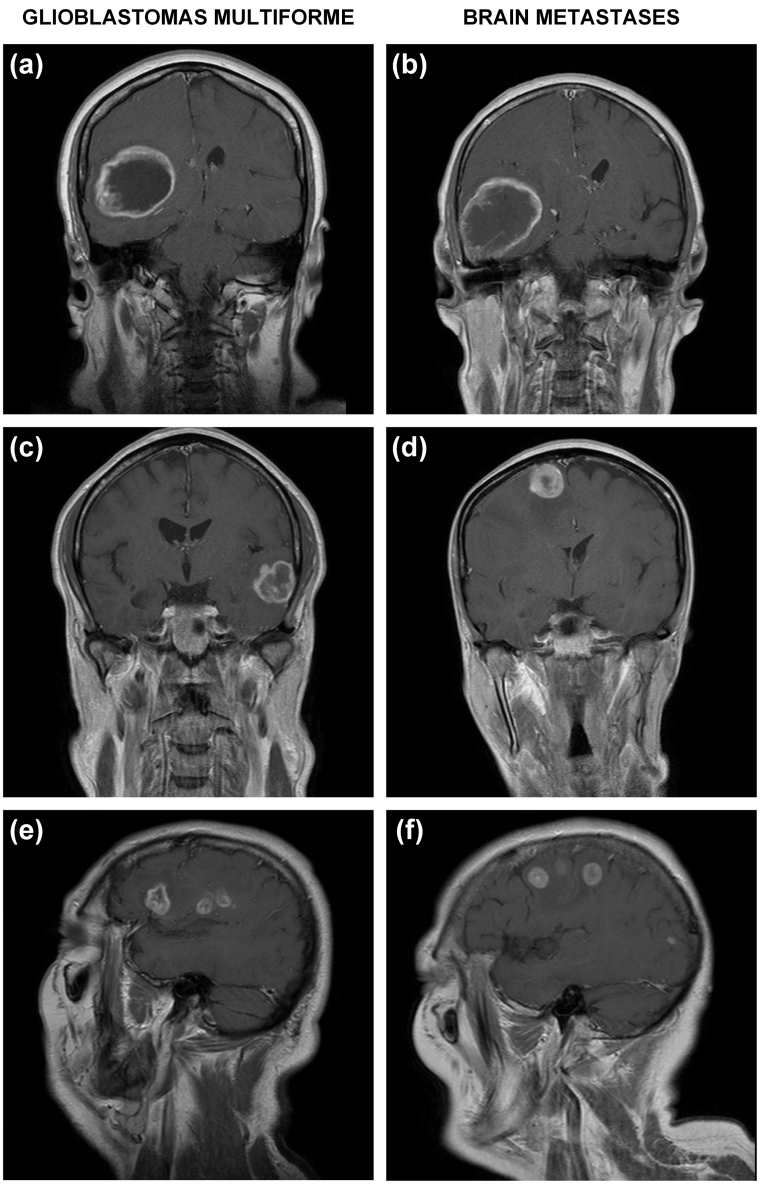



**Figure 5.1.** Examples of contrast-enhanced T1-weighted MRI scans of six different patients with GBM (left column) and BM (right column) where the lesions present similar appearances and cannot be distinguished. Patients (a) and (b) present a big ring-enhancing lesion, more typical of GBM. Patients (c) and (d) show a small solitary mass. Patients (e) and (f) have multiple lesions, typical of BM.

# 5.2. Material and Methods

## *5.2.1. Patients and Imaging Protocol*

This retrospective, single-center study was approved by the Institutional Review Board of the Hospital Universitario de La Ribera and all subjects provided written informed consent. Patients presenting single or multiple GBMs or BMs were reviewed by an expert neuroradiologist with 20 years-experience. Inclusion criteria for BMs covered: (1) pathologically confirmed primary cancer; (2) only one single primary tumor; (3) no previous treatment, biopsy or surgical resection on BM and (4) all BMs confirmed by imaging and clinical follow-up. Exclusion criteria for BM included: (1) small metastases (longest diameter < 10 mm) as texture analysis cannot capture texture information properly in small regions [13]; (2) more than one BM per patient and (3) lesion affected by MRI artifacts like motion or truncation (Gibbs phenomenon) artifacts. Inclusion criteria for GBMs comprised: (1) no previous treatment, biopsy or surgical resection on GBM and (2) all GBMs confirmed by imaging or histopathology and clinical follow-up. Exclusion criteria for GBM included: (1) more than one GBM per patient and (2) lesion affected by MRI artifacts. The first one hundred patients (63 men and 37 women, mean age 60.98 years, age range 24–86 years) who complied with inclusion criteria and not with exclusion criteria and selected between December 2010 and January 2017 were included. Since we only included one lesion per patient, one hundred baseline lesions were collected: 50 GBMs and 50 BMs. The distribution of primary known cancers for patients with BM was: lung cancer (38), renal cancer (4), breast cancer (3), colorectal cancer (3), thyroid cancer (1) and melanoma (1). The properties of the study group are shown in Table 5.1.

Imaging was conducted using a 1.5T MRI scanner (Philips Achieva 1.5T; Philips Healthcare, Best, The Netherlands) and an eight-channel sensitivity encoding (SENSE) head coil. The MRI protocol included two-dimensional turbo spin echo (TSE) coronal T1-weighted brain images. Images were acquired without magnetization transfer, after intravenous administration of a single-dose of gadobutrol (1 mmol/ml, Gadovist; Bayer Schering Pharma, Berlin, Germany). All the GBMs and BMs were scanned using the same imaging parameters since changing these parameters may influence texture analysis performance [95], [111]: repetition time/echo time (TR/TE) of 500/20 ms; flip angle of 90º; matrix size of 512×512; pixel size of 0.43×0.43 $mm^2$; and slice thickness of 7.3 mm. Some examples of the contrast-enhanced T1-weighted images of patients with GBM and BM used in this work can be seen in Figure 5.1.

**Table 5.1.** Baseline characteristics of the study group per class

| | GBM | BM | Total |
|---|---|---|---|
| Number of samples | 50 | 50 | 100 |
| Number of patients | 50 | 50 | 100 |
| Age* (years) | 61.96 ± 13.68 | 60.00 ± 11.58 | 60.98 ± 12.64 |
| Sex (Male/Female) | 26 M / 24 F | 37 M / 13 F | 63 M / 37 F |

* Continuous variables are expressed as mean ± SD

### *5.2.2. Regions of Interest*

To perform the segmentation of each lesion in 2D, the coronal slice of the T1-weighted image showing the most representative lesion area was selected and segmented by an expert neuroradiologist (20 years-experience). The rest of the slices showing lesion tissue were not segmented since 3D texture analysis was not conducted. We decided not to apply 3D texture analysis because the slice thickness of our images (7.3 mm) was very large in comparison to the in-plane resolution (0.43 mm) and in this case it is recommended to perform 2D texture analysis [51]. For considering 3D texture analysis, isotropic image resolution is required to ensure the conservation of scales and directions in all three dimensions, thus meaning that image interpolation should be applied in the third dimension and then the image would be highly transformed due to the big difference between the inter-slice and inter-pixel distances [51], [64].

Prior to the feature extraction process, we verified that there was no statistical difference between the size of the lesions in the two groups (GBM and BM), since some texture features may be influenced by the region of interest (ROI) size. In this case, it is important to ensure this fact because GBMs tend to be much larger than BMs and falsely optimistic results may be achieved due to this difference in size and its influence on the texture outcome [13], [62]. We evaluated the distribution of the ROI areas and the ROI longest diameters for both groups of lesions with the Mann-Whitney-Wilcoxon (MWW) test for independent samples. We did not find statistically significant differences when evaluating the ROI areas (GBM: 830.14 ± 420.84 $mm^2$; BM: 711.31 ± 453.22 $mm^2$; $p > 0.05$, $p = 0.088$) and the ROI longest diameters (GBM: 36.93 ± 9.73 mm; BM: 33.46 ± 10.50 mm; $p > 0.05$, $p = 0.065$) of both groups, so the size of the ROIs should not influence in the texture analysis.

Some preprocessing techniques were applied to improve texture discrimination. Firstly, normalization was applied to the MRI regions using the $\mu \pm 3\sigma$ method ($\mu$ is the mean value of the gray levels and $\sigma$ is the SD) to enhance the differences between classes, as proposed by Collewet *et al*. [72]. This method adjusts the histogram of the MRI region to $\mu \pm 3\sigma$ by rejecting the pixels with intensities out of this range.

Quantization of gray levels was also applied to the MRI regions to improve the signal-to-noise (SNR) ratio of the texture outcome [74]. This process refers to the reduction of levels of gray used to represent the image, which is originally represented by 4096 gray levels (12 bits per pixel). In particular, different number of gray levels (NGL) were tested (8, 16, 32, 64 and 128) to study the influence of the quantization process in the discriminative power of the matrix-based texture features.

The delineation and preprocessing of the ROIs in 2D was performed using a software tool developed specifically for this study in MATLAB (R2015b; The MathWorks Inc., Natick, MA, USA).

### *5.2.3. Feature Extraction*

The feature extraction process was conducted using the *Radiomics* MATLAB package implemented by Vallieres *et al*. [73] and the MATLAB implementation of the local binary pattern (LBP) operator provided by Ojala *et al.* [83]. A total of 88 texture-based features were computed for each lesion. These features derived from six different statistical methods: intensity histogram with 100 bins (6 features), gray-level co-occurrence matrix (GLCM) (9 features), gray-level run-length matrix (GLRLM) (13 features), gray-level size-zone matrix (GLSZM) (13 features), neighborhood gray-tone difference matrix (NGTDM) (5 features) and LBP (42 features). Table 5.2 summarizes the features used in this study.

When calculating the matrix-based textures, only one GLCM, GLRLM, GLSZM and NGTDM per lesion was computed with the *Radiomics* package. The GLCM and NGTDM features are originally dependent on the distance of the neighboring pixels, so only adjacent connected pixels (according to [87], distance $d = 1$) were counted when computing these matrices (8 pixel-connectivity). Additionally, the GLCM and GLRLM features are originally dependent on direction, so the neighboring properties in the four possible directions in 2D (0, 45, 90 and 135°) were summed when computing these matrices to achieve rotation invariant features. To account for discretization length differences, neighbors at a distance of $\sqrt{2}$ pixels (45º and 135º) around a center pixel

incremented the matrix by a value of $\sqrt{2}$, and neighbors at a distance of 1 pixel (0º and 90º) around a center pixel incremented the matrix by a value of 1 [73].

When computing the LBP textures, the original LBP operator (patches of 3 × 3 pixels: neighboring pixels $P = 8$, distance $R = 1$) was used to preserve the texture analysis as local as possible. Rotation invariance was achieved by performing a circular bit-wise right shift operation (rotating the neighboring pixel set clockwise) and assigning the smallest LBP binary number [83]. Using this approach, 36 unique rotation invariant histogram-based LBP features (LBP bins) were obtained, as only 36 LBP binary numbers can occur for $P = 8$. Moreover, six histogram-based measures (mean, variance, skewness, kurtosis, energy and entropy) were calculated from the LBP image and added to the LBP feature set. The MR images were not quantized to compute the LBP features since the rotation invariant LBP approach is robust to intensity variations [94].

**Table 5.2.** List of the 88 texture features used in this study.

| **Method** | **Features** | **Number of features** |
|---|---|---|
| Histogram | Mean, Variance, Skewness, Kurtosis, Energy and Entropy | 6 |
| GLCM | Energy, Contrast, Correlation, Homogeneity, Variance, Entropy, Sum Average, Dissimilarity and Autocorrelation | 9 |
| GLRLM | Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-level Non-uniformity (GLN), Run-Length Non-uniformity (RLN), Run Percentage (RP), Low Gray-level Run Emphasis (LGRE), High Gray-level Run Emphasis (HGRE), Short Run Low Gray-level Emphasis (SRLGE), Short Run High Gray-level Emphasis (SRHGE), Long Run Low Gray-level Emphasis (LRLGE), Long Run High Gray-level Emphasis (LRHGE), Gray-level Variance (GLV) and Run-Length Variance (RLV) | 13 |
| GLSZM | Small Zone Emphasis (SZE), Large Zone Emphasis (LZE), Gray-level Non-uniformity (GLN), Zone-Size Non-uniformity (ZSN), Zone Percentage (ZP), Low Gray-level Zone Emphasis (LGZE), High Gray-level Zone Emphasis (HGZE), Small Zone Low Gray-level Emphasis (SZLGE), Small Zone High Gray-level Emphasis (SZHGE), Large Zone Low Gray-level Emphasis (LZLGE), Large Zone High Gray-level Emphasis (LZHGE), Gray-level Variance (GLV) and Zone-Size Variance (ZSV) | 13 |
| NGTDM | Coarseness, Contrast, Busyness, Complexity and Strength | 5 |
| LBP | LBP histogram bins: $LBP_1$, $LBP_2$, $LBP_3$, … , $LBP_{36}$<br>LBP image statistics: Mean, Variance, Skewness, Kurtosis, Energy and Entropy | 42 |

Apart from extracting the above-mentioned features from the original T1-weighted scans, we decided to apply the discrete wavelet transform (DWT) to the original image and extract the same 88 features from the four subimages yielded after the first DWT decomposition, that is, the four matrices of coefficients that represent the approximations or low frequencies (LL: low-low) and the details or high frequencies in the vertical (LH: low-high), horizontal (HL: high-low) and diagonal (HH: high-high) directions. The Haar family of wavelets was used to perform the DWT decomposition.

At the end, 25 different datasets of texture features were obtained: five datasets, one per NGL, extracted from the original 2D regions and five datasets, one per NGL, from the four DWT images. As 88 features were computed for each ROI, a total of 2200 texture-based features (88 × 25) were extracted. All features of the 25 datasets were standardized to zero mean and unit variance to improve numerical stability when computing the model and to avoid model building being affected by the differences in the feature scales [205]. Initially, each dataset contained all 88 features, but some of these features were eliminated before applying machine learning. Firstly, those features that were not reproducible were removed because the segmentation of the ROIs was based on a manual segmentation. To evaluate the reproducibility of the features, the lesions were segmented by two different observers: the first observer was the expert neuroradiologist (20 years-experience) that originally performed the segmentations and the second observer was an expert in image processing and segmentation (5 years-experience) trained to segment BM. Then, we calculated the intraclass correlation coefficient (ICC) for all the features extracted from the 2D ROIs computed by both observers and those features with ICC < 0.75 were eliminated. This process led to the exclusion of 750 texture features, with all the excluded features corresponding to features extracted from the wavelet subimages. Table 5.3 shows the number of features per dataset that remained after checking their reproducibility. Finally, zero-variance and near-zero-variance predictors were excluded from the model training process since they are uninformative and may cause the model to crash or the fit to be unstable [205].

**Table 5.3.** Number of texture features in each of the 25 datasets after eliminating the features that were not reproducible (ICC < 0.75)

| | NGL=8 | NGL=16 | NGL=32 | NGL=64 | NGL=128 | *Number of features* |
|---|---|---|---|---|---|---|
| **Original image** | 88 | 88 | 88 | 88 | 88 | *440* |
| **Subimage LL** | 59 | 59 | 59 | 59 | 59 | *295* |
| **Subimage HL** | 49 | 50 | 51 | 49 | 46 | *245* |
| **Subimage LH** | 47 | 47 | 51 | 50 | 48 | *243* |
| **Subimage HH** | 37 | 46 | 49 | 46 | 49 | *227* |
| | | | | | | ***1450*** |

### *5.2.4. Classification Performance and Evaluation*

In the present work, two groups of patients were studied (GBM and BM), so for our study a binary classification approach was necessary. To this end, five different predictive models from different families of classifiers were tested: random forest (RF), support vector machine (SVM) with linear kernel, k-nearest neighbors (KNN), naïve Bayes classifier (NB) and multilayer perceptron (MLP). These models were chosen due to their well-known performance in application to other datasets [144]. The following hyperparameters were evaluated for each classifier:

- NB: a Gaussian kernel for estimating the probability density function was selected and any additional parameter was tuned.
- KNN: the Euclidean distance was chosen and the number of neighbors ($k$) was selected from $k \in \{1, 3, 5, ..., 13, 15\}$ in the parameter tuning process.
- RF: the number of trees (*ntree*) was set to *ntree* = 250 and the number of random variables used as candidates at each split (*mtry*) was chosen from *mtry* ∈ {2, 3, 4, ..., 11, 12} in the parameter tuning process.
- SVM: a linear kernel was evaluated and the cost parameter ($C$) was chosen from $C \in \{2^{-3}, ... ,2^{0}, ... ,2^{3}\}$ in the parameter tuning process.
- MLP: a single hidden layer was chosen and the number of neurons or nodes in the hidden layer (*size*) was selected from *size* ∈ {3, 6, 9, 12, 15} in the parameter tuning process.

To evaluate these predictive models, we decided to implement a nested cross-validation (CV) scheme, recommended when the sample size is not large enough and all the samples are needed for training and evaluating the models [173]. The structure of the nested CV used to evaluate each model is shown in Figure 5.2.

A 5-fold CV approach was implemented in the outer loop, which was used to optimize the number of features and to cross-validate the models. This resampling method randomly partitions each texture dataset into $K = 5$ equally sized subsets of samples or folds, maintaining a balanced amount of both classes in each fold. Then, five models are trained and tested so that each of the five folds is used once as the test set, while the four remaining folds are used to train the model. This process was repeated $R = 10$ times to reduce the variance of the cross validation results and to avoid possible bias in the random separation of the folds [173]. At the end 50 models ($K \times R$) were built using different sets of patients for training and testing each time. A 10-fold CV without repetitions was applied in the inner loop, and it was used to execute the hyperparameter tuning process.

The optimal number of features was evaluated in the outer loop by means of a feature selection approach. Three filter feature selection methods were employed in this work to generate a ranking of features based on some statistic and then their effectiveness was compared. In first place, we tested a simple feature selection method based on the *p*-value provided by the Mann-Whitney-Wilcoxon (MWW) test for independent groups of samples. This method was then compared with two other filter methods: the first method was based on the maximal information coefficient (MIC), which measures the strength of the linear or non-linear association between two variables and the second method was based on the ReliefF algorithm. The feature selection step was included within the model-building process to avoid overfitting [176].

The classification performance was evaluated using the average area under the curve (AUC) of the receiver operating characteristic (ROC) that resulted from averaging the AUC values obtained from the 50 iterations (mean ± SD). Good estimates of the model performance can be obtained using the validation data when the sample size is not large [173]. Assuming that *R* represents the number of repetitions and *K* represents the number of folds per repetition, the average AUC for each subset of features is calculated with the following equation:

$$AUC_{average} = \frac{\sum_{f=1}^{K} \sum_{r=1}^{R} AUC_{k,r}}{K \times R} \qquad \textbf{Equation 5.1}$$

The AUC results obtained in this training process for each of the subsets of features were summarized in a graphic called "profile curve", which represents the evolution of the average AUC as the size of the subset of features increases. Other metrics like sensitivity, specificity and accuracy were also obtained to validate the results.

The model evaluation process was implemented with the Caret package [143] in R language, version 3.2.5 (R Development Core Team, Vienna, Austria).

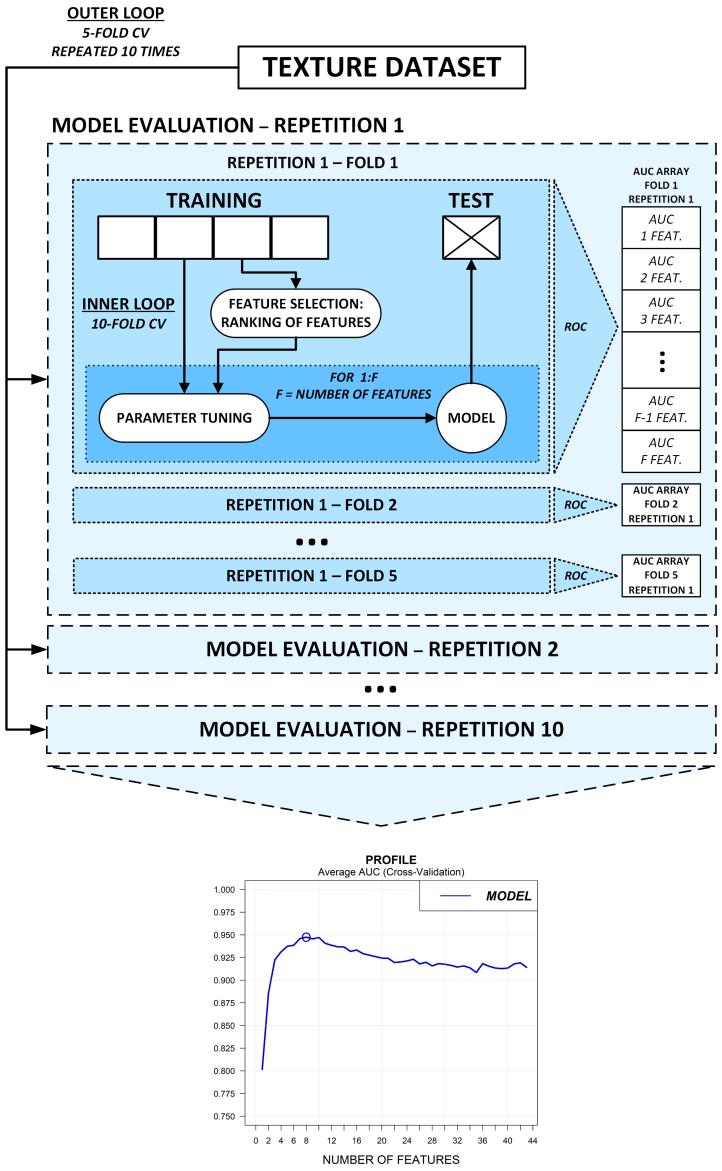



**Figure 5.2.** Structure of the nested CV method used to evaluate the different datasets of features. All the samples of each texture dataset were randomly separated R = 10 times in F = 5 folds to evaluate the model with the average AUC, examining different subsets of features.

## 5.3. Results

### *5.3.1. Influence of the Wavelet Decomposition*

The 25 different datasets were first analyzed with all the classifiers and with the feature selection method based on the *p*-value. The results indicated that the classification improved when no DWT was applied to the original image. Although the texture datasets extracted from the DWT provided good classification results in some cases (AUC > 0.8), the results always improved when employing the datasets extracted from the original images, for all the classifiers and for all NGL, as shown in Figure 5.3, Figure 5.4 and Figure 5.5.

The influence of the DWT was especially obvious when using the MLP classifier (Figure 5.5). In this case, the datasets of textures extracted from the four DWT subimages only achieved AUC < 0.8 for all NGL, while the datasets of textures extracted from the original image reached AUC > 0.85 for all NGL. Furthermore, the differences between the AUC results obtained with textures from the original image and the AUC results obtained with textures from the DWT subimages were statistically significant for all NGL when using the MLP model (MWW test, $p < 0.05$). When analyzing the performance of the other classifiers, these mentioned differences were subtler, especially for NB and KNN models (Figure 5.3).

The textures obtained from the HL subimage (horizontal details) provided AUC results very similar to those obtained with the original image when quantizing with NGL = 8 gray levels and using RF and KNN models. However, in general, as the number of gray levels increased, the differences between the AUC results were more pronounced for all DWT subimages and all classifiers.

In the following subsections, we are going to focus only on the texture datasets extracted from the original image, since applying DWT did not improve the results.

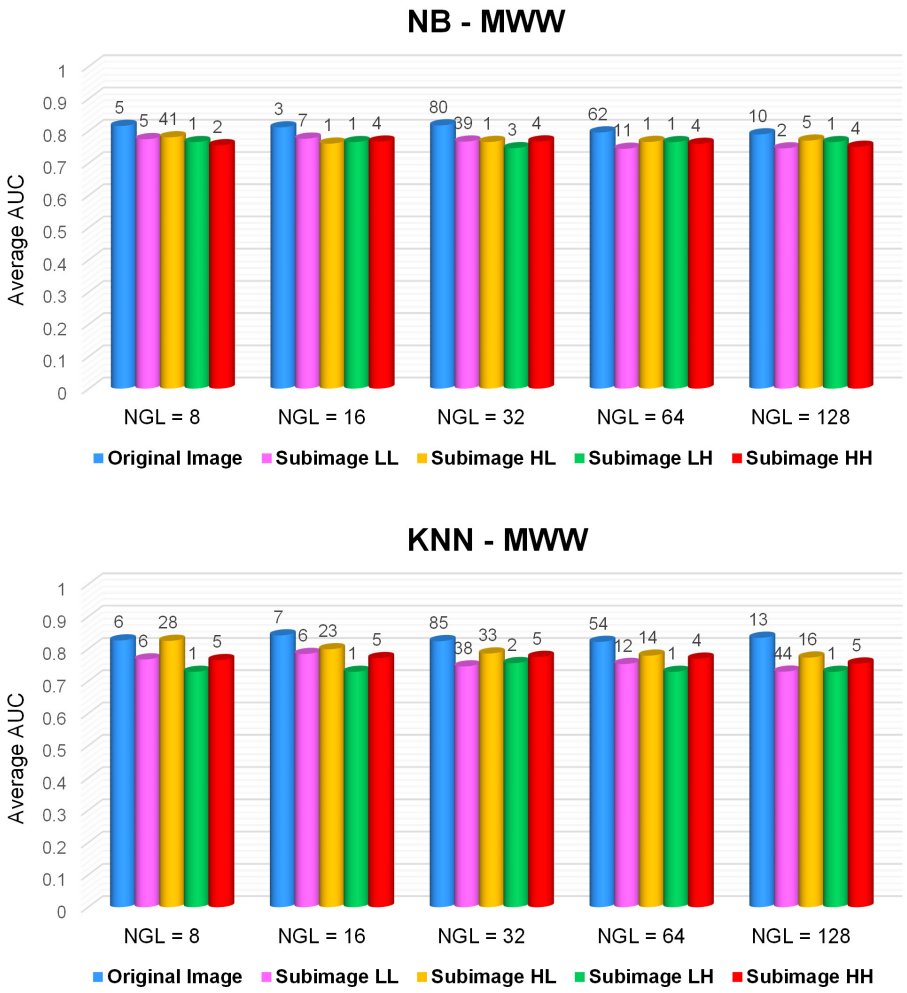



**Figure 5.3.** Average AUC results obtained for the 25 datasets when using the feature selection method based on the *p*-value and NB and KNN models. The numbers situated on the bars indicate the number of features used to achieve the maximum AUC. Notice that in most of the cases, texture features extracted from the original images provided better AUC results than features from the DWT subimages.

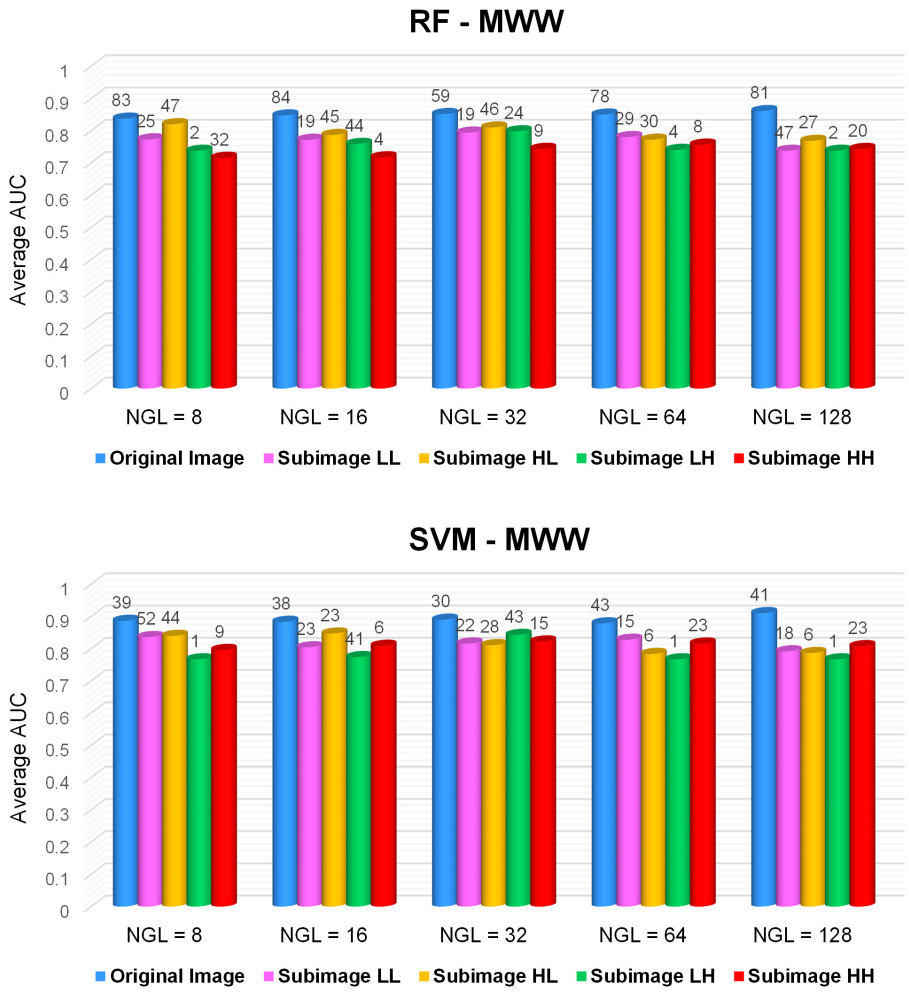



**Figure 5.4.** Average AUC results obtained for the 25 datasets when using the feature selection method based on the *p*-value and RF and SVM models. The numbers situated on the bars indicate the number of features used to achieve the maximum AUC. Notice that in most of the cases, texture features extracted from the original images provided better AUC results than features from the DWT subimages.

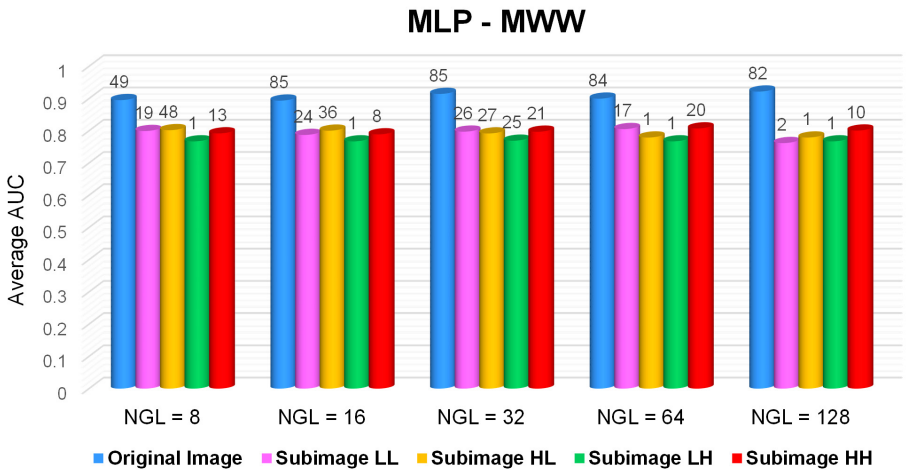



**Figure 5.5.** Average AUC results obtained for the 25 datasets when using the feature selection method based on the *p*-value and MLP model. The numbers situated on the bars indicate the number of features used to achieve the maximum AUC. Notice that in all cases, texture features extracted from the original images provided substantial better AUC results than features from the DWT subimages.

### *5.3.2. Influence of the Quantization Process and the Classifier Choice*

Depending on the classifier employed, the influence of the number of gray levels used to quantize the images was more or less subtle. However, according to the results summarized in Figure 5.6, quantizing with different number of gray levels caused that the classification results differed for all models, so matrix-based texture features were affected by this quantization process. The optimal number of gray levels for MLP, SVM and RF models was NGL = 128, but for KNN and NB classifiers the best AUC values were achieved for NGL = 16 and NGL = 32 respectively. These last results indicate that different classifiers respond differently when modifying the quantization level.

Regarding the best classifier, MLP model outperformed the rest of the models for all NGL. The best AUC result was achieved for the MLP model and the dataset of textures from images quantized with NGL = 128 (AUC = 0.912 ± 0.060, 82 features). When using the same model with textures from images quantized with NGL = 32, the results were also good (AUC = 0.904 ± 0.065, 85 features), with no statistical significance between these results and those obtained with NGL = 128 (paired *t*-test, *p* =

0.269, $p > 0.05$). However, we decided to choose the dataset that provided the highest results (NGL = 128) for the subsequent analyses since the optimal number of features was similar in both cases and the computational cost of the corresponding features was similar for both levels of quantization. Good AUC results were also achieved when using the SVM model and the dataset of textures from images quantized with NGL = 128 (AUC = 0.901 ± 0.066, 41 features), with no statistical significance between these results and those obtained with the MLP model (paired *t*-test, $p = 0.102$, $p > 0.05$). In this case, the results were considered important because the number of features used to obtain the highest AUC value was halved, thus reducing notably the computational cost.

Figure 5.7 shows the profiles curves and the ROC curves obtained for the five models under analysis when using the dataset of features extracted from images quantized with NGL = 128. Moreover, Table 5.4 shows additional metrics obtained for the two best models (SVM and MLP).

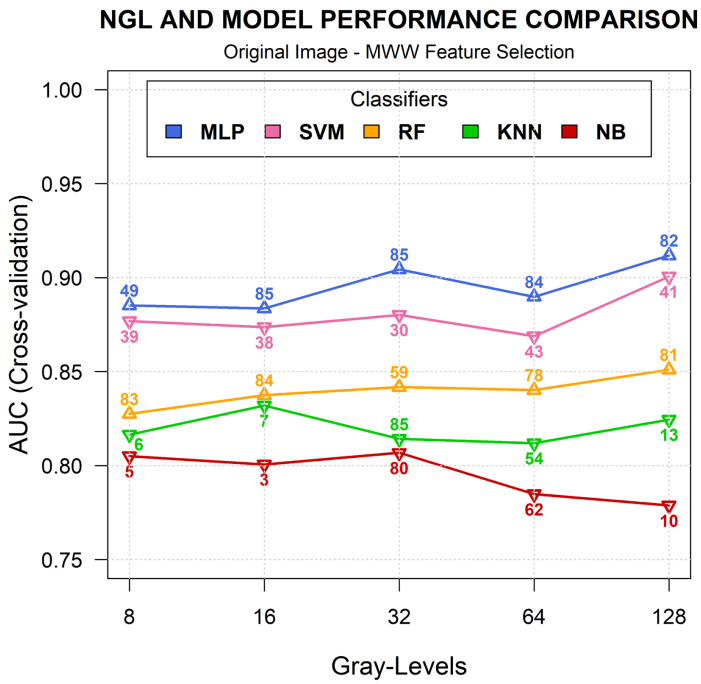



**Figure 5.6.** Comparison between the AUC results obtained for the datasets of features extracted from the original images quantized with five different NGL when applying the five models tested in this study with the *p*-value feature selection approach. The numbers on the curves indicate the number of features used to achieve the maximum AUC.

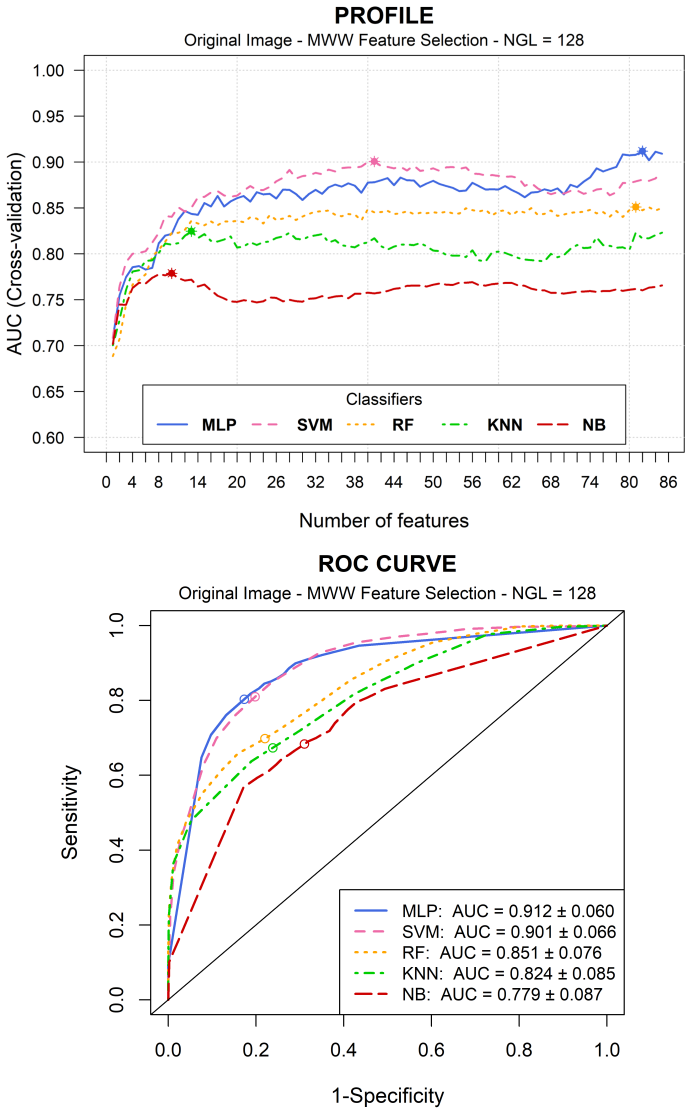



**Figure 5.7.** Profile curves and the corresponding average ROC curves for the best subset of features obtained with the five classifiers and the *p*-value feature selection method for the dataset of features extracted from the original image quantized with NGL = 128 gray levels. The highlighted points on the profile curves indicate the optimal subset of features. The highlighted points on the ROC curves indicate the optimal cutoff points that weighs both sensitivity and specificity equally computed with the "closest-to-(0,1)" criterion.

**Table 5.4.** Additional metrics obtained when using the two best models (MLP and SVM) and the *p*-value feature selection method on the dataset of features from images quantized with NGL = 128 gray levels.

| Classifier | MLP | SVM |
|---|---|---|
| Number of Features | 82 | 41 |
| AUC | 0.912 ± 0.060 | 0.901 ± 0.066 |
| Sensitivity [a] | 0.806 ± 0.135 | 0.810 ± 0.146 |
| Specificity [a] | 0.834 ± 0.142 | 0.800 ± 0.125 |
| Overall Accuracy | 0.818 ± 0.099 | 0.805 ± 0.089 |
| Kappa Index | 0.636 ± 0.199 | 0.610 ± 0.177 |

* Values are shown as mean ± SD as a result over groups' estimates.
[a] Sensitivity and specificity were computed according to the optimal cutoff point of the ROC curve computed with the "closest-to-(0,1)" criterion.

### *5.3.3. Influence of the Feature Selection Method*

Besides using the *p*-value feature selection method, we decided to test other filter feature selection methods (MIC and ReliefF algorithm) to check if ranking the features according to different statistical criteria improved the results of classification and reduced the optimal number of features. For this experiment, we used the dataset of textures extracted from the original image quantized with NGL = 128 gray levels and we tested the two best classification models according to the previous analyses: MLP and SVM classifiers.

According to the results presented in Table 5.5, changing the feature selection method did not improve significantly the AUC values. When using the MLP model, the AUC results were very similar, without statistical significance (paired *t*-test, $p > 0.05$), and the optimal number of features did not decrease. Regarding the SVM model, the AUC results were reduced when using the other feature selection methods but without statistical significance (paired *t*-test, $p > 0.05$). However, in this case, the optimal number of features to achieve the best AUC values increased notably when using other feature selection methods, which is a clear disadvantage since it implies a higher computational cost.

**Table 5.5.** Classification results obtained when using the MLP and SVM models and different feature selection methods on the dataset of features from images quantized with NGL = 128 gray levels.

| Classifier | MLP | SVM |
|---|---|---|
| **MWW – *p*-value** | | |
| Number of Features | 82 | 41 |
| AUC | 0.912 ± 0.060 | 0.901 ± 0.066 |
| **MIC** | | |
| Number of Features | 82 | 77 |
| AUC | 0.917 ± 0.056 | 0.897 ± 0.063 |
| *p*-value [a] | 0.631 | 0.817 |
| **ReliefF** | | |
| Number of Features | 84 | 85 |
| AUC | 0.912 ± 0.061 | 0.889 ± 0.065 |
| *p*-value [a] | 0.987 | 0.368 |

* Values of AUC are shown as mean ± SD as a result over groups' estimates.
[a] The *p*-value rows correspond to a paired *t*-test that evaluates whether the AUC values over groups' estimates obtained for each model and the MIC and ReliefF feature selection methods were significantly different ($p < 0.05$) from those obtained with the *p*-value feature selection method.

## 5.4. Discussion

Distinguishing reliably GBM from BM without resorting to invasive procedures and exhaustive, time-consuming neuroimaging evaluations is still a challenging task. To accelerate the correct diagnosis, it would be helpful to find radiological patterns in the conventional structural MRI acquired when screening the tumor lesions for the first time that allow classifying these lesions. To this end, in this work we proposed a 2D texture analysis approach combined with machine learning techniques to identify the lesion correctly. Our imaging dataset consisted on contrast-enhanced T1-weighted images of 50 patients diagnosed with GBM and 50 patients with BM. A total of 2200 features were extracted from the original images and from the images filtered with the DWT and quantized with 5 different numbers of gray levels. These features were analyzed with five different predictive models and three different feature selection methods. The

preliminary results showed that the combination of texture features extracted from the original images after quantizing them with NGL = 128 gray levels allows classifying GBM from BM with good accuracy (AUC > 0.9) when using SVM or MLP classifiers. Furthermore, we proved that quantizing the image with different numbers of gray levels resulted in different classification results. Finally, we evidenced that, for some models, a feature selection step is useful to reduce the optimal number of features to achieve higher accuracy, thus reducing the computational cost of the whole process.

Our work is not the first attempt to classify GBM and BM using features extracted from conventional structural MRI. Blanchet *et al.* [191] proposed a morphologic analysis based on T1-weighted images to classify these lesions, achieving a good level of accuracy (overall accuracy of 93.9%). Although using shape descriptors is a very interesting approach, the small sample size used in this study (18 GBM and 15 BM) cannot be considered statistically representative and the classification approach that they applied may lead to overoptimistic results since they did not apply machine learning techniques.

Regarding the use of texture features for identifying brain tumors, most of the related published studies focused on grading gliomas [117], [118], [206], [207], classifying primary brain tumors [103], [208], [209] or assessing the malignancy of the tumor [210], [211]. Only a few studies included both GBM and BM in their analyses, but in these studies other tumor types were also evaluated, so the comparison between their results and ours is not straightforward.

In the study of Georgiadis *et al*. [105], the authors evaluated 3D texture analysis on contrast-enhanced T1-weighted images from 67 patients with gliomas (27), meningiomas (19) and BM (19) to classify primary versus secondary tumors in a first stage and malignant versus benign tumors in a second stage. When evaluating the classification with an external cross-validation scheme comparable to ours to generalize the performance, they achieved an overall accuracy of 88.18% with only five features for classifying metastatic brain tumors from primary brain tumors. These results are promising but their analyses included WHO grade II and III gliomas and meningiomas, so comparing their results to ours is imprecise.

Sachdeva *et al*. [204] conducted an exhaustive study where they analyzed the potential of texture analysis to classify different types of brain tumors with SVM and artificial neural networks (ANN). They achieved an overall multiclass classification accuracy of 94% when classifying astrocytomas, meningiomas, medulloblastomas, GBM and BM with an ANN classifier combined with a genetic algorithm for selecting

the optimal number of features. However, they did not apply a resampling approach like LGOCV or K-fold CV and they evaluated textures extracted from 428 T1-weighted MRI scans from only 55 patients, so the results may be overoptimistic and overfitted. In our work we only took one lesion per patient to avoid overfitting and we applied a 5-fold approach repeated 10 times to generalize the process, avoid selection bias and obtain trustworthy results.

Finally, Zacharaki *et al.* [104] examined the classification of brain tumor type and grade by means of 3D texture and shape analysis and machine learning techniques. In this study, they evaluated 102 lesions from 98 patients, including 24 BM and 34 GBM. They explored several feature selection methods and classifiers within a leave-one-out CV structure. Additionally, they evaluated features extracted from several MRI modalities like T1 and T2-weighted images, fluid attenuated inversion recovery (FLAIR) images or relative cerebral blood volume maps, which is a very interesting approach. When applying a binary classification for differentiating GBM from BM, they achieved an accuracy of 81% and an AUC of 0.819 by using an SVM classifier and only 11 features selected with the SVM recursive feature elimination algorithm. We consider that their methodology was valid to obtain reliable and generalized results, and their work was very influential to the present study. However, our work presented some advantages for classifying GBM from BM like the inclusion of more patients per group and the improvement in the results obtained (AUC > 0.9).

One major concern about the present work consists on the use of 2D texture analysis instead of 3D texture analysis. As explained before, 3D texture analysis requires an image interpolation step to obtain isotropic image resolution, but when the slice thickness of the images is very large in comparison to the in-plane resolution, as in our case, 3D texture analysis should be avoided. Some studies have demonstrated that texture features extracted from volumetric regions capture more information about the lesion heterogeneity than features extracted from a single slice [51], and especially for brain tumors [102], [103], [105], so 3D texture analysis should be tested for the purpose of our work to verify if an improvement of the classification accuracy occurs. However, it is important to mention that 3D texture analysis presents some disadvantages in comparison to 2D texture analysis. First of all, the 3D segmentation of the lesion may be more complex and time-consuming than the segmentation of a single slice, so for clinicians it would be easier and faster to define 2D ROIs. Additionally, in some cases, only single slices are available, thus making 3D texture analysis unfeasible. Finally, the clinical evaluation still remains mostly based on 2D slices, so 2D texture analysis is easier to combine with this procedure [64].

Our work showed other limitations. Although the final results presented in this work are promising, we consider that further research should be carried out to increase the accuracy of a future validated final model, since the task of differentiating GBM from BM is very crucial and requires a precise diagnostic. Other features like clinical data, shape descriptors or other texture descriptors should be included in the model building to test their effectiveness and enhance the radiomics analysis. Other structural conventional MRI modalities like T2-weighted images or FLAIR images could be also examined by means of texture analysis to check if they show characteristic patterns of each lesion that are not present in T1-weighted images. Finally, more patients should be included in the analysis to confirm the results with a more representative sample size and to enable the creation of a final effective predictive model to classify GBMs and BMs.

## 5.5. Conclusion

The preliminary results presented in this study show that GBMs and BMs can be classified with a good level of accuracy by employing a set of 2D texture features extracted from structural MRI combined with a machine learning scheme. Currently, a definitive diagnostic can only be made by histopathologic analysis, but these promising results indicate that in the near future radiologists could provide a correct diagnosis with the combination of clinical, radiological and textural information derived from the first MRI evaluation where the lesion is detected. With this approach, patients could avoid invasive and exhaustive additional procedures and will be correctly diagnosed in the earliest stages of the disease.

# Chapter 6.
# Classification of brain metastases by their primary site of origin

## 6.1. Introduction and Motivation

As previously stated in Chapter 5, brain metastases (BMs) are the most common neoplasms of the central nervous system (CNS) in adults and constitute a significant cause of morbidity and mortality. The prognosis of patients diagnosed with metastatic brain tumors is poor: the median survival of these patients is estimated to be limited to months even for patients under treatment [22], [24], [184]. The incidence of brain metastases is certainly unknown, although some studies indicate that they occur in nearly 9–17% of patients with cancer. Noticeably, the incidence of brain metastases has increased in the past years mainly due to the improved availability of imaging techniques and to the prolonged survival from primary cancers, among other reasons [23], [185].

The primary tumors that metastasize more frequently to the brain are those originated in lung (≥ 50%), breast (15–25%) and skin (melanoma) (5–20%) [22]. However, recent studies indicated that there is a percentage of patients (2–14%) presenting brain metastases as the first manifestation of an unknown primary tumor [23]. Although the exact management of these patients remains unclear since scarce and not recent literature concerning this group of patients is available [186], [212], one of the main goals to achieve when diagnosing a BM from an unknown primary tumor consists in finding the primary site of origin. Generally, patients with BM from an unknown primary site have similar symptoms irrespective of the origin and they are subjected to

additional exhaustive imaging evaluations in order to detect the primary tumor. However, this approach is not always conclusive and invasive neuropathological procedures have to be carried out. In some review articles, like the ones published by Pekmezci and Perry [213] and Bekaert *et al.* [214], the neuropathology of BM is revised thoroughly, focusing on the evaluation of metastatic neoplasms of unknown primary tumor. In these reviews, the authors sustain that the distinction between carcinoma, sarcoma, lymphoma, and melanoma BM may be possible based on cell morphology alone, but when morphology is not enough, further investigation is needed. For example, immunohistochemical profiles are a widely accepted initial step in cases of metastases of unknown primary site. Despite everything, sometimes the origin of the brain metastases remains undiagnosed at the time of death [214]–[218]. Therefore, there is a clear need to detect the primary tumor in a fast, reliable and non-invasive way to determine the appropriate treatment, as even neuropathological strategies can offer contradictory results [214].

Some studies have attempted to determine the primary cancer by means of metrics extracted from advanced MRI techniques, like MR diffusion and perfusion weighted imaging or MR spectroscopy [219]–[221]. However, the results reported by these studies must be further corroborated and a combination of advanced MRI metrics should be tested in order to generate a robust model to identify the primary lesion [196], [222]. Furthermore, as stated in Chapter 5, the MR techniques proposed in these studies are not commonly performed when screening the tumor for the first time and may be quite demanding and time-consuming.

A practical approach for detecting the origin of the BM rapidly would be to find reliable imaging features or metrics extracted from the conventional structural MRI acquired when screening the metastatic lesion for the first time. Several studies have evaluated the differences between BM from some primary cancer types according to their location in the brain [223] or their appearance [224], [225]. Despite the promising conclusions obtained in these studies, their results require further validation. Another approach could be based on texture analysis applied to conventional MRI. Texture analysis has been proved to be an promising source of imaging biomarkers for identifying and classifying brain tumor lesions, including BMs [102]–[107]. Particularly, contrast-enhanced T1-weigted MRI was the main sequence evaluated in these studies as it is employed for initial brain tumor detection and contains abundant diagnostic information [107], [226].

The purpose of this work was to identify the primary site of origin of metastatic brain tumors using texture features extracted from T1-weighted MRI in combination with a machine learning classifier, based on the radiomics practice. Additionally, in this work we compared the performance of different machine learning approaches and the discriminative power of 2D and 3D texture features, and analyzed the influence of the number of gray levels used to quantize the images in the classification of these BMs.

## 6.2. Material and Methods

### *6.2.1. Patients and Imaging Protocol*

This retrospective, single-center study was approved by the Institutional Review Board of the Fundación Instituto Valenciano de Oncología and all subjects provided written informed consent. Patients showing single or multiple BMs were consecutively reviewed by an expert neuroradiologist (20 years-experience). Inclusion criteria comprised: (1) pathologically confirmed lung cancer, breast cancer or melanoma and only one single primary tumor; (2) no previous treatment, biopsy or surgical resection on BM; (3) all BMs confirmed by imaging and clinical follow-up and (4) no clear qualitative and/or systematic differences on T1-weighted images of the BM to identify the primary cancer (i.e., hyperintense in every melanoma case). Exclusion criteria were as follows: (1) small lesions (longest diameter < 9 mm) as texture analysis cannot capture texture information properly in small regions [13]; (2) more than 3 BMs per patient; (3) multiple BMs situated in the same brain area.

The first thirty-eight patients (22 men and 16 women, mean age 60.05 years, age range 24–74 years) who complied with inclusion criteria and not with exclusion criteria and selected between December 2013 and April 2016 were included. Sixty-seven baseline BMs were found in these patients: 27 derived from lung cancer, 23 from melanoma and 17 from breast cancer. The characteristics of the study group divided according to the type of lesion are displayed in Table 6.1 and an example of each type of BM is shown in Figure 6.1.

**Table 6.1.** Baseline characteristics of the study group per class

| | **Lung BM** | **Melanoma BM** | **Breast BM** | **Total** |
|---|---|---|---|---|
| Number of samples per group | 27 | 23 | 17 | 67 |
| Number of patients per group | 20 | 10 | 8 | 38 |
| Age* (years) | 57.70 ± 8.73 | 67.52 ± 12.21 | 57.71 ± 7.91 | 61.07 ± 10.84 |
| Sex (Male/Female) | 22 M / 5 F | 19 M / 4 F | 0 M / 17 F | 41 M / 26 F |

* Continuous variables are expressed as mean ± SD

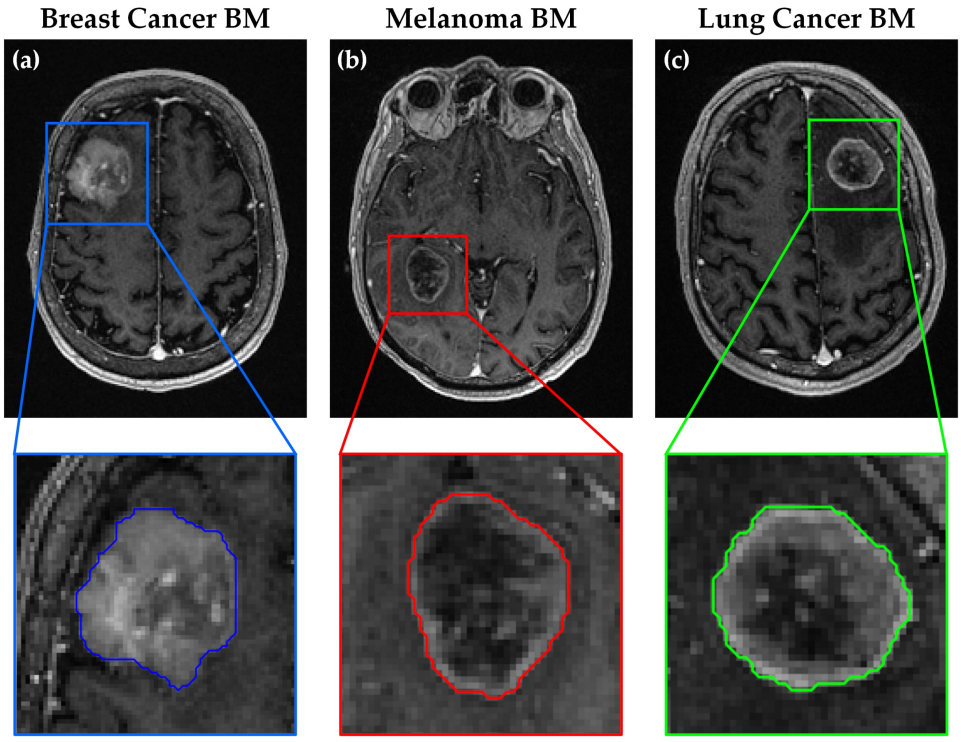



**Figure 6.1.** Examples of contrast-enhanced T1-weighted axial MRI scans of three different patients with BMs derived from a) breast cancer, b) melanoma and c) lung cancer.

Imaging was performed using a 1.5T MRI scanner (Optima MR450w; GE Medical Systems, Milwaukee, WI, USA). The MRI protocol included three-dimensional inversion recovery spoiled gradient-echo (IR-SPGR) T1-weighted brain images, according to standardized protocol [227]. Images were acquired without magnetization transfer, after intravenous administration of a single-dose of gadobenate dimeglumine (0.1 mmol/kg, MultiHance, Bracco; Milan, Italy) with a 6 minutes delay. All the BMs were scanned using the same imaging parameters since changing these parameters may lead to differences in texture analysis performance [95], [111]: repetition time/echo time (TR/TE) of 8.5/2.2 ms; flip angle of 12º; matrix size of 256×256; pixel size of 0.98×0.98 $mm^2$; and slice thickness of 1.3 mm. Partial bias field correction in raw data was conducted via the on-scanner "pre-scan normalize" option. No on-scanner gradient distortion correction was applied. As no diffusion weighted-sequences were used in this work, post processing bias field correction was not applied.

### *6.2.2. Regions of Interest*

To segment each BM in 2D, the axial slice of the 3D T1-weighted image showing the most solid lesion component was manually delineated by an expert neuroradiologist (20 years-experience). To segment each BM in 3D, all the axial slices of the 3D T1-weighted image showing tissue of the same lesion were segmented using a semiautomatic method based on the Chan-Vese algorithm [228] that takes the manually segmented 2D lesion as the initial contour. Each 3D segmented lesion was revised by the expert. The longest diameters of the volumetric lesions were normally distributed without statistical differences (One-way ANOVA *F*-test, $p > 0.05$, $p = 0.314$) between the three classes, with mean ± SD of 24.22 ± 10.67 mm (lung cancer BM), 19.92 ± 7.93 mm (melanoma BM) and 22.08 ± 10.92 mm (breast cancer BM).

Prior to the computation of texture features, some preprocessing techniques were carried out to improve texture discrimination. Firstly, normalization was applied to the MRI regions using the $\mu \pm 3\sigma$ method ($\mu$ is the mean value of the gray levels and $\sigma$ is the SD) to enhance the differences between classes, as proposed by Collewet *et al.* [72]. This method adjusts the histogram of the MRI region to $\mu \pm 3\sigma$ by rejecting the pixels with intensities out of this range.

Quantization of gray levels was also applied to the MRI regions to improve the signal-to-noise (SNR) ratio of the texture outcome [74]. This process refers to the reduction of levels of gray used to represent the image, which is originally represented by 4096 gray levels (12 bits per pixel). In particular, different number of gray levels

(NGL) were tested (8, 16, 32, 64 and 128) to study the influence of the quantization process in the discriminative power of the texture features.

Finally, volumetric regions were isotropically resampled to the in-plane resolution (voxel size = 0.98×0.98×0.98 $mm^3$) using cubic interpolation to ensure the conservation of scales and directions when extracting the 3D features [51].

Segmentation of the BM in 2D and 3D and preprocessing of the corresponding ROIs was conducted using a software tool developed specifically for this study in MATLAB (R2015b; The MathWorks Inc., Natick, MA, USA).

### *6.2.3. Feature Extraction*

Feature extraction was performed using the *Radiomics* MATLAB package [73]. Forty-three texture-based features derived from five statistical methods were computed. Three features were extracted from the intensity histogram (first-order statistics) with 100 bins and the other 40 features were extracted from the following higher-order statistical methods: gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix (GLSZM) and neighborhood gray-tone difference matrix (NGTDM). Table 6.2 compiles the features used in this study.

Only one GLCM, GLRLM, GLSZM and NGTDM per lesion was computed with the *Radiomics* package. The GLCM and NGTDM features are originally dependent on the distance of the neighboring pixels, so only adjacent connected pixels/voxels (according to [87], distance $d = 1$) were taken into account when computing these matrices (8 pixel-connectivity for 2D analysis and 26 voxel-connectivity for 3D analysis). Additionally, the GLCM and GLRLM features are originally dependent on direction, so the neighboring properties in all the possible directions (4 directions for 2D analysis and 13 directions for 3D analysis) were summed when computing these matrices to achieve rotation invariant features. To account for discretization length differences, neighbors at a distance of $\sqrt{3}$ voxels around a center voxel incremented the texture matrix by a value of $\sqrt{3}$ (only in 3D analysis), neighbors at a distance of $\sqrt{2}$ pixels/voxels around a center pixel/voxel incremented the matrix by $\sqrt{2}$, and neighbors at a distance of 1 pixel/voxel around a center pixel/voxel incremented the matrix by 1 [73].

Finally, 10 different datasets of texture features were obtained and analyzed: five datasets, one per NGL, extracted from the 2D regions and five datasets, one per NGL, from the 3D regions. All features of the 10 datasets were standardized to zero

mean and unit variance to improve numerical stability when computing the model and to avoid model building being affected by the differences in the feature scales [205].

The reproducibility of the features was evaluated since the segmentation of the ROIs was based on a manual segmentation. The lesions were segmented by two different observers: the first observer was the expert neuroradiologist (20 years-experience) that performed the segmentations for this work and the second observer was an expert in image processing and segmentation (5 years-experience) trained to segment BMs. To evaluate the reproducibility of the features, we calculated the intraclass correlation coefficient (ICC) for interrater reliability for all the features extracted from the 2D ROIs computed by both observers. In particular, we analyzed the ICC for the 43 features obtained after quantizing with the 5 different number of gray levels ($43 \times 5 = 215$ features). Only 5 features showed an ICC $< 0.75$, so 210 features achieved excellent reliability. These 5 features showed an ICC $> 0.5$, so good reliability was achieved after all.

**Table 6.2.** List of the 43 texture features used in this study.

| Method | Features | Number of features |
|---|---|---|
| Histogram | Variance, Skewness and Kurtosis | 3 |
| GLCM | Energy, Contrast, Correlation, Homogeneity, Variance, Entropy, Sum Average, Dissimilarity and Autocorrelation | 9 |
| GLRLM | Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-level Non-uniformity (GLN), Run-Length Non-uniformity (RLN), Run Percentage (RP), Low Gray-level Run Emphasis (LGRE), High Gray-level Run Emphasis (HGRE), Short Run Low Gray-level Emphasis (SRLGE), Short Run High Gray-level Emphasis (SRHGE), Long Run Low Gray-level Emphasis (LRLGE), Long Run High Gray-level Emphasis (LRHGE), Gray-level Variance (GLV) and Run-Length Variance (RLV) | 13 |
| GLSZM | Small Zone Emphasis (SZE), Large Zone Emphasis (LZE), Gray-level Non-uniformity (GLN), Zone-Size Non-uniformity (ZSN), Zone Percentage (ZP), Low Gray-level Zone Emphasis (LGZE), High Gray-level Zone Emphasis (HGZE), Small Zone Low Gray-level Emphasis (SZLGE), Small Zone High Gray-level Emphasis (SZHGE), Large Zone Low Gray-level Emphasis (LZLGE), Large Zone High Gray-level Emphasis (LZHGE), Gray-level Variance (GLV) and Zone-Size Variance (ZSV) | 13 |
| NGTDM | Coarseness, Contrast, Busyness, Complexity and Strength | 5 |

### 6.2.4. Strategies for Classification

As mentioned before, three classes of BMs were considered according to the primary site of origin (lung cancer, breast cancer and melanoma), so for our study a multiclass classification approach was needed. Random forest (RF) is a well-known ensemble learning method of the decision trees family that usually provides excellent classification results, especially when dealing with multiclass problems [144], [146]. K-nearest neighbors (KNN) is also a classifier that works easily with multiclass problems and it is usually tested due to its simplicity and good performance [104], [145]. Support vector machine (SVM) is one of the most popular classification techniques for computer-aided detection and diagnosis (CAD) in medical imaging [141], with satisfactory results reported in brain lesions classification studies [104], [106]. However, SVM is not very efficient for multiclass problems from a computational perspective since it is only directly applicable for two-class tasks, so for multiclass problems a set of binary classifiers have to be applied.

In the first stage of this study, the 10 datasets were analyzed separately using a purely multiclass approach based on RF and KNN, with the following hyperparameters:

- RF: the number of trees (*ntree*) was set to *ntree* = 250 and the number of random variables used as candidates at each split (*mtry*) was chosen from *mtry* ∈ {2, 3, 4, ..., 14, 15} in the parameter tuning process.
- KNN: the Euclidean distance was chosen the number of neighbors ($k$) was selected from $k \in \{1, 3, 5, ..., 13, 15\}$ in the parameter tuning process.

The resulting statistical metrics derived from the model performance of each dataset were compared to identify the dataset of features that provided the best classification results. Afterwards, the optimal datasets were evaluated using the one-versus-one strategy to examine the capability of these features to differentiate between individual types of BM. In this latter step, RF and KNN approaches were again tested with the same hyperparameters and an additional SVM approach was implemented with the following hyperparameters:

- SVM: a linear kernel was evaluated and the cost parameter ($C$) was chosen from $C \in \{2^{-4}, ... ,2^{0}, ... ,2^{4}\}$ in the parameter tuning process.

### 6.2.5. Model Performance and Evaluation

Considering the small sample size of our datasets, we decided to evaluate the performance of each classifier within a nested cross-validation (CV) structure (Figure 6.2). Good estimates of the model performance can be achieved using the validation data when the number of samples is not large [173]. The outer resampling loop of the nested CV structure was used to optimize the number of features and to cross-validate the models and the inner resampling loop was used to tune the model parameters.

Leave-group-out CV (LGOCV) was applied in the outer resampling loop. This resampling method randomly divides each dataset into training and test sets $N$ times, forming $N$ groups. Each group is examined independently: the samples of the training set of a group are used to build the model and then this model is evaluated using the samples of the test set of the same group. Then, the classification results provided by the estimates of all groups are averaged. A total of $N = 100$ groups were used to reduce the variance of the CV results [173]. In each group, 25% of the samples were randomly selected as test set and the remaining 75% were used as training set.

Brain metastases from the same patient were treated indistinctively in the resampling step to avoid selection bias. To support this decision, a Pearson correlation test was conducted to measure the linear dependence between random pairs of vectors of texture features from BM of the same patient ($|r| = 0.431 \pm 0.296$) and BM from different patients ($|r| = 0.424 \pm 0.248$). No statistical difference was found between the two groups (Welch's $t$-test: $p = 0.917$), suggesting that BM from the same patient are correlated in the same way that BM of different patients can be.

For the feature selection step, a filter method based on the $p$-value was employed to obtain a ranking of features with the most discriminative power. The $p$-values were obtained with the one-way analysis of variance (ANOVA) $F$-test for the multiclass strategy and the Mann-Whitney-Wilcoxon (MWW) test for the one-versus-one strategy. The ReliefF algorithm (filter method) and the mean decrease accuracy (MDA) computed in the training process of the RF model (embedded method) were also tested as feature selection methods to compare the results obtained with these approaches and our proposed filter method. To avoid overfitting, feature selection was implemented within the model-building process, that is, a different ranking of features was obtained in each group using only the training samples of each group [176]. The ranked features were progressively added one by one from most to least important and then each feature subset was used to tune the model parameters (inner 10-fold CV loop), to train the model and

to compute the metrics on the test samples of the same group. At the end, a total of $F$ = 43 sets of metrics were obtained in each group evaluation, one per each feature subset.

Although several metrics were obtained, the relevance of the classification results was estimated using the area under receiver operating characteristic curve (AUC) averaged over groups' estimates (mean ± SD). In the multiclass strategy, AUC was computed by averaging the one-versus-all statistics, as it is a simple way to extend the AUC computation to multiple classes problems [229]. Assuming that $N$ represents the number of groups chosen to perform LGOCV, the average AUC for each subset of features was calculated with the following equation:

$$AUC_{average} = \frac{\sum_{n=1}^{N} AUC_n}{N} \qquad \textbf{Equation 6.1}$$

The AUC results obtained in the training process for each of the 43 possible subsets of features were summarized in a graphic called "profile curve", which represents the evolution of the average AUC as the size of the subset of features increases.

The model evaluation process was implemented with the Caret package [143] in R language, version 3.2.5 (R Development Core Team, Vienna, Austria).

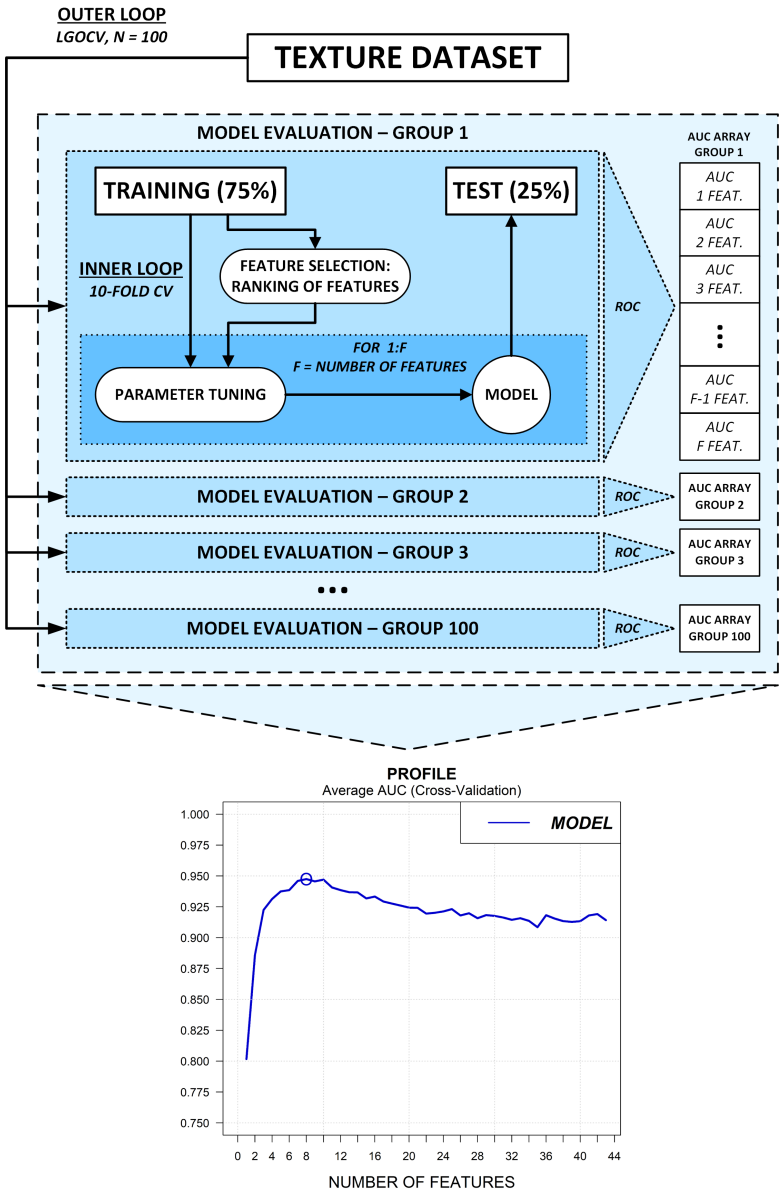



**Figure 6.2.** Structure of the nested CV method used to evaluate the different datasets of features. All the samples of each dataset were randomly separated in training and test sets $N = 100$ times to evaluate each model with the AUC, examining different subsets of features.

## 6.3. Results

### *6.3.1. Multiclass Strategy*

In general, 3D features provided better classification accuracy than 2D features in the multiclass strategy for both RF and KNN approaches and for the three feature selection methods tested (*p*-value obtained with ANOVA *F*-test, ReliefF and MDA). However, the number of gray levels used for quantization affected the model performance considerably. As it is shown in Figure 6.3, 3D features from the MRI lesions quantized with NGL of 8, 16 and 32 gray levels provided better average AUC than the equivalent 2D features. However, for NGL = 64 and NGL = 128, the resulting AUC was similar for 3D and 2D features and in some cases 2D features were even more discriminative than 3D features. Therefore, 3D features were more influenced by the quantization of the MRI regions than 2D features, losing discriminative power when increasing NGL. Additionally, for achieving the highest AUC, in general 3D analysis required fewer features than 2D analysis, which is an advantage for reducing the time complexity of the process.

Regarding the comparison between classifiers and between feature selection methods, the differences were not very noticeable. Focusing only on the 3D texture analyses, RF models combined with ANOVA *F*-test and ReliefF feature selection methods provided higher AUC values than KNN models with the same feature selection methods. The results obtained with MDA feature selection method were similar for both RF and KNN classifiers. For the following analyses, we decided to apply only the *p*-value feature selection method since it is simpler and more intuitive, and ReliefF and MDA algorithms did not improve the results.

The highest AUC was achieved when using 3D features extracted from the lesions quantized with NGL = 32 and a RF model trained with only the top four features ranked with the *p*-value feature selection method (AUC = 0.873 ± 0.064). Hence, we chose the dataset of 3D features and NGL = 32 as the optimal dataset for the following analyses. However, it is important to mention that no statistically significant difference was found when using 3D features from images quantized with 8, 16 and 32 gray levels (paired *t*-test with the arrays of AUCs obtained over groups' estimates, $p > 0.05$), so the three datasets provided comparable good results. Additionally, the paired *t*-test results indicated that the optimal dataset (3D features and NGL = 32) provided an average AUC that was significantly different from the AUCs obtained with 2D features ($p < 0.05$).

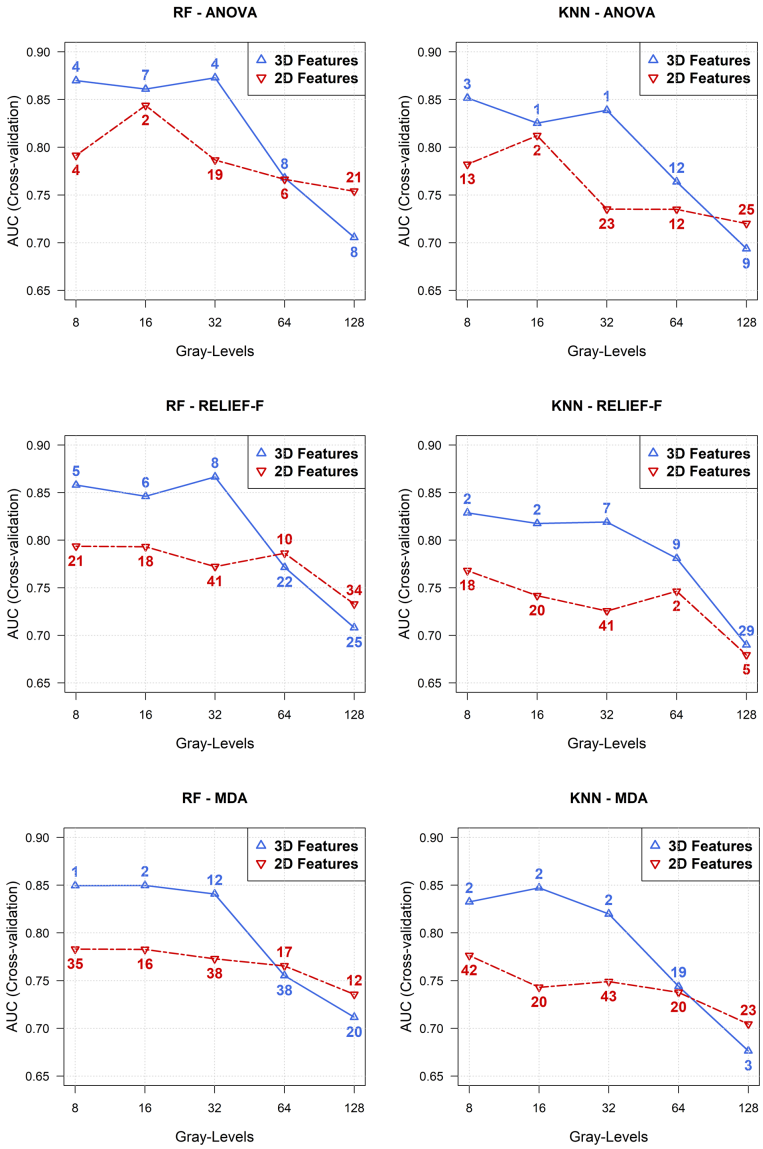



**Figure 6.3.** Comparison between the multiclass results of RF and KNN models using 2D and 3D features and the three feature selection approaches ($p$-value from ANOVA $F$-test, ReliefF and MDA) for all the number of gray levels considered in this study. The numbers on the curves indicate the number of features used to achieve the maximum AUC.

Table 6.3 shows the top ten ranked features of the optimal dataset, with the average *p*-value of each feature computed using the ANOVA *F*-test. It is important to mention that the nested CV scheme used in this study did not allow determining the exact ranking of features because the feature selection step was recomputed at each loop, so an average ranking was obtained. Table 6.3 also shows that features derived from the GLCM, GLRLM and GLSZM topped the ranking with significant *p*-value ($p < 10^{-3}$). However, when the *p*-value is obtained with the ANOVA *F*-test, the value indicates that there is a significant difference between at least two of the three classes of BMs, so additional evaluation of the difference between individual groups was needed.

**Table 6.3.** Top ten features of the dataset with the highest AUC (3D features, NGL = 32 gray levels) ranked according to their average *p*-value computed with the ANOVA *F*-test in the multiclass analysis.

| Method | Feature | Average Ranking | Average *p*-value |
|---|---|---|---|
| **GLCM** | **Variance** | **1.02** | **$< 10^{-8}$** |
| **GLSZM** | **Low Gray-level Zone Emphasis** | **2.72** | **$< 10^{-6}$** |
| **GLCM** | **Sum Average** | **3.02** | **$< 10^{-6}$** |
| **GLSZM** | **Small Zone Low Gray-level Emphasis** | **3.73** | **$< 10^{-6}$** |
| GLRLM | Short Run Low Gray-level Emphasis | 5.36 | $< 10^{-5}$ |
| GLRLM | Low Gray-level Run Emphasis | 6.72 | $< 10^{-5}$ |
| GLRLM | High Gray-level Run Emphasis | 6.86 | 0.00001 |
| GLSZM | High Gray-level Zone Emphasis | 7.37 | 0.00001 |
| GLCM | Autocorrelation | 8.52 | 0.00004 |
| GLSZM | Gray-level Non-uniformity | 10.32 | 0,00062 |

* The subset of features highlighted in bold provided the highest classification accuracy.

### 6.3.2. One-versus-one Strategy

An overall confusion matrix (CM) was obtained from the analysis presenting the highest results in the multiclass strategy (3D features, NGL = 32, RF model and 4 features from *p*-value ranking) by summing up all CMs obtained in every group's estimate (Table 6.4). The overall CM revealed that lung cancer BMs were classified correctly most of the time (82%), but breast cancer and melanoma BMs were often misclassified. This statement is in accordance with the results provided by the one-versus-one analysis.

The one-versus-one analysis revealed that it would be possible to differentiate precisely lung cancer BMs from breast cancer BMs (AUC > 0.96) and melanoma BMs (AUC > 0.92) using few features of the optimal dataset (less than 4 and 12 features respectively), no matter which classifier is used. However, poor accuracy (AUC < 0.62) was achieved when discriminating BMs from breast cancer and melanoma for the three classifiers, thus indicating that these features are not suitable for classifying those types of BMs. These results are displayed in Figure 6.4. Additional statistical metrics were computed to validate the results (Table 6.5).

**Table 6.4.** Overall confusion matrix extracted from the RF model performance using the dataset with the best results in the multiclass strategy (3D features, NGL = 32 gray levels).

| | | **Predicted Class** | | |
|---|---|---|---|---|
| | | Breast Cancer | Lung Cancer | Melanoma |
| **True Class** | Breast Cancer | 235 (58.75%) | 44 (11%) | 121 (30.25%) |
| | Lung Cancer | 55 (9.17%) | 492 (82%) | 53 (8.83%) |
| | Melanoma | 95 (19%) | 66 (13.20%) | 339 (67.80%) |

* The percentages indicate the proportion of samples of one specific class that were classified in each of the three classes, throughout all the iterations of the nested cross-validation process.

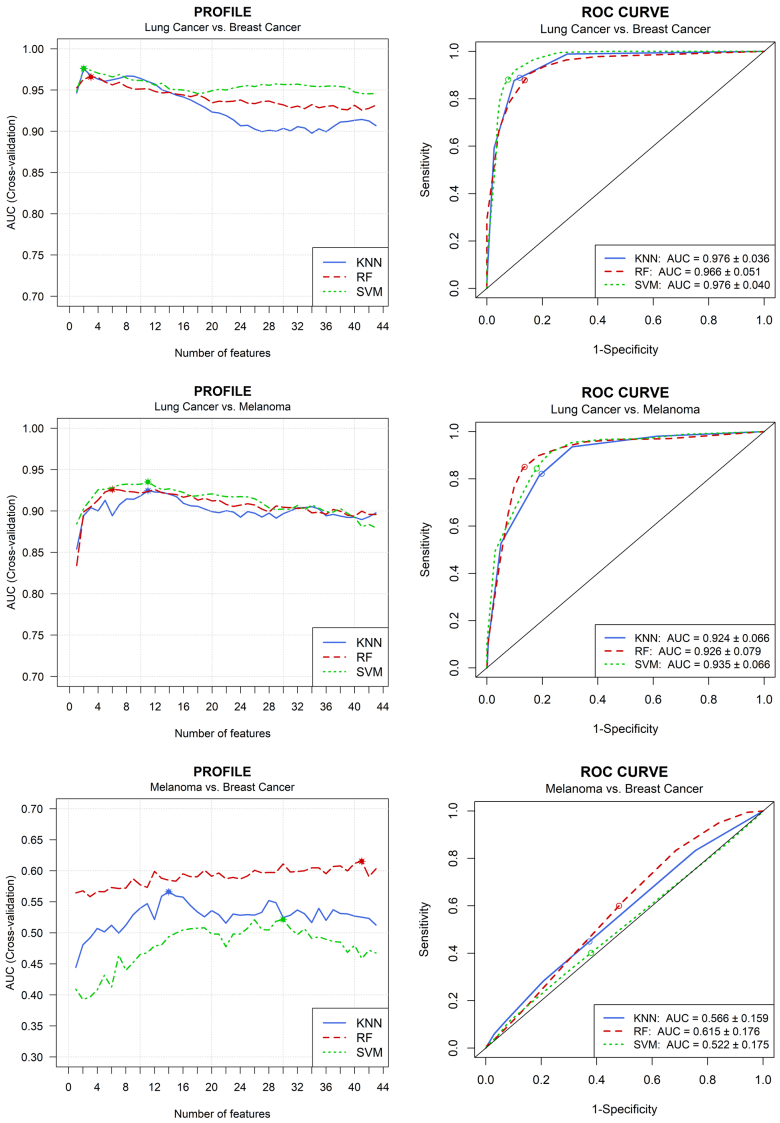



**Figure 6.4.** Profile curves and the corresponding average ROC curves for the best subset of features obtained in the one-versus-one analysis with the three classifiers (KNN, RF and SVM). The highlighted points on the profile curves indicate the optimal subset of features. The highlighted points on the ROC curves indicate the optimal cutoff points that weighs both sensitivity and specificity equally computed with the "closest-to-(0,1)" criterion.

**Table 6.5.** Additional metrics obtained in the one-versus-one analysis using the KNN, RF and SVM models on the best dataset (3D features, NGL = 32 gray levels).

| Classifier | KNN | RF | SVM |
|---|---|---|---|
| **Lung Cancer vs. Breast Cancer** | | | |
| Number of Features | 2 | 3 | 2 |
| AUC | 0.976 ± 0.036 | 0.966 ± 0.051 | 0.976 ± 0.040 |
| Sensitivity [a] | 0.908 ± 0.145 | 0.905 ± 0.165 | 0.895 ± 0.159 |
| Specificity [a] | 0.921 ± 0.099 | 0.880 ± 0.140 | 0.925 ± 0.116 |
| Overall Accuracy | 0.905 ± 0.078 | 0.870 ± 0.089 | 0.894 ± 0.081 |
| Kappa Index | 0.795 ± 0.174 | 0.728 ± 0.187 | 0.769 ± 0.182 |
| **Lung Cancer vs. Melanoma** | | | |
| Number of Features | 11 | 6 | 11 |
| AUC | 0.924 ± 0.066 | 0.926 ± 0.079 | 0.935 ± 0.066 |
| Sensitivity [a] | 0.818 ± 0.159 | 0.853 ± 0.164 | 0.845 ± 0.172 |
| Specificity [a] | 0.824 ± 0.161 | 0.876 ± 0.147 | 0.832 ± 0.165 |
| Overall Accuracy | 0.833 ± 0.093 | 0.870 ± 0.096 | 0.845 ± 0.090 |
| Kappa Index | 0.662 ± 0.188 | 0.739 ± 0.190 | 0.687 ± 0.181 |
| **Melanoma vs. Breast Cancer** | | | |
| Number of Features | 14 | 41 | 30 |
| AUC | 0.566 ± 0.159 | 0.615 ± 0.176 | 0.522 ± 0.175 |
| Sensitivity [a] | 0.363 ± 0.276 | 0.598 ± 0.248 | 0.423 ± 0.286 |
| Specificity [a] | 0.730 ± 0.181 | 0.516 ± 0.229 | 0.624 ± 0.221 |
| Overall Accuracy | 0.562 ± 0.133 | 0.550 ± 0.153 | 0.554 ± 0.098 |
| Kappa Index | 0.074 ± 0.287 | 0.082 ± 0.310 | 0.034 ± 0.208 |

***** Values are shown as mean ± SD as a result over groups' estimates.

[a] Sensitivity and specificity were computed according to the optimal cutoff point of the ROC curve computed with the "closest-to-(0,1)" criterion.

The three classifiers worked similarly and the results were comparable. However, it is important to mention that SVM classifier provided higher AUC values than RF when classifying lung cancer BMs from breast cancer BMs (1% AUC increase) and melanoma BM (0.9% AUC increase). However, when distinguishing breast cancer BMs from melanoma BMs, the results obtained with the SVM model were considerably worse (9.3% AUC decrease). The KNN results were better than the RF results for the classification of lung cancer BMs and breast cancer BMs (1% AUC increase), but worse when classifying melanoma BMs from lung cancer BMs (0.2% AUC decrease) and breast cancer BMs (4.9% AUC decrease). Based on these results, RF model would be the optimal choice because it provided the most balanced results. However, as mentioned before, poor results were achieved for discriminating BMs from breast cancer and melanoma, so none of the three classifiers would be still valid to classify those types of BMs reliably.

Regarding the top ranked features, Table 6.6 shows that the ranking of features provided by the multiclass strategy mostly coincided with the rankings computed to classify lung cancer BM from breast cancer and melanoma BM. Furthermore, the top ten features of both rankings showed significant average $p$-values ($10^{-7} < p < 10^{-2}$). However, none of the features showed significant average $p$-value ($p > 0.2$) when classifying BMs from breast cancer and melanoma. Finally, it is relevant to mention that features derived from GLCM, GLRLM and GLSZM were useful to classify these types of BMs but features derived from the histogram and the NGTDM had no influence on the final results.

**Table 6.6.** Top ten features of the best dataset (3D features, NGL = 32 gray levels) ranked according to their average *p*-value computed with the MWW test in the one-versus-one analysis.

| Lung Cancer vs. Breast Cancer | | Lung Cancer vs. Melanoma | | Breast Cancer vs. Melanoma | |
|---|---|---|---|---|---|
| Feature | Average *p*-value | Feature | Average *p*-value | Feature | Average *p*-value |
| **Sum Average** | $< 10^{-6}$ | **Small Zone Low Gray-level Emphasis** | $< 10^{-5}$ | **Gray-level Non-uniformity** [b] | 0.19556 |
| **Variance** [a] | $< 10^{-6}$ | **Variance** [a] | $< 10^{-5}$ | **Variance** [a] | 0,21753 |
| **Low Gray-level Zone Emphasis** | 0.00001 | **Low Gray-level Zone Emphasis** | $< 10^{-5}$ | Small Zone Emphasis | 0,34021 |
| **Small Zone Low Gray-level Emphasis** | 0.00001 | **Short Run Low Gray-level Emphasis** | $< 10^{-5}$ | Zone-Size Non-uniformity | 0,34021 |
| **High Gray-level Zone Emphasis** | 0.00005 | **Low Gray-level Run Emphasis** | 0.00001 | **Small Zone Low Gray-level Emphasis** | 0,32071 |
| **Autocorrelation** | 0.00007 | **Sum Average** | 0.00003 | **Sum Average** | 0,36852 |
| **High Gray-level Run Emphasis** | 0,00014 | **High Gray-level Run Emphasis** | 0.00028 | **Autocorrelation** | 0,37284 |
| **Short Run Low Gray-level Emphasis** | 0,00035 | **High Gray-level Zone Emphasis** | 0,00035 | **Low Gray-level Zone Emphasis** | 0,36283 |
| **Low Gray-level Run Emphasis** | 0,00119 | Long Run Low Gray-level Emphasis | 0,00054 | Entropy | 0,39376 |
| Long Run Low Gray-level Emphasis | 0,00294 | **Gray-level Non-uniformity** [b] | 0,00081 | Dissimilarity | 0,47033 |

* The features in bold are in accordance with those features ranked in the multiclass analysis.

[a] These features are computed from the GLCM (Gray-level co-occurrence matrix)

[b] These features are computed from the GLSZM (Gray-level size zone matrix)

## 6.4. Discussion

Determining the origin of a BM in those patients where the primary cancer is unknown without invasive, exhaustive and time-consuming procedures is still a challenging task. To this end, in this work, we proposed a novel radiomics methodology based on 2D and 3D texture analysis on structural MRI and machine learning approaches to test the feasibility of texture descriptors to identify correctly the primary site of origin of the BMs. The radiomics approach used in this study showed that 3D texture features were more suitable than 2D features for classifying lung cancer BMs from breast cancer and melanoma BMs, achieving an average AUC > 0.9 in both cases. Several classification models were tested (KNN, RF and SVM) and all of them provided similar results, but RF classifier provided more balanced results for classifying the three types of BMs under study. Furthermore, the results improved when limiting the number of features with a feature selection scheme.

Our work is not the first attempt to differentiate BMs by its primary site of origin using texture features. Béresová *et al*. [230] studied the statistical significance of 2D and 3D texture features from the histogram and the GLCM to identify the differences between lung and breast cancer BM. Our work enhances this study by exploring more texture features, including melanoma patients and considering a machine learning approach. With our results, we support the conclusions of Béresová *et al*. that texture analysis may help in the discrimination of BM from different primary tumors.

We based our work on other similar studies that showed the potential of MRI texture features combined with machine learning techniques to classify different brain lesions, including BMs. Larroza *et al.* [106] used texture features to distinguish between BM and radiation necrosis using a LGOCV structure and support vector machine classifier (AUC > 0.9). Li *et al.* [107] used texture features to differentiate BMs from different pathological types of lung cancers using K-nearest neighbors and back-propagation artificial neural network classifiers in a one-versus-one approach (AUC ≥ 0.9 when differentiating small cell lung carcinoma from other types of lung cancers). Both studies showed promising result and were very influential to our work. However, we tried to go beyond by including 3D texture features and taking into account rotation invariance for extracting the features.

Several studies have addressed the problem of classifying different brain tumor types by analyzing the potential of 3D MRI texture features in comparison with 2D features [102], [103], [105]. These studies showed an improvement in classification

accuracy when using 3D texture analysis. The conclusions in these works are clear: 3D texture descriptors capture more information about the lesion heterogeneity than 2D descriptors. In particular, the study of Fetit *et al.* [103] is very conclusive on this matter. This study mainly compares 2D and 3D texture features with several predictive models to classify different childhood brain tumors. All the models worked better with 3D features: for example, the neural network classifier showed 12% improvement in AUC and 19% in overall accuracy when using 3D texture analysis instead of 2D analysis. Nevertheless, 3D texture analysis presents some drawbacks. Firstly, the 3D segmentation of the lesion can be more complex and time-consuming than the segmentation of a single slice. Additionally, 3D texture analysis requires MRI scans as isotropic as possible to reduce the effect of the image interpolation, and the acquisition process of these scans can be very slow.

The influence of the NGL used in the quantization of MRI has been analyzed in some studies with mixed results. No difference was reported by several studies [73], [76] when comparing the effect of changing NGL on the texture outcome. However, other studies showed that the discriminative power of texture-based features were affected by the gray-level quantization. Chen *et al.* [61] found that the optimal results for characterizing breast lesions were achieved for NGL = 32. Leite *et al.* [77] observed that quantizing with NGL = 16 allowed identifying the etiology of brain white matter lesions more accurately. Mahmoud-Ghoneim *et al.* [75] analyzed the impact of varying NGL on GLCM features of brain white matter: they concluded that their classification results were influenced significantly by the NGL chosen and they obtained better results with NGL = 128 for both 2D and 3D texture analysis. Our results support the fact that the NGL should be optimized for each specific application because it can lead to better classification results.

One major concern in this work is related to the exclusion of possible significant variables like sex or age in the model building. In our study, the age of the patients was not a significant parameter; we performed an MWW test for independent samples and the results were not significant (breast cancer vs lung cancer: $p$-value = 0.779; breast cancer vs melanoma: $p$-value = 0.227; lung cancer vs melanoma: $p$-value = 0.052). In the case of sex, we conducted a chi-squared test of independence and the results were not significant when comparing lung cancer and melanoma patients ($p$-value = 1). However, we found statistical differences when comparing breast cancer patients with melanoma and lung cancer patients ($p$-value < 0.05). This occurs because our set of patients was not entirely balanced, and all our breast cancer patients were women (male breast cancer is rare, but it does exist). However, our purpose was to analyze exclusively

the potential of texture analysis to discriminate between BMs from different primary sites of origin, so the evaluation of the sex and age for classifying BMs went beyond our main purpose, as these variables do not affect the heterogeneity properties of the BMs.

Our study showed other important limitations. The main limitation was the reduced set of BMs; more samples would be needed to build and test a final predictive model. In addition, we only considered metastases derived from the most common primary sites of origin; other types of BMs like those from renal or colorectal cancer should be considered in further analyses because it is necessary to consider all possible sites of origin to build a reliable final predictive model. Moreover, we only included MR images acquired with the same scanner and imaging parameters since texture analysis can be affected by differences in scan parameters; a multicenter study on this specific application should be performed to evaluate this limitation. Finally, our study failed to classify breast cancer and melanoma BMs, so further investigation will be conducted by exploring other texture methods like Local Binary Patterns or transform methods (Wavelets, Gabor filters…) or other MRI sequences that could capture differences between BMs from different primary sites of origin. To our knowledge, a genetic or pathologic link between breast cancer and melanoma that could be related to these texture analysis results is unclear at this point, and the study of this association goes beyond the objective of this work.

## 6.5. Conclusion

Our results show that texture analysis on T1-weighted MRI in combination with a RF classifier allows differentiating accurately BMs of lung cancer origin from those of breast cancer and melanoma origin when the proper features are chosen with a feature selection scheme. Although we only included the three more common BM origins, we established a robust methodology to perform a multiclass classification that could be extended to other primary sites of origin. Our promising results reveal that, with further research, texture analysis could help in the identification of the primary site of origin in patients with BMs from an unknown primary cancer. In addition, patients with two known primary tumors could benefit from this methodology to find which tumor has metastasize to the brain. These results support the conclusions derived from other studies to encourage radiologists to use texture analysis as a new tool to improve precision in diagnosis.

# Chapter 7.
# Evaluation of new biomarkers for Alzheimer's disease

## 7.1. Introduction and Motivation

In 2011, it was estimated that 35.6 million people around the world suffered from dementia, whereas, in 2015, this figure amounted to 46.8 billion. At this rate, it is expected that this figure will almost double by 2030 and more than treble by 2050 [25], [231]. Alzheimer's Disease (AD) represents the most common type of dementia, accounting for an estimated 60 to 80 percent of cases. This neurodegenerative disorder is characterized by the presence of a progressive deterioration of the cognitive and behavioral functions, mainly in the old age [27].

The diagnosis of AD remains nowadays fundamentally clinical, which means that it cannot be diagnosed until the first symptoms appear, or even later, because these early symptoms are usually associated with consequences due to aging [25]. Definitive diagnosis can only be made with histopathological confirmation of amyloid plaques and neurofibrillary tangles, usually at autopsy [30]. This is the main reason behind exploring new biomarkers that allow an early detection of AD, as patients could benefit from more efficient treatments if the AD is diagnosed in its first stages, or even before the first symptoms appear. In the past years many studies have focused on the analysis of mild cognitive impairment (MCI) to this end, as it is considered as a prodromal stage of the disease or a transitional phase between normal ageing and AD, although not all patients with MCI develop AD [29], [232], [233]. Furthermore, several studies have proposed

the existence of a preclinical stage prior to the appearance of symptoms, during which neuropathological typical changes of AD already occur [32], [234], [235]. This pre-symptomatic phase is still under research.

Imaging has played an important role in the study of AD over the past decades. Diagnostically, imaging has moved from a minor role to a central position. In particular, structural magnetic resonance imaging (MRI) has gain more attention than other imaging techniques because it allows visualizing in life the progressive cerebral atrophy that characterize the neurodegenerative process of dementia, thus contributing to improve diagnostic accuracy [30], [236], [237]. This progressive cerebral atrophy firstly affects the medial temporal lobe [238], being the entorhinal cortex the earliest site of atrophy, closely followed by the hippocampus, amygdala, and parahippocampal gyrus. Consequently, new biomarkers for early diagnosis of AD could be defined by processing and studying structural MRI of these brain structures [239], [240].

In general, the features extracted from MRI typically used as biomarkers of AD are related to volume and/or shape changes of specific brain structures, thus only taking into account macroscopic apparent alterations that occur when neurodegeneration has already taken place [241]–[243]. However, in the past years, texture analysis has been considered as a source of imaging biomarkers for characterizing AD. In particular, several studies tried to differentiate AD patients from cognitive normal (CN) and MCI patients using 2D or 3D texture features extracted from T1-weighted MRI and focusing on the hippocampus [58], [244], [245], other brain regions or structures [119], [246], [247], or even the whole brain [248]–[251]. Most of these studies proved that AD patients could be differentiated from CN and MCI patients using texture features with good accuracy (by means of statistical analyses or machine learning techniques), but the predictive capacity of the texture features was reduced when comparing CN and MCI patients. These shared results reinforce the necessity of exploring new texture biomarkers to identify image differences in the early stages of AD, when patients transition from cognitive normal to prodromal stage [252].

The main objective of the present study consisted on the analysis of new potential biomarkers for AD through the acquisition of image parameters using 2D and 3D texture analysis on MRI. For this purpose, 3D T1-weighted MRI from three different populations were used: AD patients, MCI patients and CN subjects. The analysis was performed mainly in the hippocampal region using circular and spherical regions of interest (ROIs). The texture features were analyzed by means of statistical analysis and machine learning and several approaches were tested to obtain the optimal results.

## 7.2. Material and Methods

### *7.2.1. Patients and Imaging Protocol*

The images used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu) [253]. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org.

In this study we evaluated T1-weigthed MRI brain images of a total of 150 subjects from the ADNI2 database, comprising 50 AD patients, 50 CN patients and 50 MCI patients ranging from 57 to 91 years (74.71 ± 7.97 years, mean ± SD) who underwent MRI between January 2013 and December 2014. In the ADNI2 database, the MCI patients are subdivided in late MCI (LMCI) and early MCI (EMCI) according to the performance of the age-, sex- and education-adjusted normative mean in a standardized test for evaluating cognitive impairment [254]. In LMCI, impairment is identified using the original definition of MCI (performance of 1.5 standard deviations below the normative mean on a standardized test), whereas in EMCI, impairment is defined as performance between 1.0 and 1.5 standard deviations below the normative mean [255]. For our study, all 50 MCI patients were chosen from the EMCI group in order to take into account earlier stages of AD, closer to the pre-symptomatic phase. The specific characteristics of the study group divided according to the type of subject are displayed in Table 7.1.

**Table 7.1.** Baseline characteristics of the study group

| | **CN** | **EMCI** | **AD** | **Total** |
|---|---|---|---|---|
| Number of samples per group | 50 | 50 | 50 | 150 |
| Age* (years) | 78.34 ± 6.90 | 70.35 ± 6.15 | 75.37 ± 8.62 | 74.71 ± 7.97 |
| Sex (Male/Female) | 27 M / 23 F | 33 M / 17 F | 31 M / 19 F | 91 M / 59 F |

* Continuous variables are expressed as mean ± SD

The T1-weighted MRI images were acquired using 3T Siemens MRI scanners (Siemens Medical Solutions, Erlangen, Germany) with the following protocol: 3D magnetization prepared rapid gradient echo (MP-RAGE) sequence, repetition time/echo time (TR/TE) of 2300/2.98 ms, tilt angle of 9º, sagittal acquisition plane, in-plane resolution of 1×1 $mm^2$, slice thickness of 1.2 mm, scan matrix 256×256 and field of view of 256 mm. The different image files were downloaded in NIfTI format, a format used for storing volumetric MRI data. An example of the T1-weighted MRI brain images used in our study is shown in Figure 7.1. The difference in the brain atrophy (and specifically the hippocampus) between AD, EMCI and CN patients can be seen in this figure.

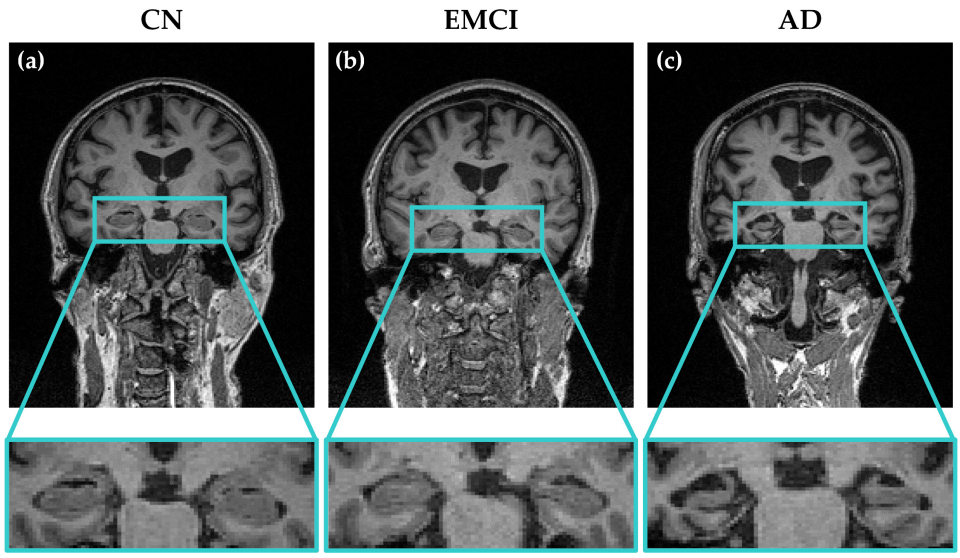



**Figure 7.1.** Examples of T1-weighted coronal MRI scans of three different subjects from the three groups of patients considered in this study: a) cognitive normal (CN), b) early mild cognitive impairment (EMCI) and c) Alzheimer's disease (AD). The progressive atrophy in the hippocampal region through the different stages of the disease can be observed.

### 7.2.2. Regions of Interest

In the present study, we decided to evaluate the right and left hippocampi of AD, EMCI and CN patients as the regions of interest (ROIs). The segmentation of the T1-weighted MRI images was carried out on the coronal slices since this plane allows imaging the hippocampus more appropriately, thus being the preferred plane for manual tracing of hippocampal borders in most of the studies [256]. Due to the complexity of finding the hippocampal borders and the difficulty of segmenting manually the hippocampus (especially in AD patients, where this structure has already been affected by atrophy) [256], circular and spherical segmentation was performed for defining the 2D and 3D ROIs respectively. For each hippocampus (right and left), three concentric circles and spheres with different radii were drawn. The center of each circle or sphere was defined manually by clicking on the center of the right and left hippocampi in the coronal slice where both hippocampi showed a larger area. The radii used were of 3, 5 and 8 pixels, for both 2D and 3D segmentation. This way, the smallest ROI (ROI1, $r$ = 3 pixels) contained only hippocampal tissue and the middle ROI (ROI2, $r$ = 5 pixels) and biggest ROI (ROI3, $r$ = 8 pixels) contained tissue from the hippocampus and from surrounding structures like the entorhinal cortex. At the end, a total of 6 ROIs were defined on each hippocampus, 3 circular ROIs (2D) and 3 spherical ROIs (3D).

Before conducting texture analysis, some preprocessing techniques were applied to the image ROIs. Firstly, the image ROIs were normalized using the $\mu \pm 3\sigma$ ($\mu$ is the mean value of the gray levels and $\sigma$ is the SD) to enhance the differences between groups, as proposed by Collewet *et al*. [72]. This method adjusts the histogram of the image ROI to $\mu \pm 3\sigma$ by rejecting the pixels with intensities out of this range. Quantization of gray levels was also applied to the image ROIs to improve the signal-to-noise (SNR) ratio of the texture outcome and to reduce the computation time of the matrix-based texture features [74]. This process refers to the reduction of levels of gray used to represent the image, which is originally represented by 4096 gray levels (12 bits per pixel). In this case, image ROIs were quantized to 32 gray levels (5 bits per pixel). However, other number of gray levels (NGL) were used to quantize the image ROIs (8, 16, 64 and 128) to test the influence of this parameter in the performance results.

The process of selecting the ROIs in the MR images was performed using a software tool developed specifically for this study in MATLAB (R2015b; The MathWorks Inc., Natick, MA, USA).

### 7.2.3. Feature Extraction

Texture analysis was conducted on each of the 12 preprocessed image ROIs (6 circular and 6 spherical) defined for every subject with the *Radiomics* MATLAB package implemented by Vallieres *et al.* [73]. A total of 46 statistical texture features were computed per ROI. According to the type of relationship between pixels/voxels quantified by each feature, three groups of parameters were established: global, local and regional [257]. The global parameters describe the whole gray-level distribution of the image ROI, and they were obtained from the intensity histogram of the ROI with 100 bins (6 features). The local parameters (14 features) describe the spatial relationship between pairs of pixels, and they were extracted from the gray-level co-occurrence matrix (GLCM) (9 features) and the neighborhood gray-tone difference matrix (NGTDM) (5 features). The regional parameters (26 features) measure the distribution of groups of connected pixels with the same gray-level values, and they were calculated from the gray-level run-length matrix (GLRLM) (13 features) and the gray-level size-zone matrix (GLSZM) (13 features). Table 7.2 shows the list of parameters evaluated in this work. All features were standardized to zero mean and unit variance.

Originally, GLCM and GLRLM features are dependent on direction, so different values may be obtained if the image is rotated. For texture characterization on MRI this fact is unacceptable since images from different patients may have different orientations. Additionally, GLCM and NGTDM features are dependent on the distance, meaning that different feature values are obtained depending on the distance of the set of neighboring pixels/voxels to the center pixel/voxel chosen to compute the matrices. To solve these problems, the *Radiomics* package only computes one GLCM, GLRLM and NGTDM per image region by considering only the neighboring connected pixels (according to [87], distance $d = 1$) and by summing the matrices computed over all directions, thus achieving rotation invariant features. Consequently, for 2D texture analysis, 8 pixel-connectivity was applied and the neighboring properties of the pixels in the 4 directions of the 2D space (0, 45, 90 and 135°) were summed. For 3D texture analysis, 26 voxel-connectivity was applied and the neighboring properties of the voxels in the 13 directions of the 3D space were summed. To account for discretization length differences, neighbors at a distance of $\sqrt{3}$ voxels around a center voxel incremented the texture matrix by a value of $\sqrt{3}$ (only for 3D texture analysis), neighbors at a distance of $\sqrt{2}$ pixels/voxels around a center pixel/voxel incremented the matrix by a value of $\sqrt{2}$, and neighbors at a distance of 1 pixel/voxel around a center pixel/voxel incremented the matrix by a value of 1 [73].

**Table 7.2.** List of the 46 texture features used in this study.

| Scale | Method | Features | Number of features |
|---|---|---|---|
| Global | Histogram | Mean, Variance, Skewness, Kurtosis, Energy and Entropy | 6 |
| Local | GLCM | Energy, Contrast, Correlation, Homogeneity, Variance, Entropy, Sum Average, Dissimilarity and Autocorrelation | 9 |
| | NGTDM | Coarseness, Contrast, Busyness, Complexity and Strength | 5 |
| Regional | GLRLM | Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-level Non-uniformity (GLN), Run-Length Non-uniformity (RLN), Run Percentage (RP), Low Gray-level Run Emphasis (LGRE), High Gray-level Run Emphasis (HGRE), Short Run Low Gray-level Emphasis (SRLGE), Short Run High Gray-level Emphasis (SRHGE), Long Run Low Gray-level Emphasis (LRLGE), Long Run High Gray-level Emphasis (LRHGE), Gray-level Variance (GLV) and Run-Length Variance (RLV) | 13 |
| | GLSZM | Small Zone Emphasis (SZE), Large Zone Emphasis (LZE), Gray-level Non-uniformity (GLN), Zone-Size Non-uniformity (ZSN), Zone Percentage (ZP), Low Gray-level Zone Emphasis (LGZE), High Gray-level Zone Emphasis (HGZE), Small Zone Low Gray-level Emphasis (SZLGE), Small Zone High Gray-level Emphasis (SZHGE), Large Zone Low Gray-level Emphasis (LZLGE), Large Zone High Gray-level Emphasis (LZHGE), Gray-level Variance (GLV) and Zone-Size Variance (ZSV) | 13 |

### *7.2.4. Statistical Analysis*

A preliminary statistical study was carried out to evaluate the discriminative power of each feature, that is to say, if it offered a good differentiation between the three populations considered in this work: AD, EMCI and CN. The final purpose of this analysis was to check the feasibility of these features as biomarkers of AD.

To compare the distributions of the texture parameters for the three classes, a statistical test was applied to each feature in order to evaluate the difference between individual groups. Specifically, the Mann-Whitney-Wilcoxon (MWW) test, also called Mann-Whitney *U* test or Wilcoxon rank sum test, was used to compare the populations in pairs. This non-parametric test, analog to the independent samples *t*-test, does not require the normality assumption of the *t*-test and it is recommended when the sample sizes are relatively small [258].

In statistics, when the number of statistical tests performed increases, the contrast test became more permissive, rejecting the void hypothesis more easily and increasing this way the number of false positives by rising the probability of obtaining randomly a significant result [258]. This problem is usually referred as the multiple comparisons problem. To counter this effect, we decided to apply two multiple comparisons correction methods before determining which features were statistically significant. The first method applied was the Bonferroni correction, which controls the family-wise error rate. This method compensates the type I error (incorrectly rejecting the null hypothesis) and attempts to limit the probability of even one false discovery, so it is relatively strong (conservative) and, in some cases, it may lead to a very high rate of false negatives, thus increasing the type II error (accepting the null hypothesis when the alternative is true). The second method used was the Benjamini-Hochberg (BH) procedure, which controls the false discovery rate. This method attempts to control the expected proportion of false discoveries, that is, the proportion of discoveries (significant results) that are actually false positives, thus being less sensitive than the Bonferroni correction.

### *7.2.5. Machine Learning Analysis*

In this work, three classes of subjects were considered (CN, EMCI and AD), so for our study a multiclass classification approach was necessary. Since we wanted to evaluate the difference between individual groups, we decided to apply directly a one-versus-one approach. To this end, three different predictive models from different families of classifiers were tested: random forest (RF), support vector machine (SVM) with linear kernel and multilayer perceptron (MLP). These models were chosen due to their well-known performance in application to other datasets [144]. The following hyperparameters were evaluated for each classifier:

- RF: the number of trees (*ntree*) was set to *ntree* = 250 and the number of random variables used as candidates at each split (*mtry*) was chosen from *mtry* ∈ {2, 3, 4, ..., 11, 12} in the parameter tuning process.
- SVM: a linear kernel was evaluated and the cost parameter (*C*) was chosen from $C \in \{2^{-3}, ... ,2^{0}, ... ,2^{3}\}$ in the parameter tuning process.
- MLP: a single hidden layer was chosen and the number of neurons or nodes in the hidden layer (*size*) was selected from *size* ∈ {3, 6, 9, 12, 15} in the parameter tuning process.

To evaluate these predictive models, we decided to implement a nested cross-validation (CV) scheme. The outer loop was used to evaluate the optimal number of features and the inner loop was used to evaluate the hyperparameters of each model. The structure of the nested CV approach chosen for this analysis is the same as the structure implemented in Chapter 5 (Figure 5.2).

A 5-fold CV approach was implemented in the outer loop. This resampling method randomly partitions each texture dataset into five equally sized subsets of samples or folds, maintaining a balanced amount of both classes in each fold. Then, five models are trained and tested so that each of the five folds is used once as the test set, while the four remaining folds are used to train the model. This process was repeated 10 times to reduce the variance of the cross validation results and to avoid possible bias in the random separation of the folds [173]. At the end, 50 models (5 test folds × 10 repetitions) were built using different sets of patients for training and testing each time.

The optimal number of features was evaluated in the outer loop by means of a filter feature selection approach based on the ReliefF algorithm. This feature selection step was included within the model-building process to avoid overfitting [176]. The hyperparameter tuning process was implemented in the inner loop and a 10-fold CV without repetitions was applied in this inner loop.

The classification performance was evaluated using the average area under the curve (AUC) of the receiver operating characteristic (ROC) that resulted from averaging the AUC values obtained from the 50 iterations (mean ± SD), as previously indicated in Equation 5.1. Good estimates of the model performance can be obtained using the validation data when the sample size is not large [173].

The entire classification process was implemented with the Caret package [143] in R language, version 3.2.5 (R Development Core Team, Vienna, Austria).

## 7.3. Results

### *7.3.1. Results from the Statistical Analysis*

The number of statistically significant parameters ($p < 0.05$) varied notably depending on the type and dimensionality of the ROI. Based on the results shown in Figure 7.2, texture parameters seem to be appropriate for differentiating AD patients from CN and EMCI subjects as many features turned out to be significant when comparing these groups for both circular and spherical ROIs. However, texture analysis did not seem to be useful for comparing CN and EMCI since only three features in total for both corrections turned out to be statistically significant and with borderline significance ($p \approx 0.05$).

For both right and left hippocampus, the spherical ROIs provided more significant parameters, and both ROI2 and ROI3 appeared to be effective ROIs regarding the size. In contrast, features extracted from ROI1 had a low or null significance, thus indicating that, in this case, small ROIs comprising only hippocampal tissue in MRI are not appropriate for analyzing differences between different stages of the disease with textures. Regarding the analysis of right and left hippocampus, the results were not very conclusive since the number of significant parameters was similar for both cases.

Of all the features extracted from the circular ROIs that turned out to be statistically significant ($p < 0.05$), four can be outlined: Correlation and Autocorrelation from the GLCM, and Strength and Busyness from the NGTDM. These local parameters were statistically significant after applying Bonferroni correction (more restrictive) for ROI2 and ROI3 and for right and left hippocampi when distinguishing AD from CN and EMCI. In general, in 2D ROIs local parameters were more statistically significant for all cases and regional parameters showed low or null significance.

Regarding the analysis of the spherical ROIs, the features that can be highlighted are the following: Variance and Entropy from the histogram; Contrast, Correlation, Dissimilarity, Autocorrelation and Homogeneity from the GLCM; Contrast, Busyness, Coarseness, Strength and Complexity from the NGTDM; and ZP from the GLSZM. These features were all statistically significant ($p$-value $< 0.05$) after applying Bonferroni correction when differentiating AD from CN and EMCI for ROI2 and ROI3 and for both hippocampi. In general, in 3D ROIs local parameters were also more statistically significant for all cases and it is important to mention that all NGTDM features were statistically significant for all the cases.

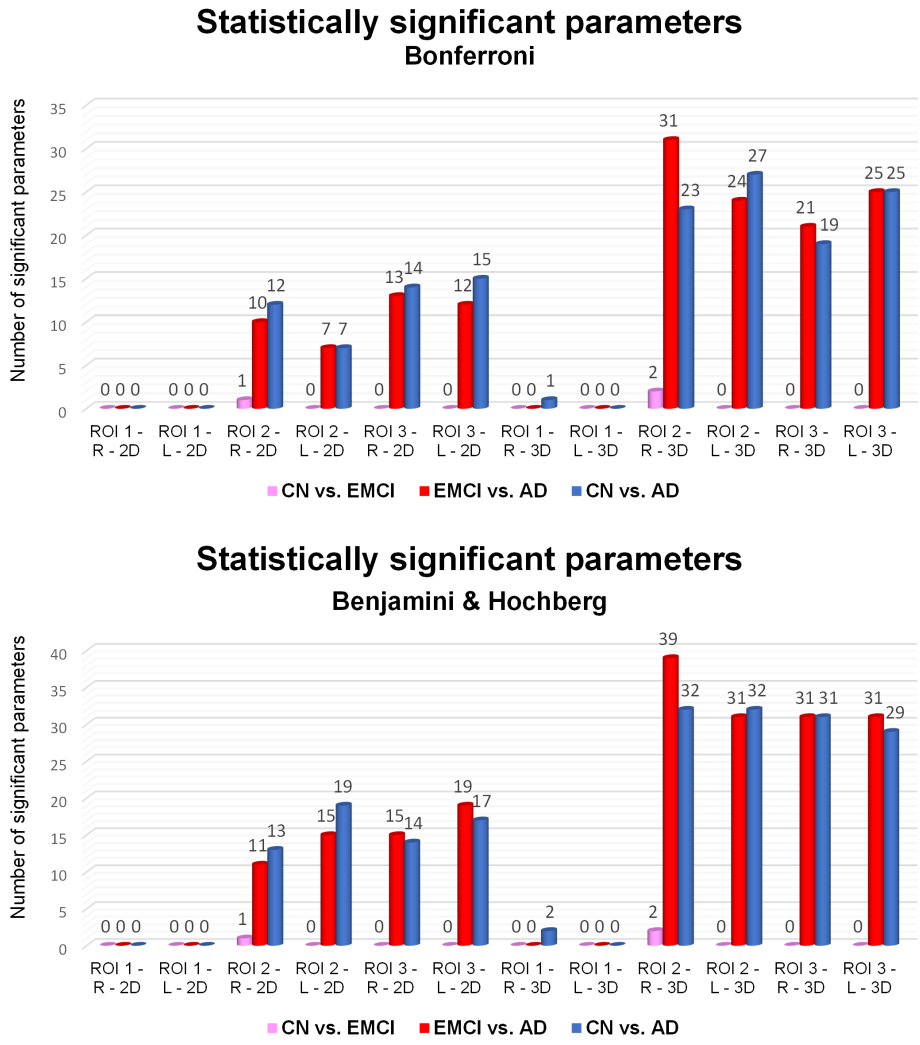



**Figure 7.2.** Number of significant parameters ($p < 0.05$) for each ROI with Bonferroni and the BH corrections. ROI1, ROI2 and ROI3 refer to the ROIs with $r = 3$, 5 and 8 pixels respectively. The suffixes "R" and "L" correspond to those ROIs placed on the right and left hippocampus respectively. The suffixes "2D" and "3D" refer to circular and spherical ROIs respectively.

Although the statistical analysis identified several texture parameters that were useful for distinguishing AD patients, we decided to apply a machine learning approach with the 3D features to confirm the statistical results and to prove if the proper combination of features allowed classifying these subjects correctly.

### *7.3.2. Results from the Machine Learning Analysis*

The results provided by the machine learning analysis confirmed that AD patients can be classified accurately from CN and EMCI subjects by using a combination of 3D texture features using non-small ROIs in both hippocampi (AUC > 0.75 in all of the cases). However, unsatisfactory results were achieved when combining texture parameters for classifying CN and EMCI subjects (AUC < 0.75 in all of the cases). Table 7.3 shows the AUC results obtained for each dataset with the three classifiers (RF, SVM and MLP).

Regarding the ROI size, parameters extracted from ROI1 were not useful to classify the groups under analysis, thus confirming the results achieved with the statistical analysis. The combination of features extracted from both ROI2 and ROI3 produced notable results. Regarding the side of the hippocampus, both right and left hippocampi produced similar results, but, in general, features extracted from the right hippocampus produced better results for ROI2 and features extracted from the left hippocampus produced better results for ROI3.

The best classification approach for distinguishing EMCI and AD subject was achieved for the RF classifier with 41 features from the ROI2 on the right hippocampus (AUC = 0.823 ± 0.080), but good results were also achieved for SVM (AUC = 0.819 ± 0.086, 13 features) and MLP (AUC = 0.816 ± 0.089, 7 features) classifiers using fewer features from the ROI3 on the left hippocampus. When classifying CN from AD subject, the best classification results were also obtained for features extracted from the ROI3 on the left hippocampus with SVM (AUC = 0.869 ± 0.071, 7 features) and MLP (AUC = 0.865 ± 0.072, 8 features) classifiers. Considering these results, a combination of features extracted from the ROI3 on the left hippocampus and an SVM or an MLP classifier is the best option for differentiating AD patients from both CN and EMCI subjects.

**Table 7.3.** AUC values obtained in the one-versus-one analysis using the RF, SVM and MLP models on the datasets of 3D features

| Classifier | RF | SVM | MLP |
|---|---|---|---|
| **CN vs. EMCI** | | | |
| ROI 1 – Right Hippocampus | < 0.6 | < 0.6 | < 0.6 |
| ROI 1 – Left Hippocampus | < 0.5 | < 0.5 | < 0.5 |
| ROI 2 – Right Hippocampus | < 0.7 | 0.710 ± 0.105 (46 features) | < 0.7 |
| ROI 2 – Left Hippocampus | < 0.7 | < 0.6 | < 0.6 |
| ROI 3 – Right Hippocampus | < 0.5 | < 0.5 | < 0.5 |
| ROI 3 – Left Hippocampus | < 0.6 | < 0.6 | < 0.7 |
| **EMCI vs. AD** | | | |
| ROI 1 – Right Hippocampus | < 0.7 | < 0.6 | < 0.6 |
| ROI 1 – Left Hippocampus | < 0.7 | < 0.6 | < 0.7 |
| ROI 2 – Right Hippocampus | 0.823 ± 0.080 (41 features) | 0.802 ± 0.084 (7 features) | 0.787 ± 0.090 (4 features) |
| ROI 2 – Left Hippocampus | 0.794 ± 0.104 (45 features) | 0.793 ± 0.099 (17 features) | 0.779 ± 0.102 (4 features) |
| ROI 3 – Right Hippocampus | 0.759 ± 0.101 (44 features) | 0.778 ± 0.096 (13 features) | 0.776 ± 0.105 (40 features) |
| ROI 3 – Left Hippocampus | 0.780 ± 0.101 (5 features) | 0.819 ± 0.086 (13 features) | 0.816 ± 0.089 (7 features) |
| **CN vs. AD** | | | |
| ROI 1 – Right Hippocampus | < 0.7 | < 0.7 | < 0.7 |
| ROI 1 – Left Hippocampus | < 0.6 | < 0.6 | < 0.6 |
| ROI 2 – Right Hippocampus | 0.820 ± 0.091 (38 features) | 0.836 ± 0.102 (34 features) | 0.834 ± 0.102 (23 features) |
| ROI 2 – Left Hippocampus | 0.807 ± 0.087 (22 features) | 0.804 ± 0.098 (10 features) | 0.799 ± 0.098 (3 features) |
| ROI 3 – Right Hippocampus | 0.807 ± 0.101 (44 features) | 0.863 ± 0.080 (44 features) | 0.843 ± 0.086 (45 features) |
| ROI 3 – Left Hippocampus | 0.829 ± 0.091 (41 features) | 0.869 ± 0.071 (7 features) | 0.865 ± 0.072 (8 features) |

* Values are shown as mean ± SD as a result over groups' estimates.
* Values of AUC < 0.7 were not expressed in the "mean ± SD" form since we considered that they are unsatisfactory and irrelevant for the conclusions of the work.

When analyzing different levels of quantization, we found that in some cases quantizing the regions with other number of gray levels improved the classification results achieved for regions quantized with NGL = 32 gray levels. In particular, we focused on the regions that produced the best classification results (ROI3 on the left hippocampus) and quantized them with other NGL (8, 16, 64 and 128 gray levels).

When using the SVM classifier, 3D textures extracted from NGL = 128 regions provided better results (AUC = 0.890 ± 0.069 with 26 features) than 3D textures from NGL = 32 regions with statistical significance (paired *t*-test, $p = 0.002$) when classifying CN from AD subjects. When classifying EMCI from AD patients with SVM, the AUC values were similar for all NGL, with no statistical significance (paired *t*-test, $p > 0.05$). When using the MLP, 3D textures extracted from NGL = 8 and NGL = 16 regions provided better results (AUC = 0.852 ± 0.074 with 46 features and AUC = 0.851 ± 0.082 with 46 features respectively) than 3D textures from NGL = 32 regions with statistical significance (paired *t*-test, $p = 0.001$ for both cases) when classifying EMCI from AD subjects. When classifying CN from AD patients with MLP, the AUC values were similar for all NGL, with no statistical significance (paired *t*-test, $p > 0.05$). Figure 7.3 shows the influence of the quantization process in the classification performance.

These results indicate that the number of gray levels used to quantize the ROIs may influence in the texture outcome and in the classification performance. However, despite this improvement, these higher AUC values were achieved by using a higher number of texture features.

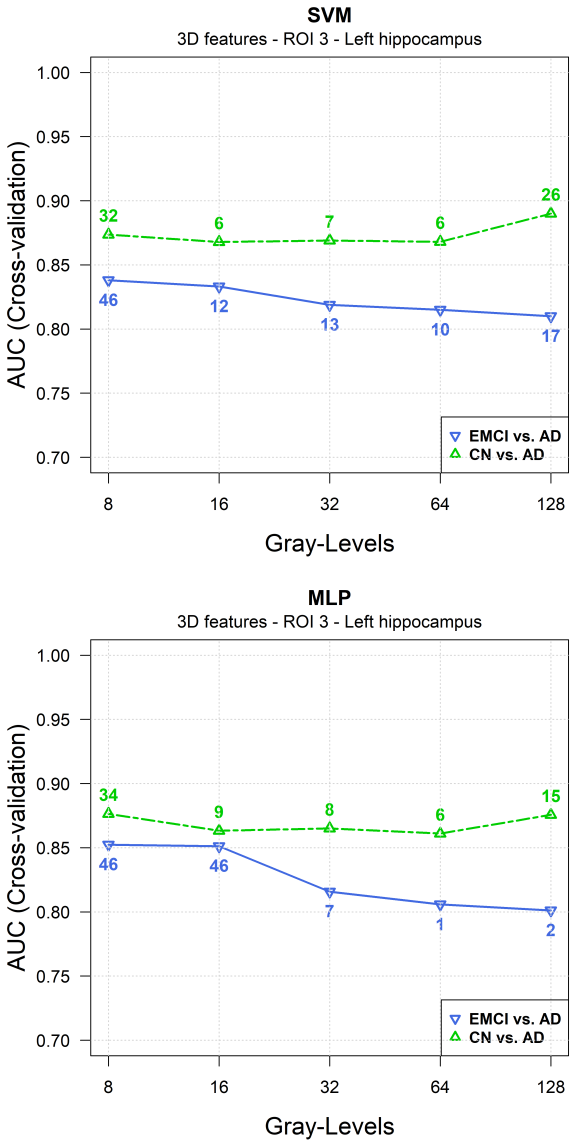



**Figure 7.3.** Comparison between the classification results obtained for different the number of gray levels when using SVM and MLP models in combination with 3D features extracted from ROI3 on the left hippocampus for differentiating AD patients from CN and EMCI subjects. The numbers on the curves indicate the number of features used to achieve the maximum AUC.

## 7.4. Discussion

Due to difficulties in accessing the brain, the diagnosis of AD is based mainly on clinical and neuropsychological tests. However, structural changes within the brain occur years before the first clinical symptoms appear and consequently the brain tissue may be damaged by the time the patient is diagnosed with AD. Therefore, there is a need to find new biomarkers of AD in its early stages and texture analysis applied on MRI may be a good approach. For this purpose, in the present study we studied the statistical significance of 2D and 3D texture features extracted from the hippocampus of T1-weighted images for differentiating between three stages of the AD: pre-symptomatic (CN), prodromal (EMCI), and advanced (AD) stages. We also evaluated with a machine learning approach if the combination of these parameters could be useful to generate classification models to distinguish between these three groups.

In the light of the obtained preliminary results, we can affirm that texture analysis is a very powerful tool that could supplement and improve AD diagnosis to a great extent. A large number of the features obtained through 2D and 3D texture analysis resulted to be statistically significant to differentiate between subjects suffering from AD and subjects from the other two populations (CN and EMCI). However, the number of significant parameters was much higher for 3D texture features. When combining these 3D texture features with machine learning techniques, good accuracy results (AUC > 0.75) were obtained. These results indicate that texture features could be helpful for detecting AD presence. However, few statistically significant parameters and poor accuracy results (AUC < 0.75) were obtained to differentiate between CN and EMCI subjects, so, with the features evaluated in this study, this first stage of the illness could not be identified. This last result is shared with most of the studies that applied machine learning to neuroimaging for detecting AD, so more efforts in this field are still needed [252]. The results also indicate that the size of the ROI has a big relevance when analyzing textures, obtaining worse results for small ROIs only including hippocampal tissue because the region is too small to capture texture differences. Finally, the number of gray levels used to quantize the ROIs also influenced in the results.

Several studies have addressed the problem of identifying the presence of early AD in MRI scans by studying texture analysis. Although some studies have focused on studying brain structures like the corpus callosum and thalamus [119], brain regions like white or gray matter [246], [251] or even the whole brain [248], most of the analyses coincide on selecting the hippocampus as the brain structure of interest. In particular, Simões *et al*. proposed a method to identify and localize early-stage AD in 3D MRI

volumes with a classification scheme based on Local Binary Patterns 3D patches [249] and local feature maps voxels [250]. In both studies they observed that the patches and voxels located at the hippocampi region were highly discriminative when detecting mild AD, specifically at the left hippocampus. On the contrary, Martínez-Murcia *et al*. [247] applied texture analysis to 90 cortical and subcortical regions in order to differentiate AD and CN patients, and they concluded that the texture measures from the right hippocampus provided higher classification results. In our study, a definitive conclusion about which hippocampus is better for classification cannot be clearly stated, since both hippocampi provided good results depending on the ROI size. However, we considered that the best results were achieved for the left hippocampus when selecting the biggest ROI (ROI3). Differences in the performance of texture analysis for both hippocampi can be explained since some studies point out that AD patients present hippocampal asymmetry, with the left hippocampus deteriorating at a higher rate, so this asymmetry may imply an early sign of the presence of AD [259], [260].

When comparing our work with related similar studies that apply texture analysis to the hippocampal region in T1-weighted MRI images, our results are in accordance with those provided by these studies. In Zhang *et al*. [58] they studied the differentiation between 17 AD and 17 CN patients with a classification approach based on histogram, gradient, GLCM and GLRLM features extracted from spherical regions situated in the hippocampi and entorhinal cortex. They determined that too small ROIs offered worse results than those of major size, a conclusion supported by the results obtained in the present work. Additionally, although the texture features tested by us and by them were not exactly the same, they highlighted four GLCM features (Difference entropy, Contrast, Homogeneity and Dissimilarity) as important features. In our case, for spherical ROIs, we highlighted thirteen features as important, and ten of them were local features (five features extracted from the GLCM and five from the NGTDM). Therefore, local heterogeneity information of the hippocampal region may play an important role for characterizing AD. Although their results were very promising, we tried to go beyond their study by increasing the number of subjects per group, including EMCI patients, comparing 2D and 3D ROIs and considering rotation invariance when computing the textures. Furthermore, we tested a more sophisticated machine learning approach, with a nested CV scheme to ensure that the results were more reliable.

In Li *et al.* [244], the hippocampi of 12 CN, 12 EMCI and 12 AD patients were segmented manually in 3D and four rotation invariant features (two GLCM features and two GLRLM features) were extracted from these ROIs. By means of statistical tests, they determined that the features RLN and GLN extracted from the right hippocampus were statistically significant ($p < 0.05$) for distinguishing between the three populations, although they speculated that their results might be biased due to the limited sample size. Additionally, they did not apply any multiple comparisons correction the *p*-value results. In our study, we increased the number of patients and we applied Bonferroni and BH corrections to the *p*-value results in order to reduce the number false positives, and this is probably the main reason why they found differences between CN and EMCI groups and we did not, thus being our analysis more accurate. Additionally, the number of extracted parameters in our work increased to 46.

One major concern about this work is related to the ROI definition. In this study, we decided to work with circular and spherical ROIs instead of segmenting manually or automatically the whole hippocampal region. We decided to analyze only textures extracted from circular and spherical ROIs because segmenting the hippocampus is a problematic, challenging task that could result in imprecise hippocampal borders, especially in 3D [256]. Moreover, delineating manual ROIs in all the patients for both hippocampi could result in a difficult, time-consuming process, not translatable to the clinical practice. In future analyses, automatic segmentation techniques for segmenting brain structures (atlas-based segmentation, for example) in a fast and accurate way should be explored. This approach would be of interest because some studies report that analyzing the whole region tissue without including surrounding structures may offer better texture characterization of the tissue. However, for small regions like the hippocampus that may present different sizes between groups of patients, geometric regions of a predefined size are recommended too. It is important to take into account that the ROI size should be sufficiently large to capture the texture information and that several texture features may be dependent on the ROI size, thus probably leading to false results due to the differences in the ROI sizes between groups [13].

Another methodological issue consisted on the lack of a spatial normalization process before defining the ROIs. As reported by Kovalev *et al.* [261], when working in the characterization or discrimination of small, equal-sized ROIs, spatial normalization can be omitted. Additionally, in the study of Zhang *et al.* [58], where they also used spherical ROIs for defining the hippocampal region, the authors stated that normalization might distort the ROIs of the MR images and destroy the texture properties of the tissue. According to these reports, we decided to keep the brain MR images in their own space

and we did not register them to the standard brain for spatial normalization because our ROIs were small in comparison to the whole brain and in order to preserve the original texture properties. In future analyses, we intend to incorporate a spatial normalization process in our pipeline to perform a registration to the standard brain and then apply an atlas-based segmentation for extracting only the hippocampal tissue. However, before including this functionality, we will have to study how to integrate this process in our MATLAB-based pipeline so that the segmentation of patients does not become a time-consuming process and we will have to analyze the impact of the spatial normalization process on the hippocampal ROI distortion and the texture outcome.

Our work showed other limitations or issues. Firstly, our study failed to find individual texture features or predictive models based on a combination of these features that were useful to classify CN and EMCI patients, meaning that a solution to accurately identify the first structural changes of the AD in the hippocampal region was not found. Further analyses should be carried out by including a wider range of texture analysis methods (Local Binary Patterns or Wavelet and Gabor transforms, for example) or by analyzing other MRI modalities in order to look for reliable texture biomarkers for the early detection of AD. Also, more patients should be included to empower the analysis.

## 7.5. Conclusion

In conclusion, our preliminary results show that 3D texture features are effective for detecting the presence of the advance stage of the AD. With further research and validation, 3D texture features could be used as biomarkers to complement the identification of the presence of AD and the specific stage of the disease in a fast and reliable way.

# Chapter 8.
# Characterization of ischemic stroke

## 8.1. Introduction and Motivation

In brain magnetic resonance imaging (MRI) of patients with small vessel disease (SVD) and older individuals, the presence of pathological attributes of similar imaging characteristics hampers the accuracy of algorithms developed to differentially assess them [262]–[264]. For example, stroke lesions (either acute or old, symptomatic or asymptomatic) in MRI can present signal intensities similar to those presented by white matter hyperintensities (WMHs), which are radiological findings of abnormal change in appearance of white matter that become more common with advancing age, and are thought to have diverse etiologies such as ischemic or demyelinating [77], [265]. Both stroke lesions and WMHs show up as areas of increased brightness when visualized by T2-weighted and fluid-attenuated inversion recovery (FLAIR) MRI sequences (Figure 8.1), so, therefore, stroke lesions may be accidentally considered as WMHs by image processing methods. Distinguishing them to disentangle the effect that each of them have in cognitive and health indicators is crucial for individual prognosis of stroke outcomes and for understanding the pathophysiology of stroke and ageing. For example, brain tissue atrophy and stroke lesion volume, but not WMH volume, have been cited as neuroimaging determinants of poststroke cognitive performance [266], and there is evidence that WMH and not old stroke lesion volume is associated with brain atrophy and cognitive decline in normal ageing [267], [268].

Routine clinical stroke neuroimaging protocols usually incorporate diffusion-weighted imaging (DWI) as it facilitates the identification and differentiation of SVD and stroke lesions. However, it has been reported that this technique does not identify the presence of the stroke on approximately a third of patients seen in clinics with a non-disabling stroke [269]. Additionally, DWI is an advanced MRI technique that is not part of the neuroimaging protocols for studies of other related processes such as ageing and dementia. Therefore, an approach for capturing the differences between the imaging profiles of different SVD related lesions and other normal ageing processes on more conventional MRI sequences would be of interest. In this work we hypothesize that texture analysis may be helpful for this purpose.

Focusing on the study of SVD and stroke lesions, several studies have successfully applied texture analysis to different tasks using MRI [77], [98], [122], [123], [270]. In particular, a recent study that used texture analysis to investigate SVD found evidences that texture features in normal-appearing tissues were able to stratify patients according to their SVD and WMH burden and to differentiate patients that had a lacunar stroke from those that had a cortical stroke, which are both subtypes of ischemic stroke [123]. Subtyping ischemic strokes is still a challenge since lacunar strokes (small infarcts resulting from the occlusion of one of the small penetrating arteries that supply the subcortical regions of the brain) can be clinically confused with cortical ischemic strokes (strokes affecting the cortical regions of the cerebral cortex). This differentiation is clinically important because the etiology and clinical management of these types of strokes may differ [35], [271], [272].

Given the effect that a stroke is known to be present not only in the affected region, but also in unaffected tissue, the main purpose of this project was to investigate the feasibility of using texture analysis in normal-appearing tissues to identify the presence or absence of a previous stroke on conventional brain MRI (T1-weighted, T2-weighted and FLAIR images). Moreover, as WMHs (i.e., a confound for the automatic identification of ischemic stroke lesions) have been defined as having mainly a vascular origin [273], we also analyzed whether texture in WMHs can help increasing the likelihood of accurately identifying the presence of a major and more sudden ischemic lesion. The main hypothesis of this project was that a proper trained classifier based on texture features in WMH or normal-appearing tissues could discriminate the brain MRI of individuals that had a stroke from those who had not, although the type of stroke (i.e., cortical versus lacunar) might be difficult to be ascertained. Specifically, our research questions were: 1) Can a texture-based automatic classifier discriminate a routine clinical structural brain MRI scan of a patient with a recent stroke of type lacunar from another

brain MRI scan but from a patient with a recent mild cortical stroke? and 2) Can a texture-based automatic classifier discriminate a structural brain MRI scan from an individual who had a previous stroke from another structural MRI scan from an individual of similar age who never had a stroke?

## 8.2. Material and Methods

### *8.2.1. Patients and Imaging Protocol*

All studies that provided data and involved human participants were conducted in accordance with the 1964 Helsinki declaration and its later amendments, with protocols and ethical standards approved by the following Scottish Research Ethics Committees: Lothian Research Ethics Committee (09/S1101/54, LREC/2003/2/29, REC 09/81101/54), the NHS Lothian R+D Office (2009/W/NEU/14), and the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) [274], [275].

To answer our two research questions we used MRI data from individuals enrolled in two different prospective studies: one study of stroke mechanisms [275] and one study of cognitive ageing [276]. The dataset extracted from the first study included MRI data from 100 patients (54 women and 46 men, mean age 65.3 years old, SD 11 years) which had a lacunar (50 patients) or mild cortical (50 patients) ischemic stroke less than 2 weeks prior to the MRI scan (i.e., post-acute stage). The dataset from the second study included MRI data from 100 individuals from a year-of-birth cohort (53 women and 47 men, mean age 73.2 years old, SD 0.6 years) who were either stroke free (50 subjects) or had a prior ischemic stroke in the non-acute phase identifiable on imaging (50 subjects). The data selection was conducted randomly and fully automatically, only taking into account that the subgroups were equal sized. The final four subgroups were: 1) recent lacunar stroke, 2) recent cortical stroke, 3) no stroke and 4) old stroke (Figure 8.1). To evaluate the influence of age in the classification into having a stroke or not, we used brain MRI data from 36 individuals from another year-of-birth cohort also enrolled in a study of cognitive ageing [277] (20 women and 16 men, mean age 91, SD 0.5 years), from which 22 never had a stroke, at least identifiable in imaging, and 14 had a previous stroke.

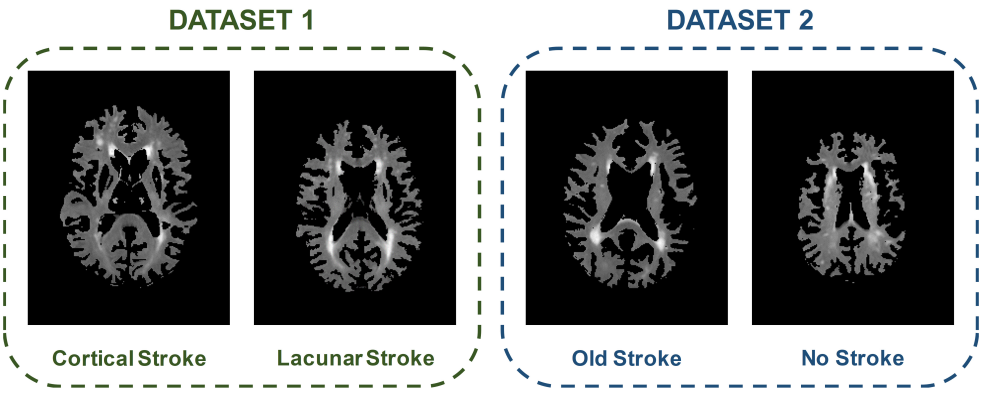



**Figure 8.1.** Examples of FLAIR MRI axial scans showing the brain white matter of four different subjects from the four groups of patients considered in this study: recent cortical stroke and recent lacunar stroke from the first dataset, and old stroke and no stroke from the second dataset. The four scans present areas of abnormal increased brightness, which may correspond to stroke lesions or to WMHs of a different nature.

All brain MRI data were acquired on a 1.5T MRI clinical scanner (Signa LX; GE Medical Systems, Milwaukee, WI, USA General Electric) equipped with a self-shielding gradient set and manufacturer supplied eight-channel-phased array heal coil. The MRI acquisition protocols of the studies that provided data for these analyses were different. The MRI sequences considered in this work were 3D T1-weighted (T1W) inversion recovery spoiled gradient echo (IR-SPGR), axial 2D T2-weighted (T2W) and axial 2D FLAIR brain images. For the stroke study the T1W sequence had repetition time/echo time (TR/TE) of 7.3/2.9 ms, flip angle of 8°, field of view of 33×21.5 $cm^2$, acquisition matrix of 256×146 and slice thickness of 1.8 mm; the T2W sequence had TR/TE of 6000/90 ms, field of view of 24×24 $cm^2$, acquisition matrix of 384×384 and slice thickness of 5 mm; and the FLAIR sequence had TR/TE of 9000/153 ms, field of view of 24×24 $cm^2$, acquisition matrix of 384×224 and slice thickness of 5 mm. Both year-of-birth cohort (normal ageing) studies had the same MRI acquisition protocol: T1W had TR/TE of 9.7/3.984 ms, field of view of 25.6×25.6 $cm^2$, acquisition matrix of 192×192 and slice thickness of 1.3 mm; T2W had TR/TE of 11320/102 ms, field of view of 25.6×25.6 $cm^2$, acquisition matrix of 256×256 and slice thickness of 2 mm; and FLAIR had TR/TE of 9000/140 ms, field of view of 25.6×25.6 $cm^2$, acquisition matrix of 256×192 and slice thickness of 4 mm.

### 8.2.2. Image Processing and Segmentation

The segmentation of the brain tissues and structures was performed following the protocol described by Valdés-Hernández *et al.* [274]. Briefly, binary masks of normal appearing white matter (NAWM) and WMH were obtained using a multispectral segmentation method [278] followed by manual editing to correct for possible errors. The structures of the basal ganglia and thalami were fully automatically extracted using a combination of three tools from the FMRIB software library (FSL) [279]: Smallest Univalue Segment Assimilating Nucleus (SUSAN), FMRIB's Linear Image Registration Tool (FLIRT) and a model-based segmentation/registration tool (FIRST), combined on an automatic pipeline developed in-house, and also manually corrected if necessary. Binary masks of NAWM, WMH and subcortical structures (SS) were mapped into the T1W, T2W and FLAIR sequences as illustrated in Figure 8.2.

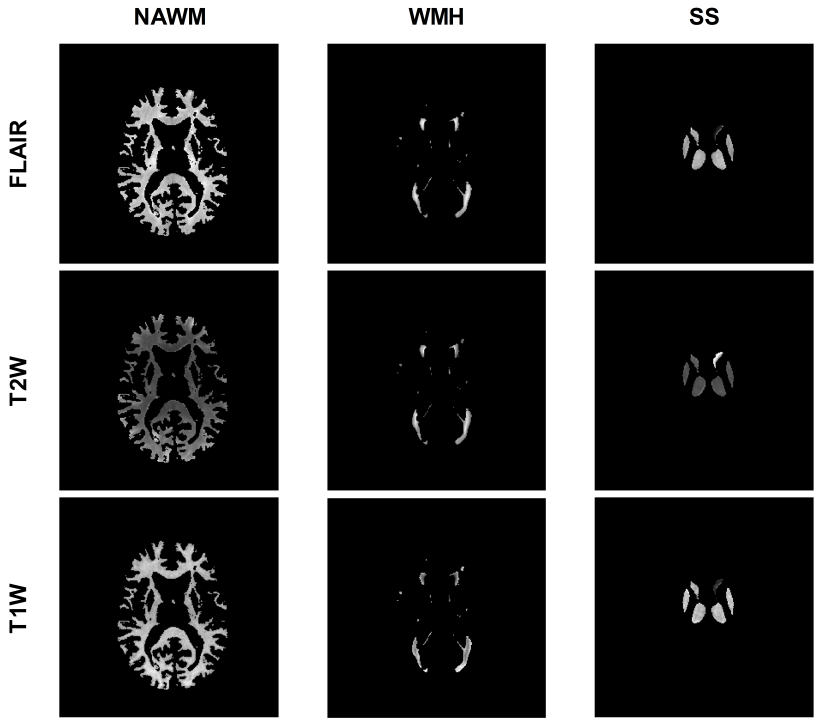



**Figure 8.2.** Set of images obtained for each patient after defining the regions of interest. In this representative case, T1-weighted (T1W), T2-weighted (T2W) and fluid-attenuated inversion recovery (FLAIR) brain images of normal appearing white matter (NAWM), white matter hyperintensities (WMH) and subcortical structures (SS) of a lacunar stroke patient are presented

### *8.2.3. 3D Texture Analysis*

A total of 18 different sets of MR images (2 prospective studies × 3 MRI sequences × 3 brain tissues/structures) were processed with texture analysis. A simple approach to capture the volumetric information of each 3D image was implemented: we first extracted the 2D texture features from each slice of each 3D image, and then the 3D texture features of the image were obtained by computing the median of the values of all the slices. This process is illustrated in Figure 8.3. Using this approach, the gray-level distributions in the third dimension are not considered; however, it has been shown that features computed with this 2D averaging method are more discriminative than features extracted from a single slice [13]. Additionally, all features were standardized to zero mean and unit variance to improve numerical stability in the model training process. Also, zero-variance and near-zero-variance predictors were removed for the same reason [205]. Finally, some features failed to give a valid numeric value for some patients (e.g., while attempting to be calculated on very small WMH clusters), so these features were also removed to avoid computational problems in the training process.

The feature extraction process was performed in MATLAB (R2015b; The MathWorks Inc., Natick, MA, USA) taking as a reference the code implemented by Alegre *et al.* [97].

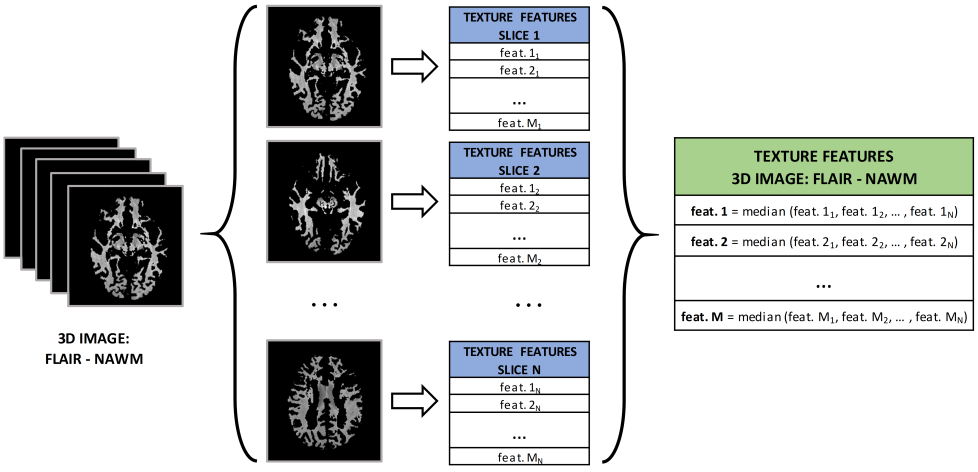



**Figure 8.3.** Process followed to extract the 3D features of a FLAIR image of the NAWM. The same process is applied to all the image of each MRI modality (FLAIR, T1W and T2W) and of each tissue/structure (NAWM, WMH and SS)

### 8.2.4. Texture Descriptors

A total of 114 features were extracted from each of the 1800 MR images (18 sets of MR images × 100 subjects enrolled in this project), and grouped into five different sets of textural features according to the texture analysis method employed: gray-level co-occurrence matrix features (GLCM: 13 parameters), gray-level run-length matrix features (GLRLM: 11 parameters), local binary patterns features (LBP: 40 parameters), wavelet statistical features (WSF: 26 parameters) and wavelet co-occurrence features (WCF: 24 parameters). Table 8.1 shows all textural features extracted from each method.

**Table 8.1.** List of the 114 texture features used in this study.

| Method | Features | Number of features |
|---|---|---|
| GLCM | Energy, Contrast, Correlation, Homogeneity, Variance, Entropy, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, First information measure of correlation (FIMC), Second information measure of correlation (SIMC) | 13 |
| GLRLM | Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-level Non-uniformity (GLN), Run-Length Non-uniformity (RLN), Run Percentage (RP), Low Gray-level Run Emphasis (LGRE), High Gray-level Run Emphasis (HGRE), Short Run Low Gray-level Emphasis (SRLGE), Short Run High Gray-level Emphasis (SRHGE), Long Run Low Gray-level Emphasis (LRLGE) and Long Run High Gray-level Emphasis (LRHGE) | 11 |
| WSF | Mean_OI, SD_OI (OI: Original image)<br>Mean_$LL_i$, Mean_$LH_i$, Mean_$HL_i$ and Mean_$HH_i$, for $i$ = 1, 2, 3<br>SD_$LL_i$, SD_$LH_i$, SD_$HL_i$ and SD_$HH_i$, for $i$ = 1, 2, 3 | 26 |
| WCF | Energy_$LL_1$, Contrast_$LL_1$, Correlation_$LL_1$, Homogeneity_$LL_1$, Entropy_$LL_1$, Variance_$LL_1$<br>Energy_$LH_1$, Contrast_$LH_1$, Correlation_$LH_1$, Homogeneity_$LH_1$, Entropy_$LH_1$, Variance_$LH_1$<br>Energy_$HL_1$, Contrast_$HL_1$, Correlation_$HL_1$, Homogeneity_$HL_1$, Entropy_$HL_1$, Variance_$HL_1$<br>Energy_$HH_1$, Contrast_$HH_1$, Correlation_$HH_1$, Homogeneity_$HH_1$, Entropy_$HH_1$, Variance_$HH_1$ | 24 |
| LBP | LBP histogram bins: $LBP_1$, $LBP_2$, $LBP_3$, … , $LBP_{36}$<br>LBP image statistics: Mean, Variance, Skewness, Kurtosis | 40 |

The GLCM quantifies the local relationship between gray levels in an image by counting the pairs of pixels separated by a predefined distance ($d$) and direction ($\theta$) that have the same distribution of gray-level values. Each pixel of the resulting matrix represents the number of times that the gray level of a reference pixel and the gray level of the neighbor pixel in the predefined distance $d$ and direction $\theta$ are seen in the image. In this study, images were uniformly quantized to 32 gray levels to reduce the computational cost of the feature extraction process and to improve the signal-to-noise (SNR) ratio [74]. A distance of $d = 1$ pixel was chosen to enhance mainly the local properties when computing the GLCM. To achieve rotation invariance, the features extracted from the GLCMs in the four directions of the 2D space ($\theta = 0°$, 45°, 90° and 135°) were averaged. Rotation invariance is important in the context of our work because some texture methods like GLCM are dependent on the direction and different texture values could be obtained if the image is rotated, thus affecting the results when images from different patients have different orientations [13].

The GLRLM describes regional heterogeneity information by examining the times that each gray level value is seen consecutively in an image in a predefined direction. The GLRLM is constructed by detecting and counting the runs (sequences of consecutive pixels with the same gray level) of different gray levels and lengths in the image. To compute the GLRLMs, images were previously quantized to 32 gray levels as in the case of GLCMs. The GLRLM features are also affected by the orientation of the image, so features extracted from the GLCMs in the four directions of the 2D space ($\theta = 0°$, 45°, 90° and 135°) were averaged to achieve rotation invariance.

The LBP method labels each pixel of the image under analysis by comparing its gray level with the gray levels of the surrounding pixels and then assigning a specific binary number. This binary number for each pixel is obtained by allocating a value of 1 to those surrounding pixels with a greater gray level value and a 0 to those surrounding pixels with a lower gray level value. Originally, LBP was defined for patches of 3×3 pixels, but it was later extended for blocks of $P$ surrounding pixels separated by a distance $R$. In this work, the original LBP operator (patches of $3 \times 3$ pixels: $P = 8$, $R = 1$) was employed to preserve the texture analysis as local as possible because regions like WMH are not very large. Rotation invariance was achieved by performing a circular bit-wise right shift operation (rotating the neighbor pixel set clockwise) and assigning the smallest LBP binary number [83]. Using this approach, 36 unique rotation invariant histogram-based LBP features were obtained, as only 36 LBP binary numbers can occur for $P = 8$. Additionally, 4 statistics derived directly from the LBP image (mean, variance, skewness and kurtosis) were added to the LBP features set. The MR images were not

quantized to compute the LBP texture feature since the rotation invariant LBP approach is robust to intensity variations [94].

The DWT examines the spatial frequency patterns of an image within different scales and frequency directions, considering that frequency is directly proportional to gray level variations in an image. The DWT applied to an image produces four matrices of coefficients (subimages) that represent the approximations or low frequencies (LL: low-low) and the details or high frequencies in the vertical (LH: low-high), horizontal (HL: high-low) and diagonal (HH: high-high) directions. An example of this matrices is shown in Figure 8.4. The DWT can be repeated consecutively to achieve a major image decomposition: the first level of decomposition ($LL_1$, $LH_1$, $HL_1$ and $HH_1$) is applied to the original image as mentioned before and the subsequent levels are applied to the matrix of approximations of the previous level ($LL_i$, $LH_i$, $HL_i$ and $HH_i$, where $i$ is the level of decomposition). In this work we examined two groups of texture features derived from the DWT. The first group was the Wavelet statistical features (WSF), consisting of 26 descriptors that are the mean and SD of the histograms of the original image and the subimages yielded after three levels of decomposition. The second group was the Wavelet co-occurrence features (WCF), consisting of 24 descriptors that are obtained by extracting six of the GLCM features (energy, contrast, correlation, homogeneity, entropy and variance) from the subimages yielded after the first DWT decomposition. The Haar family of wavelets was used to perform the DWT decomposition.

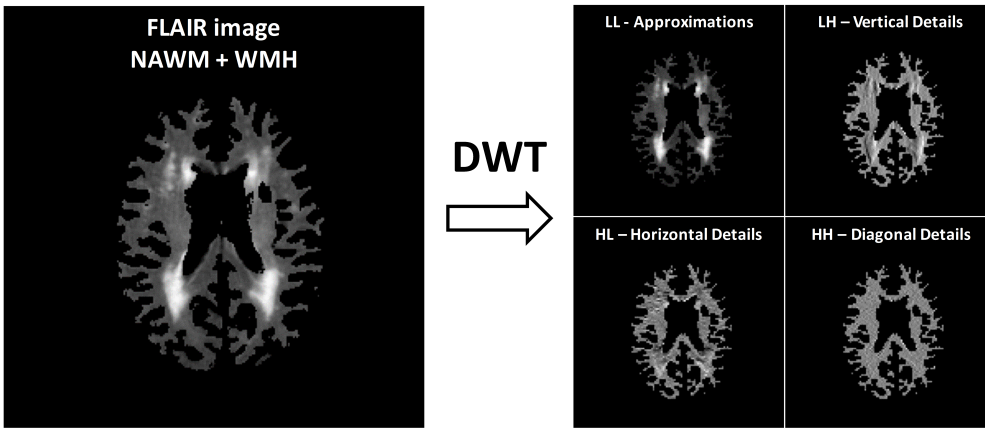



**Figure 8.4.** First DWT level of decomposition of a FLAIR image of the white matter tissue (NAWM and WMH) of a single brain slice.

### 8.2.5. Statistical Analysis

Prior to the evaluation of the texture features using machine learning techniques, a preliminary statistical analysis was conducted to evaluate the discriminative power of each feature independently between populations. Its final purpose was to assess the feasibility of these features individually as biomarkers of stroke. To compare the distributions of each textural feature for each of the classes, the Mann-Whitney-Wilcoxon (MWW) test, also called Mann-Whitney *U* test, was applied. This non-parametric test is equivalent to the independent samples *t*-test but without the requirement of the normality assumption and is recommended for relatively small sample sizes [258]. As the number of statistical tests performed increases, the contrast test becomes more permissive, thus rejecting the null hypothesis more easily and increasing the number of false positives [258]. To counter this effect, usually known as the multiple comparisons problem, we decided to apply the Holm-Bonferroni correction before assuming the statistical significance of the features. This relatively strong (conservative) method controls the family-wise error rate, thus compensating the Type I error (incorrectly rejecting the null hypothesis) and attempting to limit the probability of even one false discovery.

### 8.2.6. Classification Approach

Two well-known conventional classifiers were trained and evaluated in this work: Support Vector Machine with linear kernel and Random Forest. The Support Vector Machine (SVM) classifier, in a binary classification task like ours, tries to maximize the margin distance between the classification boundary (i.e., hyperplane) and the closest samples of both classes by adjusting internal parameters in the training process. One of these parameters is the cost $C$, which controls the trade-off between misclassification of the training data and the size of the margins. Values of $C = 2^{-3}$, $2^{-2}$, $2^{-1}$, 1, 2, $2^2$ and $2^3$ were tuned to obtain the optimal classification results. We used a linear kernel after an initial evaluation where non-linear kernels did not produce notably better results even after a lengthy training process. The Random Forest (RF) classifier combines the results of a multitude of independent and decorrelated decision trees in the training process, thus improving generalization of the model and robustness against overfitting especially in small sample sizes problems like ours. The parameter *mtry*, which identifies the number of random variables used in each tree, controls the strength (how accurate the individual trees are) and the correlation (the dependence between trees) of the RF model. Another tuning parameter is the number of trees to be built. In

this work, values of *mtry* = 2, 4, 6, 8, 10 and 12 were evaluated and the number of trees was set to 250, as higher values of this parameter did no produce notably better results on a preliminary evaluation.

For evaluating the efficiency of the classification models, we employed a 5-fold cross-validation (CV) approach. This resampling method randomly partitions each texture dataset into five equally sized subsets of samples or folds, maintaining a balanced amount of both classes in each fold. Then, five models are trained and tested so that each of the five folds is used once as the test set, while the four remaining folds are used to train the model. This process was repeated ten times to reduce the variance of the cross validation results and to avoid possible bias in the random separation of the folds [173], so at the end 50 models (5 test folds × 10 repetitions) were built using different sets of patients for training and testing each time. The classification performance was evaluated using the averaged area under the curve (AUC) of the receiver operating characteristic (ROC) that resulted from averaging the AUC values obtained from the 50 iterations (mean ± SD). Good estimates of the model performance can be obtained using the validation data when the sample size is not large [173]. Other metrics like sensitivity, specificity and accuracy were also obtained to validate the results.

A total of 90 texture dataset (18 sets of MR images × 5 texture analysis methods) were firstly examined with the classifiers without excluding any texture feature. However, the texture combinations that provided the highest AUC values were analyzed again using the same cross validation structure with a feature selection step included within the model-building process to avoid overfitting [176]. This way, we could test if reducing the number of features improved the classification results. Two filter feature selection methods were applied to obtain rankings of features based on the discriminative power of each feature independently without analyzing the relation between features and without involving any predictive model [162]. The first method used the *p*-value provided by the MWW test for independent groups of samples. The second method used the Maximal Information Coefficient (MIC), which measures the strength of the linear or non-linear association between two variables.

The model evaluation process was implemented with the Caret package [143] in R language, version 3.2.5 (R Development Core Team, Vienna, Austria). The structure of the classification approach chosen for this analysis is the same as the one schematically represented in Chapter 5 (Figure 5.2).

## 8.3. Results

### *8.3.1. Discrimination between cortical and lacunar stroke patients*

The first analysis consisted on finding differences between patients diagnosed with recent cortical stroke and lacunar stroke. Firstly, a basic statistical analysis was carried out before applying the machine learning approach to estimate the discrimination power of the features individually. Texture features did not seem to be useful to discriminate between recent cortical and lacunar stroke patients individually for any MRI sequence and any brain tissue or structure. Sixty-one texture features of a total of 1026 features (114 features × 3 MRI sequences × 3 brain tissues/structures) were statistically significant ($p < 0.05$) when applying an MWW test for independent groups of samples, but only two features derived from the GLCM (FIMC and SIMC, with $p = 0.0218$ and $p = 0.0096$ respectively) were significant after applying a Holm-Bonferroni correction for multiple comparisons. Table 8.2 shows the distribution of significant features according to the MRI sequence and the brain tissue/structure. Based on the data presented in this table, T1W images seem to be the images where texture information allows discriminating cortical and lacunar stroke patients more accurately, especially when analyzing the brain SS. Nevertheless, the texture data extracted from these images and these brain tissues/structures did not seem to have enough discriminative power to classify precisely cortical and lacunar stroke patients in general.

**Table 8.2.** Number of significant features ($p < 0.05$) for discriminating CS vs LS patients before (numerator) and after (denominator) Holm-Bonferroni correction for multiple comparisons per MRI sequence and brain region (tissue or structure).

| REGION / SEQUENCE | NAWM | SS | WMH |
|---|---|---|---|
| **FLAIR** | 0 / 0 | 16 / 0 | 1 / 0 |
| **T2W** | 3 / 0 | 9 / 0 | 1 / 0 |
| **T1W** | 11 / 0 | 19 / 2 | 1 / 0 |

The classification results confirmed the assumption provided by the previous statistical analysis: the texture features tested in this study are not useful to discriminate between cortical and lacunar stroke patients. Table 8.3 shows the averaged AUC (mean ± SD) computed from the 50 iterations when examining all the texture datasets with the two models under analysis (SVM with linear kernel and RF), and for all the MRI sequences and brain tissues/structures. A relevant AUC value could not be obtained (AUC < 0.7) in any case. The best result was obtained using the GLCM parameters in the T2W images of the NAWM region using SVM with linear kernel (AUC = 0.667 ± 0.117), but this value was not accurate enough to determine that a good classification can be achieved using these data.

### *8.3.2. Discrimination between patients with and without stroke*

The second analysis consisted on finding differences between patients who presented an old stroke and patients who had not suffer a stroke at the time of the imaging evaluation. A basic statistical analysis was also performed before applying the machine learning approach to estimate the discrimination power of the features individually. Many texture features showed capability for discriminating between “no stroke” and “old stroke” patients. In this case, 349 texture features out of 1026 features (114 features × 3 MRI sequences × 3 brain tissues/structures) were statistically significant ($p < 0.05$) when applying an MWW test for independent groups of samples, and 235 features remained significant after applying a Holm-Bonferroni correction for multiple comparisons. Table 8.4 shows the distribution of significant features according to the MRI sequence and the brain tissue/structure. The information collected in this table indicates that the texture features extracted from the brain SS are more effective in discriminating between “no stroke” and “old stroke” patients, regardless of the MRI sequence is used.

**Table 8.3.** Results of the classification analysis for cortical and lacunar stroke patients. The AUC values are shown for the two models and for all the MRI sequences and brain tissues/structures when using the texture features extracted from the 5 texture analysis methods.

| **AUC: Mean (SD)** | | ***FLAIR*** | | ***T2W*** | | ***T1W*** | |
|---|---|---|---|---|---|---|---|
| | | **RF** | **SVM** | **RF** | **SVM** | **RF** | **SVM** |
| ***GLRLM*** | **NAWM** | <0.6 | <0.6 | <0.6 | <0.5 | <0.5 | <0.5 |
| | **SS** | <0.6 | <0.6 | <0.6 | 0.622 (0.125) | <0.5 | <0.5 |
| | **WMH** | <0.6 | <0.5 | <0.6 | <0.6 | <0.5 | <0.6 |
| ***GLCM*** | **NAWM** | <0.6 | <0.5 | <0.6 | **0.667 (0.117)** | <0.6 | <0.6 |
| | **SS** | <0.6 | <0.6 | 0.611 (0.121) | <0.5 | <0.6 | 0.637 (0.140) |
| | **WMH** | <0.6 | <0.5 | <0.5 | <0.5 | <0.6 | <0.6 |
| ***LBP*** | **NAWM** | <0.5 | <0.6 | <0.5 | <0.6 | <0.6 | <0.6 |
| | **SS** | <0.6 | 0.604 (0.121) | <0.5 | <0.5 | <0.6 | <0.6 |
| | **WMH** | <0.6 | <0.6 | 0.616 (0.107) | <0.6 | <0.5 | 0.616 (0.092) |
| ***WCF*** | **NAWM** | <0.5 | <0.6 | <0.5 | <0.6 | <0.6 | <0.6 |
| | **SS** | 0.600 (0.118) | <0.5 | <0.6 | 0.604 (0.129) | <0.6 | <0.5 |
| | **WMH** | <0.5 | <0.5 | <0.6 | <0.6 | <0.5 | <0.5 |
| ***WSF*** | **NAWM** | <0.5 | <0.6 | <0.4 | 0.621 (0.140) | **0.650 (0.114)** | 0.618 (0.128) |
| | **SS** | <0.5 | <0.5 | 0.605 (0.117) | <0.6 | <0.6 | <0.5 |
| | **WMH** | <0.5 | <0.5 | <0.4 | **0.661 (0.132)** | <0.5 | <0.6 |

* Values are shown as mean ± SD as a result of averaging the results of the validation data.
* Values in bold indicate the three best AUC results (AUC > 0.65).

**Table 8.4.** Number of significant features ($p < 0.05$) for discerning between "old stroke" and "no stroke" patients before (numerator) and after (denominator) Holm-Bonferroni correction for multiple comparisons per MRI sequence and brain tissue/structure.

| REGION / SEQUENCE | NAWM | SS | WMH |
|---|---|---|---|
| **FLAIR** | 5 / 1 | 79 / 71 | 30 / 9 |
| **T2W** | 1 / 0 | 79 / 72 | 34 / 3 |
| **T1W** | 20 / 4 | 71 / 66 | 30 / 9 |

The classification results were more optimistic in this analysis: according to the AUC values obtained, we can state that certain groups of textures allowed "no stroke" and "old stroke" patients to be classified with a good degree of precision. Table 8.5 shows the averaged AUC (mean ± SD) obtained from the 50 iterations when examining all the texture datasets with the two models (SVM with linear kernel and RF), and for all the MRI sequences and brain tissues/structures. Good results were not achieved with all groups of textures, but in several cases, AUCs higher than 0.75 were obtained. For example, LBP features extracted from T2W and FLAIR images of the SS delivered good results as expected from the previous statistical analysis. However, other feature datasets like GLRLM features extracted from FLAIR images of the WMH or WCF features extracted from FLAIR images of the NAWM provided satisfactory results although the previous statistical analyses were not very optimistic with features extracted from these groups of images. It should be noted that parameters extracted from the T1W images as well as parameters derived from the GLCM did not provide relevant AUCs values in any case (AUC < 0.7). The predictive model employed for classifying the patients influenced the results, but there was not a firm conclusion on which model was better as SVM worked better with some texture datasets and RF with others.

**Table 8.5.** Results of the classification analysis for "old stroke" and "no stroke" patients. The AUC values are shown for the two models and for all the MRI sequences and brain tissues/structures when using the texture features extracted from the 5 texture analysis methods.

| **AUC: Mean (SD)** | | ***FLAIR*** | | ***T2W*** | | ***T1W*** | |
|---|---|---|---|---|---|---|---|
| | | **RF** | **SVM** | **RF** | **SVM** | **RF** | **SVM** |
| ***GLRLM*** | **NAWM** | <0.6 | <0.6 | <0.5 | 0.665 (0.084) | 0.609 (0.104) | <0.6 |
| | **SS** | 0.691 (0.109) | 0.676 (0.097) | 0.643 (0.099) | 0.738 (0.121) | 0.654 (0.112) | 0.662 (0.104) |
| | **WMH** | 0.674 (0.108) | **0.770 (0.089)** | <0.6 | 0.646 (0.128) | <0.6 | <0.5 |
| ***GLCM*** | **NAWM** | <0.6 | <0.5 | <0.5 | 0.601 (0.138) | <0.6 | 0.642 (0.126) |
| | **SS** | 0.612 (0.099) | 0.666 (0.090) | 0.641 (0.107) | 0.644 (0.111) | 0.662 (0.091) | <0.6 |
| | **WMH** | 0.608 (0.111) | 0.614 (0.124) | 0.617 (0.102) | <0.5 | 0.659 (0.113) | <0.5 |
| ***LBP*** | **NAWM** | <0.5 | <0.5 | <0.5 | <0.6 | 0.667 (0.125) | <0.6 |
| | **SS** | 0.742 (0.100) | **0.751 (0.103)** | 0.680 (0.112) | **0.763 (0.116)** | 0.649 (0.120) | 0.676 (0.122) |
| | **WMH** | <0.6 | 0.682 (0.136) | 0.608 (0.116) | 0.671 (0.122) | 0.611 (0.140) | 0.630 (0.126) |
| ***WCF*** | **NAWM** | **0.761 (0.097)** | 0.637 (0.121) | <0.5 | <0.5 | 0.624 (0.105) | 0.628 (0.143) |
| | **SS** | 0.647 (0.099) | <0.6 | 0.680 (0.097) | 0.608 (0.157) | 0.682 (0.109) | <0.6 |
| | **WMH** | 0.702 (0.108) | <0.6 | **0.752 (0.097)** | <0.5 | 0.664 (0.115) | <0.6 |
| ***WSF*** | **NAWM** | 0.669 (0.114) | 0.637 (0.137) | <0.5 | <0.6 | <0.6 | <0.6 |
| | **SS** | 0.635 (0.094) | <0.6 | 0.705 (0.116) | 0.737 (0.103) | <0.5 | <0.5 |
| | **WMH** | <0.6 | <0.6 | 0.665 (0.103) | 0.677 (0.123) | <0.6 | <0.6 |

* Values are shown as mean ± SD as a result of averaging the results of the validation data.
* Values in bold indicate the five best AUC results (AUC > 0.75).

**Influence of the feature selection on the classification results**

We applied two filter feature selection methods to the five texture datasets that yielded better results in terms of AUC (AUC > 0.75) to see if better classification results were achieved when reducing the number of features. Rankings of features based on the maximal information coefficient (MIC) and the *p*-value provided by the MWW test were computed from the training folds in each of the 50 iterations of the CV procedure. Table 8.6 shows the new AUC values obtained when reducing the number of features according to the computed rankings in the best texture datasets. In general, better AUC values were obtained when reducing the number of features. In particular, it is remarkable the substantial improvement achieved for LBP descriptors extracted from T2W images of the SS when using the SVM model: a final value of AUC = 0.828 ± 0.075 was obtained when only using the 7 more relevant LBP characteristics according to the MIC statistic. Figure 8.5 shows the classification performance profile, which reflects the AUC values obtained for all possible subsets of features according to the MIC ranking, and the ROC curves provided by the model when using all the features and when using the optimal number of features.

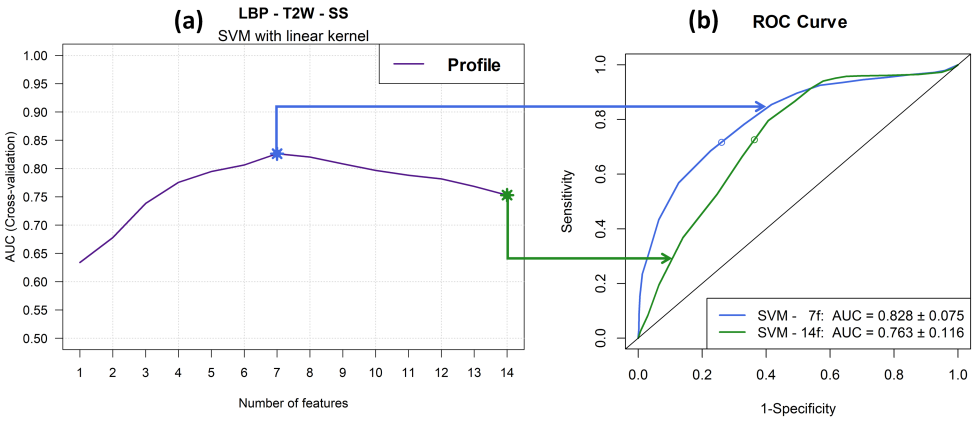



**Figure 8.5.** Results of applying the feature selection method based on MIC to the texture dataset of LBP features extracted from T2W images of SS when training the SVM model. The profile of AUC values obtained for all possible subsets of features according to the MIC ranking is illustrated in (a). The ROC curves provided by the model when using all the features (14 features) and when using the optimal number of features (7 features) is shown in (b).

**Table 8.6.** Values of AUC obtained when analyzing the best texture datasets without applying feature selection, that is, using all the features of the dataset, and with feature selection, that is, reducing the number of features based on two metrics: the *p*-value and the MIC.

| AUC: Mean (SD) | | RF | SVM |
|---|---|---|---|
| ***GLRLM FLAIR - WMH*** | **All features** | 0.674 (0.108) | 0.770 (0.089) |
| | ***p*-value** | = | 0.773 (0.089) |
| | **MIC** | = | 0.773 (0.093) |
| ***LBP FLAIR - SS*** | **All features** | 0.742 (0.100) | 0.751 (0.103) |
| | ***p*-value** | = | = |
| | **MIC** | 0.744 (0.104) | 0.759 (0.103) |
| ***LBP T2W - SS*** | **All features** | 0.680 (0.112) | 0.763 (0.116) |
| | ***p*-value** | 0.693 (0.101) | 0.774 (0.099) |
| | **MIC** | 0.714 (0.113) | 0.828 (0.075) |
| ***WCF FLAIR - NAWM*** | **All features** | 0.761 (0.097) | 0.637 (0.121) |
| | ***p*-value** | 0.766 (0.099) | 0.713 (0.125) |
| | **MIC** | 0.766 (0.086) | 0.712 (0.112) |
| ***WCF T2W - WMH*** | **All features** | 0.752 (0.097) | <0.5 |
| | ***p*-value** | = | <0.6 |
| | **MIC** | = | = |

* Values are shown as mean ± SD as a result of averaging the results of the validation data.
* The symbol "=" is used when no improvement is obtained by reducing the number of features.

### Influence of age in the classification results

To study the influence of age in the performance of the classifier, we introduced 36 additional datasets of images of older patients (91 years against 73 years of mean age) acquired with the same imaging parameters. We added the texture features extracted from these patients to the previous texture datasets and repeated the classification performance for the best texture datasets to study how these older patients affect the results. Table 8.7 shows the results obtained with and without including the texture data from the older patients in the texture datasets that performed better in the previous analysis. The results show that the classification performance got worse when introducing older patients in general, suggesting that the age influence the classification results by increasing the misclassification rate, possibly because in older patients the SVD is more manifest and the images present more SVD markers.

**Table 8.7.** Values of AUC obtained when analyzing the best texture datasets (without feature selection) with and without including the textures extracted from the additional older patients.

| AUC: Mean (SD) | | RF | SVM |
|---|---|---|---|
| ***GLRLM FLAIR - WMH*** | **WITHOUT older patients** | 0.674 (0.108) | 0.770 (0.089) |
| | **WITH older patients** | 0.682 (0.089) [a] | 0.736 (0.084) |
| ***LBP FLAIR - SS*** | **WITHOUT older patients** | 0.742 (0.100) | 0.751 (0.103) |
| | **WITH older patients** | 0.655 (0.098) | 0.644 (0.106) |
| ***LBP T2W - SS*** | **WITHOUT older patients** | 0.680 (0.112) | 0.763 (0.116) |
| | **WITH older patients** | 0.623 (0.078) | 0.670 (0.083) |
| ***WCF FLAIR - NAWM*** | **WITHOUT older patients** | 0.761 (0.097) | 0.637 (0.121) |
| | **WITH older patients** | 0.645 (0.086) | 0.580 (0.106) |
| ***WCF T2W - WMH*** | **WITHOUT older patients** | 0.752 (0.097) | <0.5 |
| | **WITH older patients** | 0.678 (0.074) | <0.5 |

* Values are shown as mean ± SD as a result of averaging the results of the validation data.
[a] Exception where the AUC increased after adding older patients

**Influence of the type of clinical stroke in the classification results**

The patients for which the best models performed well 80% or more of the times across the 50 validation iterations were identified. From these best models, the patients for which both classifiers (i.e., SVM and RF) performed well 80% or more of the times, were also identified. For each patient, we extracted the following data: 1) proportion of times the images were correctly (and wrongly) classified, 2) proportion of times in which both classifiers correctly (and incorrectly) classified the images using the same descriptors, 3) clinical stroke classification into no stroke, large cortical, small cortical or lacunar. The pattern of the classification performance of the different stroke subtypes was similar irrespective of the classifier used, and when the analysis accounted for whether the images were correctly (or incorrectly) classified by both classifiers, “no stroke” images achieved the greatest proportion of well classified, followed by “lacunar”, “small cortical and “large cortical” (Figure 8.6).

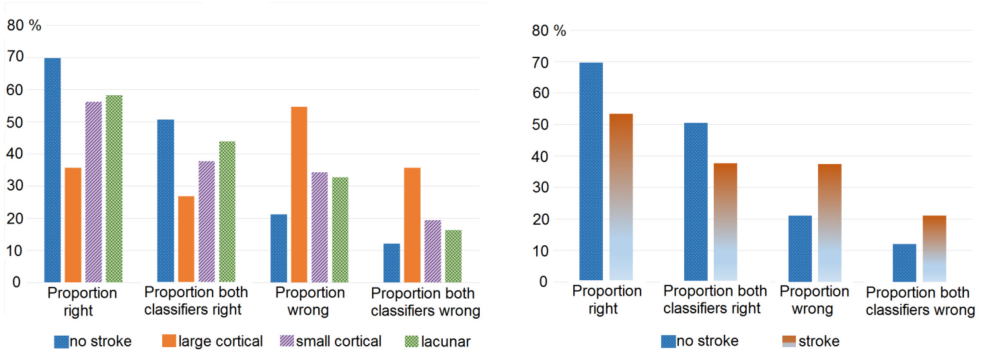



**Figure 8.6.** Pattern of the classification performance of the best models (i.e., for which the accuracy was above 80%) per stroke subtype (i.e., no stroke, large cortical, small cortical or lacunar) (left) and per stroke occurrence (i.e., had stroke or not) (right).

## 8.4. Discussion

In this project, the performance of a substantial amount of texture features extracted from different brain tissues and structures (i.e., NAWM, WMH and SS) in different MRI structural sequences (i.e., FLAIR, T2W and T1W) were analyzed, in two conventional machine learning approaches, to identify the presence of stroke in the images of post-acute stroke patients and normal ageing individuals. Differentiation of post-acute cortical versus lacunar stroke subtypes using texture analysis was examined as well as classification of images with and without chronic stroke lesions. We were not able to find a proper machine learning approach to discriminate between patients with cortical and lacunar stroke with texture features (AUC $< 0.7$). On the contrary, the machine learning scheme used in this study provided promising results for discerning between patients presenting an old stroke and normal-ageing patients who never had a stroke (AUC $> 0.75$). The results improved when using a feature selection approach to reduce the number and identify those features that may obscure the classification performance, especially when evaluating LBP descriptors extracted from T2W images of the SS (AUC $> 0.8$).

Regarding the classification of cortical and lacunar stroke patients, only two textural features from the SS in T1W images resulted with high discriminatory power between images from post-acute lacunar vs. cortical stroke in the previous statistical analysis. These features were the first and second information measures of correlation

(FIMC and SIMC) derived from the GLCM, which quantify the linear dependency or correlation between intensities, thus representing homogeneity but adding some desirable properties that are not represented by the original correlation descriptor extracted from the GLCM [87]. A previous study by Valdés-Hernández *et al.* that evaluated the use of texture analysis as an alternative for characterizing SVD and assessing possible blood brain barrier leakage [123] reported differences in the texture outcome of the FLAIR deep gray matter between post-acute lacunar and cortical stroke patients, but only with borderline significance. This study reported that the texture pattern in the deep gray matter was more homogeneous in patients with recent lacunar stroke compared to those who had a cortical type. Statistically significant differences between the FLAIR images from both groups of patients were only found in the texture features measured in the post-acute stroke lesions. Our motivation was to explore whether the texture in normal (as opposed to abnormal) tissues could have enough discriminatory power to be used in machine learning schemes to identify the stroke subtype and if there was a stroke. Therefore, we did not analyze the texture outcome in the stroke lesions exclusively. Our analysis managed to find features with borderline statistical significance to discriminate between cortical and lacunar stroke patients in T1W MRI data but failed to find a conventional machine learning model to classify these patients accurately. The reason behind these results may lie in the fact that both types of stroke can be seen simultaneously in many cases, as reported by Xu [280].

Regarding the classification of patients with and without stroke evidences, the statistical analysis and the classification results indicated that texture features are suitable for characterizing the presence of an old stroke lesion in MRI against imaging patterns of normal-ageing patients. The inclusion of older patients in the analysis (91 years against 73 years of mean age) influenced negatively in the classification, thus suggesting that, in patients of advanced age, the imaging patterns for detecting an old stroke may be confounded with normal-ageing imaging patterns because of the presence of more SVD markers due to age. When evaluating the stroke subtypes, the images from “no stroke” patients were, in general, better identified by the classifiers as opposed to the images that had “large cortical” chronic stroke lesions, which resulted in them being the less well classified, and, instead, were classed as not having any stroke lesion at all. It might result paradoxically, given that the images of individuals with chronic lacunar lesions (i.e., small lesions mainly in the region crossed by the corticospinal tracts [281], which can be confounded by WMH), were the second best classified. However, lacunar, and not large cortical, strokes have been associated with blood brain barrier impairment, manifested in abnormal extracellular leakage [275]. Also, textural features in normal and

abnormal tissues have been reported as being useful in detecting the subtle differences that this mechanism causes. Therefore, our results are in-line with the relevant clinical literature.

One methodological limitation of this work was the impossibility of combining both datasets to analyze if images showing recent cortical or lacunar strokes could be distinguished from images of patients without stroke and patients with an old stroke lesion. This is due to the fact that variations in acquisition parameters may result in differences in the texture outcome that are not due to the underlying biological characteristics of the tissues expressed by the texture patterns [71], [108]. Image normalization techniques help reducing these differences in imaging acquisition settings, but some residual effects may not be totally suppressed, thus obscuring the true texture differences due to the tissue properties only. Therefore, combining texture features extracted from both databases in the proposed machine learning pipeline may lead to overoptimistic results caused by the differences in imaging acquisition protocols.

Other limitation of the present work consists on the 3D texture analysis approach based on the median of 2D texture features. Although pure 3D texture analysis is usually preferred because it allows capturing more heterogeneity information of the tissue under analysis, this approach is not always feasible, especially when the slice thickness of the images is very large compared to the in-plane resolution [51], as in our case for T2W and FLAIR images. In these situations, 2D texture approaches or approaches like the one carried out are recommended.

## 8.5. Conclusion

In this work we conducted a very detailed texture analysis study for identifying and characterizing ischemic stroke lesions in structural brain MRI data by considering several regions or tissues and by testing a large amount of quantitative texture descriptors. The number of patients per group was sufficiently large to draw reliable conclusions and the machine learning pipeline was designed to avoid overoptimistic and overfitted results. We achieved promising preliminary results that suggested that texture features may help in the detection of a stroke lesions.

# Chapter 9.
# Final conclusions

Radiomics analyses have shown an increased interest in the past years and the reasons behind this interest seem obvious. The intrinsic information provided by medical images has been missed and omitted for too long, but now, the new and expanding advances in technology for acquiring high-quality images and processing the underlying information in a rapid and intelligent way have permitted exploiting these "hidden" data, thus revolutionizing the field of radiology. In this thesis, we provide four original contributions to help clinicians in the evaluation of different brain disorders by means of a radiomics approach based on texture analysis in conventional MRI. Our preliminary results prove the potential of this practice for defining and characterizing brain lesions in a fast, reliable and non-invasive way. It is important to remark that the objective of conducting these feasibility projects is not to substitute the radiologic assessment made by clinicians but to provide more diagnostic information complementing the work of clinicians so as to facilitate the decision-making tasks.

The question now is clear: *are we on the right track or are we wasting our time in an endless road?* We strongly believe that radiomics analyses are the future of radiology and the preliminary results presented in this thesis support this conclusion. However, we still have a long way to go since there is still a clear need for standardization and validation. The technology necessary for accomplishing radiomics tasks is already available, but now we need larger imaging datasets standardized across institutions in order to validate the models created for helping in the diagnosis and assessment of diseases. Furthermore, the corresponding demographic, clinical, histologic or genomic data should be obtained when possible for each imaging evaluation with the purpose of correlating and combining these data with the imaging information. The latter would increase the reliability of the models and might result in important discoveries about the direct connection between parameters that in principle did not seem to show any kind of relationship.

Regarding the correlation of data of different nature, some studies have attempted to reflect on the connection between texture features and histopathologic features. The question to answer in these studies is clear: *are textural and histopathological parameters linked because they measure the same biology in different ways?* Most of these studies suggest that, whereas a single feature cannot still be directly linked to a specific biologic process, it is possible to assume that a combination of textural parameters may be closely related to underlying pathophysiologic processes. Nevertheless, although several texture features in structural and functional imaging have been shown to characterize diseases or lesions, predict treatment response or be associated with survival, the biologic correlates of texture features are still largely unknown. There is a need to carefully investigate the correlation between texture features from different imaging modalities and histopathologic features that may influence image texture, either in a preclinical model or in humans for each specific application when tissue is available for exhaustive histologic analysis. Based on our results, we only can affirm that the texture features tested in each project were able to capture image heterogeneity differences between the populations analyzed for each specific task and we preferred not to make conclusions about the correlation between our results and the histopathologic analysis of the tissues under examination since we did not have all the information that an analysis of these dimensions would require. The latter should be considered as the next step of radiomics analyses.

Another important concern that arose during the performance of this thesis is related to the implemented machine learning methodology: *why using machine learning approaches instead of trending deep learning techniques?* From 2016 onwards, deep learning methods have dominated the advances in the medical image analysis field since they provide optimal results without the need for much user interaction. However, we preferred to keep using machine learning methods in our analyses mainly for two reasons. In first place, the requirement of still expensive hardware and large amounts of annotated data limited the applicability of deep learning methods in our studies. In second place, deep learning models are seen as "black boxes" where intrinsic quantitative features cannot be extracted, thus limiting further studies regarding the pathologic meaning of the imaging features, as previously discussed. We consider that texture features still have a great deal to offer in the medical imaging field, so we decided to stick to our initial plan in this thesis.

In conclusion, this thesis shows four different applications in which texture analysis applied to MRI is capable of assessing different brain lesions and diseases by means of machine learning implementations. Our preliminary results set the stage for a possible future scenario where texture features could be used as imaging biomarkers for helping in the accurate diagnosis and evaluation of brain diseases like brain tumors, Alzheimer's disease and stroke.

# Chapter 10.
# References


[1] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data.," *Radiology*, vol. 278, no. 2, pp. 563–77, Feb. 2016.

[2] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. P. M. Van Stiphout, P. Granton, C. M. L. Zegers, R. Gillies, *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[3] S. S. F. Yip and H. J. W. L. Aerts, "Applications and limitations of radiomics.," *Phys. Med. Biol.*, vol. 61, no. 13, pp. R150-66, 2016.

[4] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, Oct. 2017.

[5] M. Zhou, J. Scott, B. Chaudhury, L. Hall, D. Goldgof, K. W. Yeom, M. Iv, Y. Ou, *et al.*, "Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches," *Am. J. Neuroradiol.*, vol. 39, no. 2, pp. 208–216, Feb. 2018.

[6] A. Kotrotsou, P. O. Zinn, and R. R. Colen, "Radiomics in Brain Tumors," *Magn. Reson. Imaging Clin. N. Am.*, vol. 24, no. 4, pp. 719–729, Nov. 2016.

[7] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. W. L. Aerts, *et al.*, "Radiomics: the process and the challenges," *Magn. Reson. Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[8] E. J. Limkin, R. Sun, L. Dercle, E. I. Zacharaki, C. Robert, S. Reuzé, A. Schernberg, N. Paragios, *et al.*, "Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology," *Ann. Oncol.*, vol. 28, no. 6, pp. 1191–1206, Jun. 2017.

[9] M. Avanzo, J. Stancanello, and I. El Naqa, "Beyond imaging: The promise of radiomics," *Phys. Medica*, vol. 38, pp. 122–139, Jun. 2017.

[10] R. T. H. M. Larue, G. Defraene, D. De Ruysscher, P. Lambin, and W. van Elmpt, "Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures," *Br. J. Radiol.*, vol. 90, no. 1070, p. 20160665, Feb. 2017.

[11] S. Ranjbar and J. Ross Mitchell, "An Introduction to Radiomics: An Evolving Cornerstone of Precision Medicine," in *Biomedical Texture Analysis: Fundamentals, Tools and Challenges*, 1st ed., A. Depeursinge, O. S. Al-Kadi, and J. Ross Mitchell, Eds. London, UK: Elsevier/Academic Press, 2017, pp. 223–245.

[12] G. Castellano, L. Bonilha, L. M. Li, and F. Cendes, "Texture analysis of medical images.," *Clin. Radiol.*, vol. 59, no. 12, pp. 1061–9, 2004.

[13] A. Larroza, V. Bodí, and D. Moratal, "Texture analysis in magnetic resonance imaging: review and considerations for future applications," in *Assessment of cellular and organ function and dysfunction using direct and derived MRI methodologies*, C. Constantinides, Ed. Rijeka, Croatia: InTech, 2016, pp. 75–106.

[14] A. R. Crossman and R. Tunstall, "Overview of the nervous system," in *Gray's Anatomy: The Anatomical Basis of Clinical Practice*, 41st ed., S. Standring, Ed. London, UK: Elsevier, 2016, pp. 227–237.

[15] N. M. Borden, S. E. Forseen, and C. Stefan, Eds., "Introduction to the Development, Organization, and Function of the Human Brain," in *Imaging Anatomy of the Human Brain: A Comprehensive Atlas Including Adjacent Structures*, New York, NY, USA: Demos Medical, 2016, pp. 1–16.

[16] K. T. Patton and G. A. Thibodeau, Eds., "The nervous system," in *The Human Body in Health and & Disease*, 6th ed., St. Louis, MO, USA: Elsevier/Mosby, 2014, pp. 238–279.

[17] E. N. Marieb and K. Hoehn, Eds., "The Central Nervous System," in *Human Anatomy & Physiology*, 10th ed., Edinburgh, UK: Pearson Education Limited, 2016, pp. 450–504.

[18] G. C. Ribas, "Cerebral hemispheres," in *Gray's Anatomy: The Anatomical Basis of Clinical Practice*, 41st ed., S. Standring, Ed. London, UK: Elsevier, 2016, pp. 373–403.

[19] R. S. Porter and J. L. Kaplan, Eds., *The Merck Manual of Diagnosis and Therapy*, 19th ed. Kenilworth, NJ, USA: Merck Sharp & Dohme Corp., 2011.

[20] D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, *et al.*, "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary," *Acta Neuropathol.*, vol. 131, no. 6, pp. 803–820, Jun. 2016.

[21] Q. T. Ostrom, H. Gittleman, P. Liao, T. Vecchione-Koval, Y. Wolinsky, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS Statistical Report: Primary brain and other

central nervous system tumors diagnosed in the United States in 2010–2014," *Neuro. Oncol.*, vol. 19, no. suppl_5, pp. v1–v88, Nov. 2017.

[22] R. Soffietti, P. Cornu, J. Y. Delattre, R. Grant, F. Graus, W. Grisold, J. Heimans, J. Hildebrand, *et al.*, "EFNS Guidelines on diagnosis and treatment of brain metastases: report of an EFNS Task Force," *Eur. J. Neurol.*, vol. 13, no. 7, pp. 674–681, 2006.

[23] L. Nayak, E. Q. Lee, and P. Y. Wen, "Epidemiology of brain metastases," *Curr. Oncol. Rep.*, vol. 14, no. 1, pp. 48–54, 2012.

[24] Q. T. Ostrom, C. H. Wright, and J. S. Barnholtz-Sloan, "Brain metastases: epidemiology," in *Metastatic Disease of the Nervous System*, Handbook of Clinical Neurology, vol. 149, D. Schiff and M. J. van den Bent, Eds. Elsevier, 2018, pp. 27–42.

[25] M. Wortmann, "Dementia: a global health priority - highlights from an ADI and World Health Organization report.," *Alzheimers. Res. Ther.*, vol. 4, no. 5, p. 40, 2012.

[26] C. Patterson, *World Alzheimer Report 2018: The state of the art of dementia research: New frontiers*. London, UK: Alzheimer's Disease International (ADI), 2018.

[27] Alzheimer's Association, "2017 Alzheimer's disease facts and figures," *Alzheimer's Dement.*, vol. 13, no. 4, pp. 325–373, Apr. 2017.

[28] B. Winblad, P. Amouyel, S. Andrieu, C. Ballard, C. Brayne, H. Brodaty, A. Cedazo-Minguez, B. Dubois, *et al.*, "Defeating Alzheimer's disease and other dementias: a priority for European science and society," *Lancet Neurol.*, vol. 15, no. 5, pp. 455–532, Apr. 2016.

[29] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, *et al.*, "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 280–292, May 2011.

[30] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk, "Brain Imaging in Alzheimer Disease," *Cold Spring Harb. Perspect. Med.*, vol. 2, no. 4, p. a006213, Apr. 2012.

[31] C. R. Jack, D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, P. Vemuri, *et al.*, "Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers," *Lancet Neurol.*, vol. 12, no. 2, pp. 207–216, Feb. 2013.

[32] V. L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K. A. Ellis, O. Salvado, C. Szoeke, S. L. Macaulay, *et al.*, "Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study," *Lancet Neurol.*, vol. 12, no. 4, pp. 357–367, Apr. 2013.

[33] C. M. Miller, "Stroke Epidemiology," in *Translational Research in Stroke*, P. A. Lapchak and G.-Y. Yang, Eds. Singapore: Springer, 2017, pp. 41–49.

[34] A. G. Thrift, T. Thayabaranathan, G. Howard, V. J. Howard, P. M. Rothwell, V. L. Feigin, B. Norrving, G. A. Donnan, *et al.*, "Global stroke statistics," *Int. J. Stroke*, vol. 12, no. 1, pp. 13–32, Jan. 2017.

[35] J. M. Wardlaw, "What causes lacunar stroke?," *J. Neurol. Neurosurg. Psychiatry*, vol. 76, no. 5, pp. 617–619, May 2005.

[36] M. Saini, K. Ikram, S. Hilal, A. Qiu, N. Venketasubramanian, and C. Chen, "Silent Stroke: Not Listened to Rather Than Silent," *Stroke*, vol. 43, no. 11, pp. 3102–3104, Nov. 2012.

[37] Y. Shi and J. M. Wardlaw, "Update on cerebral small vessel disease: a dynamic whole-brain disease," *Stroke Vasc. Neurol.*, vol. 1, no. 3, pp. 83–92, Sep. 2016.

[38] F. Bloch, W. W. Hansen, and M. Packard, "The Nuclear Induction Experiment," *Phys. Rev.*, vol. 70, no. 7–8, pp. 474–485, Oct. 1946.

[39] E. M. Purcell, H. C. Torrey, and R. V. Pound, "Resonance Absorption by Nuclear Magnetic Moments in a Solid," *Phys. Rev.*, vol. 69, no. 1–2, pp. 37–38, Jan. 1946.

[40] R. Damadian, "Tumor Detection by Nuclear Magnetic Resonance," *Science*, vol. 171, no. 3976, pp. 1151–1153, Mar. 1971.

[41] P. C. Lauterbur, "Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance," *Nature*, vol. 242, no. 5394, pp. 190–191, Mar. 1973.

[42] P. Mansfield, "Multi-planar image formation using NMR spin echoes," *J. Phys. C Solid State Phys.*, vol. 10, no. 3, pp. L55–L58, Feb. 1977.

[43] D. Moratal, M. E. Brummer, L. Martí-Bonmatí, and A. Vallés-Lluch, "NMR Imaging," in *Wiley Encyclopedia of Biomedical Engineering*, M. Akay, Ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006, pp. 2590–2606.

[44] S. C. Bushong and G. Clarke, *Magnetic Resonance Imaging: Physical and Biological Principles*, 4th ed. St. Louis, MO, USA: Elsevier/Mosby, 2015.

[45] D. W. McRobbie, E. A. Moore, M. J. Graves, and M. R. Prince, *MRI from Picture to Proton*, 2nd ed. Cambridge, UK: Cambridge University Press, 2006.

[46] M. Symms, H. R. Jäger, K. Schmierer, and T. A. Yousry, "A review of structural magnetic resonance neuroimaging," *J. Neurol. Neurosurg. Psychiatry*, vol. 75, no. 9, pp. 1235–1244, Sep. 2004.

[47] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," *Phys. Med. Biol.*, vol. 58, no. 13, pp. R97–R129, Jul. 2013.

[48] A. Materka, "Texture analysis methodologies for magnetic resonance imaging.," *Dialogues Clin. Neurosci.*, vol. 6, no. 2, pp. 243–250, 2004.

[49] P. M. Szczypinski, M. Strzelecki, A. Materka, and A. Klepaczko, "MaZda - A software package for image texture analysis.," *Comput Methods Programs Biomed*, vol. 94, no. 1, pp. 66–76, 2009.

[50] A. Materka, "Statistical Methods," in *Texture Analysis for Magnetic Resonance Imaging*, 1st ed., M. Hajek, M. Dezortova, A. Materka, and R. Lerski, Eds. Prague, Czech Republic: med4publishing, 2006, pp. 81–106.

[51] A. Depeursinge, A. Foncubierta-Rodriguez, D. Van De Ville, and H. Müller, "Three-dimensional solid texture analysis in biomedical imaging: review and opportunities," *Med. Image Anal.*, vol. 18, no. 1, pp. 176–196, 2014.

[52] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat. Commun.*, vol. 5, p. 4006, Jun. 2014.

[53] E. Orphanidou-Vlachou, N. Vlachos, N. P. Davies, T. N. Arvanitis, R. G. Grundy, and A. C. Peet, "Texture analysis of T1 - and T2 -weighted MR images and use of probabilistic neural network to discriminate posterior fossa tumours in children," *NMR Biomed.*, vol. 27, no. 6, pp. 632–639, Jun. 2014.

[54] Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D. B. Goldgof, *et al.*, "Reproducibility and Prognosis of Quantitative Features Extracted from CT Images," *Transl. Oncol.*, vol. 7, no. 1, pp. 72–87, Feb. 2014.

[55] R. T. H. Leijenaar, S. Carvalho, E. R. Velazquez, W. J. C. van Elmpt, C. Parmar, O. S. Hoekstra, C. J. Hoekstra, R. Boellaard, *et al.*, "Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability," *Acta Oncol. (Madr).*, vol. 52, no. 7, pp. 1391–1397, Oct. 2013.

[56] C. Parmar, E. Rios Velazquez, R. Leijenaar, M. Jermoumi, S. Carvalho, R. H. Mak, S. Mitra, B. U. Shankar, *et al.*, "Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation," *PLoS One*, vol. 9, no. 7, p. e102107, Jul. 2014.

[57] K. K. Holli, L. Harrison, P. Dastidar, M. Wäljas, S. Liimatainen, T. Luukkaala, J. Öhman, S. Soimakallio, *et al.*, "Texture analysis of MR images of patients with Mild Traumatic Brain Injury," *BMC Med. Imaging*, vol. 10, no. 1, p. 8, Dec. 2010.

[58] J. Zhang, C. Yu, G. Jiang, W. Liu, and L. Tong, "3D texture analysis on MRI images of Alzheimer's disease," *Brain Imaging Behav.*, vol. 6, no. 1, pp. 61–69, 2012.

[59] H. Liu, Y. Shao, D. Guo, Y. Zheng, Z. Zhao, and T. Qiu, "Cirrhosis Classification Based on Texture Classification of Random Features," *Comput. Math. Methods Med.*, vol. 2014, no. 536308, pp. 1–8, 2014.

[60] S. A. Waugh, C. A. Purdie, L. B. Jordan, S. Vinnicombe, R. A. Lerski, P. Martin, and A. M. Thompson, "Magnetic resonance imaging texture analysis classification of primary breast cancer," *Eur. Radiol.*, vol. 26, no. 2, pp. 322–330, Feb. 2016.

[61] W. Chen, M. L. Giger, H. Li, U. Bick, and G. M. Newstead, "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images," *Magn. Reson. Med.*, vol. 58, no. 3, pp. 562–571, 2007.

[62] M. Sikiö, L. Harrison, and H. Eskola, "The Effect of Region of Interest Size on Textural Parameters A study with clinical magnetic resonance images and artificial noise images," in *9th International Symposium on Image and Signal Processing and Analysis (ISPA 2015)*, 2015, pp. 149–153.

[63] C. C. Reyes-Aldasoro and A. Bhalerao, "Volumetric Texture Analysis in Biomedical Imaging," in *Biomedical Diagnostics and Clinical Technologies: Applying High-Performance Cluster and Grid Computing*, M. Pereira and M. Freire, Eds. Hershey, PA, USA: IGI Global, 2011, pp. 200–248.

[64] S. J. Savio, L. C. Harrison, T. Luukkaala, T. Heinonen, P. Dastidar, S. Soimakallio, and H. J. Eskola, "Effect of slice thickness on brain magnetic resonance image texture analysis," *Biomed. Eng. Online*, vol. 9, no. 1, p. 60, 2010.

[65] F. Wagner, A. Gryanik, R. Schulz-Wendtland, P. A. Fasching, and T. Wittenberg, "3D Characterization of Texture: Evaluation for the Potential Application in Mammographic Mass Diagnosis," *Biomed. Eng. / Biomed. Tech.*, vol. 57, no. Suppl. 1, pp. 490–493, Jan. 2012.

[66] D. Assefa, H. Keller, C. Ménard, N. Laperriere, R. J. Ferrari, and I. Yeung, "Robust texture features for response monitoring of glioblastoma multiforme on T1-weighted and T2-FLAIR MR images: A preliminary investigation in terms of identification and segmentation," *Med. Phys.*, vol. 37, no. 4, pp. 1722–1736, Mar. 2010.

[67] D. Jirák, M. Dezortová, and M. Hájek, "Phantoms for texture analysis of MR images. Long-term and multi-center study," *Med. Phys.*, vol. 31, no. 3, pp. 616–622, Feb. 2004.

[68] L. R. Schad and L. Arvid, "Influence of the resolution and signal to noise ratio on MR image texture," in *Texture Analysis for Magnetic Resonance Imaging*, M. Hajek, M. Dezortova, A. Materka, and R. Lerski, Eds. Prague, Czech Republic: med4publishing, 2006, pp. 129–150.

[69] J. D. De Certaines, T. Larcher, D. Duda, N. Azzabou, P.-A. Eliat, L. M. Escudero, A. M. Pinheiro, G. Yang, *et al.*, "Application of texture analysis to muscle MRI: 1 – What kind of information should be expected from texture analysis?," *EPJ Nonlinear Biomed. Phys.*, vol. 3, no. 1, p. 3, Mar. 2015.

[70] R. A. Lerski, J. D. de Certaines, D. Duda, W. Klonowski, G. Yang, J. L. Coatrieux, N. Azzabou, and P.-A. Eliat, "Application of texture analysis to muscle MRI: 2 – Technical recommendations," *EPJ Nonlinear Biomed. Phys.*, vol. 3, no. 1, p. 2, Mar.

2015.

[71] M. E. Mayerhoefer, P. Szomolanyi, D. Jirak, A. Materka, and S. Trattnig, "Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study," *Med. Phys.*, vol. 36, no. 4, pp. 1236–1243, Mar. 2009.

[72] G. Collewet, M. Strzelecki, and F. Mariette, "Influence of MRI acquisition protocols and image intensity normalization methods on texture classification," *Magn. Reson. Imaging*, vol. 22, no. 1, pp. 81–91, 2004.

[73] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Phys. Med. Biol.*, vol. 60, pp. 5471–96, 2015.

[74] P. Gibbs and L. W. Turnbull, "Textural analysis of contrast-enhanced MR images of the breast," *Magn. Reson. Med.*, vol. 50, no. 1, pp. 92–98, Jul. 2003.

[75] D. Mahmoud-Ghoneim, M. K. Alkaabi, J. D. De Certaines, and F.-M. Goettsche, "The impact of image dynamic range on texture classification of brain white matter," *BMC Med. Imaging*, vol. 8, no. 8, pp. 1–8, 2008.

[76] A. Ahmed, P. Gibbs, M. Pickles, and L. Turnbull, "Texture analysis in assessment and prediction of chemotherapy response in breast cancer," *J. Magn. Reson. Imaging*, vol. 38, no. 1, pp. 89–101, Jul. 2013.

[77] M. Leite, L. Rittner, S. Appenzeller, H. H. Ruocco, and R. Lotufo, "Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging," *J. Med. Imaging*, vol. 2, no. 1, p. 014002, Feb. 2015.

[78] D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Can. J. Remote Sens.*, vol. 28, no. 1, pp. 45–62, 2002.

[79] X. Xie and M. Mirmehdi, "A Galaxy of Texture Features," in *Handbook of Texture Analysis*, M. Mirmehdi, X. Xie, and J. Suri, Eds. London, UK: Imperial College Press, 2008, pp. 375–406.

[80] W. H. Nailon, "Texture Analysis Methods for Medical Image Characterisation," in *Biomedical Imaging*, Y. Mao, Ed. Rijeka, Croatia: InTech, 2010, pp. 75–100.

[81] A. Depeursinge and J. Fageot, "Biomedical Texture Operators and Aggregation Functions," in *Biomedical Texture Analysis: Fundamentals, Tools and Challenges*, 1st ed., A. Depeursinge, O. S. Al-Kadi, and J. Ross Mitchell, Eds. London, UK: Elsevier/Academic Press, 2017, pp. 55–94.

[82] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, and  for the I. B. S. Initiative, "Image biomarker standardisation initiative," Dec. 2016.

[83] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation

invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[84] L. Zhang, D. V. Fried, X. J. Fave, L. A. Hunter, J. Yang, and L. E. Court, "IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics," *Med. Phys.*, vol. 42, no. 3, pp. 1341–1353, Feb. 2015.

[85] F. Davnall, C. S. P. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K. A. Miles, *et al.*, "Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?," *Insights Imaging*, vol. 3, no. 6, pp. 573–589, Dec. 2012.

[86] A. Materka, "What is the texture?," in *Texture Analysis for Magnetic Resonance Imaging*, M. Hajek, M. Dezortova, A. Materka, and R. Lerski, Eds. Prague, Czech Republic: med4publishing, 2006, pp. 11–44.

[87] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[88] M. M. Galloway, "Texture analysis using gray level run lengths," *Comput. Graph. Image Process.*, vol. 4, no. 2, pp. 172–179, Jun. 1975.

[89] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern Recognit. Lett.*, vol. 11, no. 6, pp. 415–419, Jun. 1990.

[90] B. V. Dasarathy and E. B. Holder, "Image characterizations based on joint gray level—run length distributions," *Pattern Recognit. Lett.*, vol. 12, no. 8, pp. 497–502, Aug. 1991.

[91] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari, "Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification," in *Pattern Recognition and Information Processing*, 2009, pp. 140–145.

[92] M. Amadasun and R. King, "Textural features corresponding to textural properties," *IEEE Trans. Syst. Man. Cybern.*, vol. 19, no. 5, pp. 1264–1274, 1989.

[93] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[94] D. Unay, A. Ekin, M. Cetin, R. Jasinschi, and A. Ercil, "Robustness of Local Binary Patterns in Brain MR Image Analysis," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 2098–2101.

[95] S. A. Waugh, R. A. Lerski, L. Bidaut, and A. M. Thompson, "The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms," *Med. Phys.*, vol. 38, no. 9, pp. 5058–5066, Aug.

2011.

[96] S. Arivazhagan and L. Ganesan, "Texture classification using wavelet transform," *Pattern Recognit. Lett.*, vol. 24, no. 9–10, pp. 1513–1521, Jun. 2003.

[97] E. Alegre, V. González-Castro, R. Alaiz-Rodríguez, and M. T. García-Ordás, "Texture and moments-based classification of the acrosome integrity of boar spermatozoa images," *Comput. Methods Programs Biomed.*, vol. 108, no. 2, pp. 873–881, Nov. 2012.

[98] V. González-Castro, M. del C. Valdés-Hernández, F. M. Chappell, P. A. Armitage, S. Makin, and J. M. Wardlaw, "Reliability of an automatic classifier for brain enlarged perivascular spaces burden and comparison with human performance," *Clin. Sci.*, vol. 131, no. 13, pp. 1465–1481, Jul. 2017.

[99] D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, *et al.*, "Measuring Computed Tomography Scanner Variability of Radiomics Features," *Invest. Radiol.*, vol. 50, no. 11, pp. 757–765, Nov. 2015.

[100] B. Zhao, Y. Tan, W.-Y. Tsai, J. Qi, C. Xie, L. Lu, and L. H. Schwartz, "Reproducibility of radiomics for deciphering tumor phenotype with imaging," *Sci. Rep.*, vol. 6, no. 1, p. 23428, Sep. 2016.

[101] I. Havsteen, A. Ohlhues, K. H. Madsen, J. D. Nybing, H. Christensen, and A. Christensen, "Are Movement Artifacts in Magnetic Resonance Imaging a Real Problem? - A Narrative Review," *Front. Neurol.*, vol. 8, p. 232, 2017.

[102] D. Mahmoud-Ghoneim, G. Toussaint, J. M. Constans, and J. D. De Certaines, "Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas," *Magn. Reson. Imaging*, vol. 21, no. 9, pp. 983–987, 2003.

[103] A. E. Fetit, J. Novak, A. C. Peet, and T. N. Arvanitis, "Three-dimensional textural features of conventional MRI improve diagnostic classification of childhood brain tumours," *NMR Biomed.*, vol. 28, no. 9, pp. 1174–1184, 2015.

[104] E. I. Zacharaki, S. Wang, S. Chawla, D. S. Yoo, R. Wolf, E. R. Melhem, and C. Davatzikos, "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," *Magn. Reson. Med.*, vol. 62, no. 6, pp. 1609–1618, 2009.

[105] P. Georgiadis, D. Cavouras, I. Kalatzis, D. Glotsos, E. Athanasiadis, S. Kostopoulos, K. Sifaki, M. Malamas, *et al.*, "Enhancing the discrimination accuracy between metastases, gliomas and meningiomas on brain MRI by volumetric textural features and ensemble pattern recognition methods," *Magn. Reson. Imaging*, vol. 27, no. 1, pp. 120–130, 2009.

[106] A. Larroza, D. Moratal, A. Paredes-Sánchez, E. Soria-Olivas, M. L. Chust, L. A. Arribas, and E. Arana, "Support vector machine classification of brain metastasis and radiation necrosis based on texture analysis in MRI," *J. Magn. Reson. Imaging*, vol.

42, no. 5, pp. 1362–8, 2015.

[107] Z. Li, Y. Mao, H. Li, G. Yu, H. Wan, and B. Li, "Differentiating brain metastases from different pathological types of lung cancers using texture analysis of T1 postcontrast MR," *Magn. Reson. Med.*, vol. 76, no. 5, pp. 1410–1419, Nov. 2016.

[108] L. R. Schad, "Problems in texture analysis with magnetic resonance imaging.," *Dialogues Clin. Neurosci.*, vol. 6, no. 2, pp. 235–42, Jun. 2004.

[109] M. Ganzetti, N. Wenderoth, and D. Mantini, "Intensity Inhomogeneity Correction of Structural MR Images: A Data-Driven Approach to Define Input Algorithm Parameters," *Front. Neuroinform.*, vol. 10, p. 10, 2016.

[110] A. Materka and M. Strzelecki, "On the Importance of MRI Nonuniformity Correction for Texture Analysis," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2013*, 2013, pp. 118–123.

[111] M. E. Mayerhoefer, M. J. Breitenseher, J. Kramer, N. Aigner, S. Hofmann, and A. Materka, "Texture analysis for tissue discrimination on T1-weighted MR images of the knee joint in a multicenter study: Transferability of texture features and comparison of feature selection methods and classifiers," *J. Magn. Reson. Imaging*, vol. 22, no. 5, pp. 674–680, Nov. 2005.

[112] A. Jethanandani, T. A. Lin, S. Volpe, H. Elhalawani, A. S. R. Mohamed, P. Yang, and C. D. Fuller, "Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review," *Front. Oncol.*, vol. 8, p. 131, 2018.

[113] J. P. B. O'Connor, C. J. Rose, J. C. Waterton, R. A. D. Carano, G. J. M. Parker, and A. Jackson, "Imaging Intratumor Heterogeneity: Role in Therapy Response, Resistance, and Clinical Outcome," *Clin. Cancer Res.*, vol. 21, no. 2, pp. 249–257, Jan. 2015.

[114] P. Kuess, P. Andrzejewski, D. Nilsson, P. Georg, J. Knoth, M. Susani, J. Trygg, T. H. Helbich, *et al.*, "Association between pathology and texture features of multi parametric MRI of the prostate," *Phys. Med. Biol.*, vol. 62, no. 19, pp. 7833–7854, Sep. 2017.

[115] H.-J. Meyer, S. Schob, A. K. Höhn, and A. Surov, "MRI Texture Analysis Reflects Histopathology Parameters in Thyroid Cancer – A First Preliminary Study," *Transl. Oncol.*, vol. 10, no. 6, pp. 911–916, Dec. 2017.

[116] A. Kassner and R. E. Thornhill, "Texture analysis: a review of neurologic MR imaging applications," *AJNR. Am. J. Neuroradiol.*, vol. 31, no. 5, pp. 809–816, 2010.

[117] K. Skogen, A. Schulz, J. B. Dormagen, B. Ganeshan, E. Helseth, and A. Server, "Diagnostic performance of texture analysis on MRI in grading cerebral gliomas," *Eur. J. Radiol.*, vol. 85, no. 4, pp. 824–829, Apr. 2016.

[118] K. L.-C. Hsieh, C.-M. Lo, and C.-J. Hsiao, "Computer-aided grading of gliomas

based on local and global MRI features," *Comput. Methods Programs Biomed.*, vol. 139, pp. 31–38, Feb. 2017.

[119] M. S. de Oliveira, M. L. F. Balthazar, A. D'Abreu, C. L. Yasuda, B. P. Damasceno, F. Cendes, and G. Castellano, "MR Imaging Texture Analysis of the Corpus Callosum and Thalamus in Amnestic Mild Cognitive Impairment and Mild Alzheimer Disease," *Am. J. Neuroradiol.*, vol. 32, no. 1, pp. 60–66, Jan. 2011.

[120] J. Zhang, L. Tong, L. Wang, and N. Li, "Texture analysis of multiple sclerosis: a comparative study," *Magn. Reson. Imaging*, vol. 26, no. 8, pp. 1160–1166, Oct. 2008.

[121] L. C. V. Harrison, M. Raunio, K. K. Holli, T. Luukkaala, S. Savio, I. Elovaara, S. Soimakallio, H. J. Eskola, *et al.*, "MRI Texture Analysis in Multiple Sclerosis: Toward a Clinical Analysis Protocol," *Acad. Radiol.*, vol. 17, no. 6, pp. 696–707, Jun. 2010.

[122] A. Kassner, F. Liu, R. E. Thornhill, G. Tomlinson, and D. J. Mikulis, "Prediction of hemorrhagic transformation in acute ischemic stroke using texture analysis of postcontrast T1-weighted MR images," *J. Magn. Reson. Imaging*, vol. 30, no. 5, pp. 933–941, Nov. 2009.

[123] M. del C. Valdés-Hernández, V. González-Castro, F. M. Chappell, E. Sakka, S. Makin, P. A. Armitage, W. H. Nailon, and J. M. Wardlaw, "Application of Texture Analysis to Study Small Vessel Disease and Blood–Brain Barrier Integrity," *Front. Neurol.*, vol. 8, p. 327, Jul. 2017.

[124] A. Larroza, A. Materka, M. P. López-Lereu, J. V. Monmeneu, V. Bodí, and D. Moratal, "Differentiation between acute and chronic myocardial infarction by means of texture analysis of late gadolinium enhancement and cine cardiac magnetic resonance imaging," *Eur. J. Radiol.*, vol. 92, pp. 78–83, Jul. 2017.

[125] L. P. Kotu, K. Engan, R. Borhani, A. K. Katsaggelos, S. Ørn, L. Woie, and T. Eftestøl, "Cardiac magnetic resonance image-based classification of the risk of arrhythmias in post-myocardial infarction patients," *Artif. Intell. Med.*, vol. 64, no. 3, pp. 205–215, Jul. 2015.

[126] X. Sun, B. He, X. Luo, Y. Li, J. Cao, J. Wang, J. Dong, X. Sun, *et al.*, "Preliminary Study on Molecular Subtypes of Breast Cancer Based on Magnetic Resonance Imaging Texture Analysis," *J. Comput. Assist. Tomogr.*, vol. 42, no. 1, pp. 531–535, Apr. 2018.

[127] A. Wibmer, H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr, J. Zheng, D. Goldman, *et al.*, "Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores," *Eur. Radiol.*, vol. 25, no. 10, pp. 2840–2850, Oct. 2015.

[128] X. Niu, Z. Chen, L. Chen, J. Li, T. Peng, and X. Li, "Clinical Application of Biparametric MRI Texture Analysis for Detection and Evaluation of High-Grade

Prostate Cancer in Zone-Specific Regions," *Am. J. Roentgenol.*, vol. 210, no. 3, pp. 549–556, Mar. 2018.

[129] T. L. Kline, P. Korfiatis, M. E. Edwards, K. T. Bae, A. Yu, A. B. Chapman, M. Mrug, J. J. Grantham, *et al.*, "Image texture features predict renal function decline in patients with autosomal dominant polycystic kidney disease," *Kidney Int.*, vol. 92, no. 5, pp. 1206–1216, Nov. 2017.

[130] T. Yokoo, T. Wolfson, K. Iwaisako, M. R. Peterson, H. Mani, Z. Goodman, C. Changchien, M. S. Middleton, *et al.*, "Evaluation of Liver Fibrosis Using Texture Analysis on Combined-Contrast-Enhanced Magnetic Resonance Images at 3.0T," *Biomed Res. Int.*, vol. 2015, no. 387653, pp. 1–12, 2015.

[131] J. W. MacKay, P. J. Murray, B. Kasmai, G. Johnson, S. T. Donell, and A. P. Toms, "MRI texture analysis of subchondral bone at the tibial plateau," *Eur. Radiol.*, vol. 26, no. 9, pp. 3034–3045, Sep. 2016.

[132] T. K. Chuah, E. Van Reeth, K. Sheah, and C. L. Poh, "Texture analysis of bone marrow in knee MRI for classification of subjects with bone marrow lesion — Data from the Osteoarthritis Initiative," *Magn. Reson. Imaging*, vol. 31, no. 6, pp. 930–938, Jul. 2013.

[133] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015.

[134] L. Deng and X. Li, "Machine Learning Paradigms for Speech Recognition: An Overview," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 21, no. 5, pp. 1060–1089, May 2013.

[135] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Q.*, vol. 36, no. 4, p. 1165, 2012.

[136] J. Bagnell, D. Bradley, D. Silver, B. Sofman, and A. Stentz, "Learning for Autonomous Navigation," *IEEE Robot. Autom. Mag.*, vol. 17, no. 2, pp. 74–84, Jun. 2010.

[137] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 223–230, Jan. 2011.

[138] M. Wernick, Y. Yang, J. Brankov, G. Yourganov, and S. Strother, "Machine Learning in Medical Imaging," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 25–38, Jul. 2010.

[139] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine Learning for Medical Imaging," *RadioGraphics*, vol. 37, no. 2, pp. 505–515, Mar. 2017.

[140] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer Series in Statistics, 2009.

[141] S. Wang and R. M. Summers, "Machine learning and radiology," *Med. Image Anal.*,

vol. 16, no. 5, pp. 933–951, 2012.

[142] A. Palmer, R. Jimenez, and E. Gervill, "Data Mining: Machine Learning and Statistical Techniques," in *Knowledge-Oriented Applications in Data Mining*, K. Funatsu, Ed. Rijeka, Croatia: InTech, 2011, pp. 373–396.

[143] M. Kuhn, "Building predictive models in R using the caret package," *J. Stat. Softw.*, vol. 28, no. 5, pp. 1–26, 2008.

[144] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, and D. Amorim Fernández-Delgado, "Do we need hundreds of classifiers to solve real world classification problems?," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.

[145] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. da Fontoura Costa, "A Systematic Comparison of Supervised Classifiers," *PLoS One*, vol. 9, no. 4, p. e94137, Apr. 2014.

[146] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning - ICML '08*, 2008, pp. 96–103.

[147] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 11, pp. 1457–1466, Nov. 2006.

[148] T. A. Almeida, J. Almeida, and A. Yamakami, "Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers," *J. Internet Serv. Appl.*, vol. 1, no. 3, pp. 183–200, Feb. 2011.

[149] K. M. Al-Aidaroo, A. A. Bakar, and Z. Othman, "Medical Data Classification with Naive Bayes Approach," *Inf. Technol. J.*, vol. 11, no. 9, pp. 1166–1174, Sep. 2012.

[150] M. Kuhn and K. Johnson, "Nonlinear Classification Models," in *Applied predictive modeling*, 1st ed., New York, NY, USA: Springer-Verlag, 2013, pp. 329–367.

[151] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008.

[152] W. S. Noble, "Support vector machine applications in computational biology," in *Kernel Methods in Computational Biology*, 1st ed., B. Schoelkopf, K. Tsuda, and J. P. Vert, Eds. Cambridge, MA, USA: MIT Press, 2004, pp. 71–92.

[153] A. Criminisi, "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning," *Found. Trends® Comput. Graph. Vis.*, vol. 7, no. 2–3, pp. 81–227, 2012.

[154] F. Schroff, A. Criminisi, and A. Zisserman, "Object Class Segmentation using Random Forests," in *Procedings of the British Machine Vision Conference 2008*, 2008, p. 54.1-54.10.

[155] M. Kuhn and K. Johnson, “Classification Trees and Rule-Based Models,” in *Applied predictive modeling*, 1st ed., New York, NY, USA: Springer-Verlag, 2013, pp. 369–413.

[156] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[157] P. Azimi, H. R. Mohammadi, E. C. Benzel, S. Shahzadi, S. Azhari, and A. Montazeri, “Artificial neural networks in neurosurgery,” *J. Neurol. Neurosurg. Psychiatry*, vol. 86, no. 3, pp. 251–256, Mar. 2015.

[158] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, “Artificial neural networks in medical diagnosis,” *J. Appl. Biomed.*, vol. 11, no. 2, pp. 47–58, 2013.

[159] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: a review of classification and combining techniques,” *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, 2006.

[160] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, 1st ed. Cambridge, UK: Cambridge University Press, 2012.

[161] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[162] M. Kuhn and K. Johnson, “An introduction to feature selection,” in *Applied predictive modeling*, 1st ed., New York, NY, USA: Springer-Verlag, 2013, pp. 487–519.

[163] Z. M. Hira and D. F. Gillies, “A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data,” *Adv. Bioinformatics*, vol. 2015, no. ID198363, pp. 1–13, 2015.

[164] Y. Saeys, I. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.

[165] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, Jan. 2014.

[166] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, *et al.*, “Detecting Novel Associations in Large Data Sets,” *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.

[167] R. Ge, M. Zhou, Y. Luo, Q. Meng, G. Mai, D. Ma, G. Wang, and F. Zhou, “McTwo: a two-step feature selection algorithm based on maximal information coefficient,” *BMC Bioinformatics*, vol. 17, no. 1, p. 142, Dec. 2016.

[168] H. Lyu, M. Wan, J. Han, R. Liu, and C. Wang, “A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining,” *Comput. Biol. Med.*, vol. 89, pp. 264–274, Oct. 2017.

[169] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-

based feature selection: Introduction and review," *J. Biomed. Inform.*, vol. 85, pp. 189–203, Sep. 2018.

[170] K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," in *Machine Learning: Proceedings of the Ninth International Workshop (ML92)*, 1992, pp. 249–256.

[171] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *European Conference on Machine Learning (ECML-94)*, 1994, pp. 171–182.

[172] A. Hapfelmeier and K. Ulm, "A new variable selection approach using Random Forests," *Comput. Stat. Data Anal.*, vol. 60, pp. 50–69, Apr. 2013.

[173] M. Kuhn and K. Johnson, "Over-fitting and model tuning," in *Applied predictive modeling*, 1st ed., New York, NY, USA: Springer-Verlag, 2013, pp. 61–92.

[174] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA, USA: Springer US, 2009, pp. 532–538.

[175] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, Aug. 2005.

[176] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 10, pp. 6562–6566, 2002.

[177] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," *Chemom. Intell. Lab. Syst.*, vol. 56, no. 1, pp. 1–11, Apr. 2001.

[178] M. Kuhn and K. Johnson, "Measuring performance in classification models," in *Applied predictive modeling*, 1st ed., New York, NY, USA: Springer-Verlag, 2013, pp. 247–273.

[179] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[180] Q. T. Ostrom, H. Gittleman, L. Stetson, S. M. Virk, and J. S. Barnholtz-Sloan, "Epidemiology of Gliomas," in *Current Understanding and Treatment of Gliomas*, Cancer Treatment and Research, vol. 163, J. Raizer and A. Parsa, Eds. Springer, Cham, 2015, pp. 1–14.

[181] V. A. Venur, D. M. Peereboom, and M. S. Ahluwalia, "Current Medical Treatment of Glioblastoma," in *Current Understanding and Treatment of Gliomas*, Cancer Treatment and Research, vol. 163, J. Raizer and A. Parsa, Eds. Springer, Cham, 2015, pp. 103–115.

[182] E. C. A. Kaal, M. J. B. Taphoorn, and C. J. Vecht, "Symptomatic management and imaging of brain metastases," *J. Neurooncol.*, vol. 75, no. 1, pp. 15–20, 2005.

[183] B. D. Fox, V. J. Cheung, A. J. Patel, D. Suki, and G. Rao, "Epidemiology of metastatic brain tumors," *Neurosurg. Clin. N. Am.*, vol. 22, no. 1, pp. 1–6, 2011.

[184] K. J. Stelzer, "Epidemiology and prognosis of brain metastases.," *Surg. Neurol. Int.*, vol. 4, no. Suppl 4, pp. S192--202, 2013.

[185] I. T. Gavrilovic and J. B. Posner, "Brain metastases: epidemiology and pathophysiology," *J. Neurooncol.*, vol. 75, no. 1, pp. 5–14, 2005.

[186] S. Campos, P. Davey, A. Hird, B. Pressnail, J. Bilbao, R. I. Aviv, S. Symons, F. Pirouzmand, *et al.*, "Brain metastasis from an unknown primary, or primary brain tumour? A diagnostic dilemma.," *Curr. Oncol.*, vol. 16, no. 1, pp. 62–6, Jan. 2009.

[187] S. K. Carlsson, S. P. Brothers, and C. Wahlestedt, "Emerging treatment strategies for glioblastoma multiforme," *EMBO Mol. Med.*, vol. 6, no. 11, pp. 1359–1370, Nov. 2014.

[188] M. Davis, "Glioblastoma: Overview of Disease and Treatment," *Clin. J. Oncol. Nurs.*, vol. 20, no. 5, pp. S2–S8, Oct. 2016.

[189] D. A. Hardesty and P. Nakaji, "The Current and Future Treatment of Brain Metastases," *Front. Surg.*, vol. 3, p. 30, May 2016.

[190] T. K. Owonikoko, J. Arbiser, A. Zelnak, H.-K. G. Shu, H. Shim, A. M. Robin, S. N. Kalkanis, T. G. Whitsett, *et al.*, "Current approaches to the treatment of metastatic brain tumours," *Nat. Rev. Clin. Oncol.*, vol. 11, no. 4, pp. 203–222, Apr. 2014.

[191] L. Blanchet, P. W. T. Krooshof, G. J. Postma, A. J. Idema, B. Goraj, A. Heerschap, and L. M. C. Buydens, "Discrimination between metastasis and glioblastoma multiforme based on morphometric analysis of MR images.," *AJNR. Am. J. Neuroradiol.*, vol. 32, no. 1, pp. 67–73, Jan. 2011.

[192] I. Tsougos, P. Svolos, E. Kousi, K. Fountas, K. Theodorou, I. Fezoulidis, and E. Kapsalaki, "Differentiation of glioblastoma multiforme from metastatic brain tumor using proton magnetic resonance spectroscopy, diffusion and perfusion metrics at 3 T," *Cancer Imaging*, vol. 12, no. 3, pp. 423–436, 2012.

[193] S. Cha, J. M. Lupo, M.-H. Chen, K. R. Lamborn, M. W. McDermott, M. S. Berger, S. J. Nelson, and W. P. Dillon, "Differentiation of glioblastoma multiforme and single brain metastasis by peak height and percentage of signal intensity recovery derived from dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging.," *AJNR. Am. J. Neuroradiol.*, vol. 28, no. 6, pp. 1078–84, Jun. 2007.

[194] B. C. Jung, J. Arevalo-Perez, J. K. Lyo, A. I. Holodny, S. Karimi, R. J. Young, and K. K. Peck, "Comparison of Glioblastomas and Brain Metastases using Dynamic Contrast-Enhanced Perfusion MRI," *J. Neuroimaging*, vol. 26, no. 2, pp. 240–246, Mar. 2016.

[195] J. G. Smirniotopoulos, F. M. Murphy, E. J. Rushing, J. H. Rees, and J. W. Schroeder, "Patterns of Contrast Enhancement in the Brain and Meninges," *RadioGraphics*, vol.

27, no. 2, pp. 525–551, Mar. 2007.

[196] W. B. Pope, “Brain metastases: neuroimaging,” in *Metastatic Disease of the Nervous System*, Handbook of Clinical Neurology, vol. 149, D. Schiff and M. J. van den Bent, Eds. Elsevier, 2018, pp. 89–112.

[197] W. B. Pope, I. Djoukhadar, and A. Jackson, “Neuroimaging,” in *Gliomas*, Handbook of Clinical Neurology, vol. 134, M. S. Berger and M. Weller, Eds. Elsevier, 2016, pp. 27–50.

[198] J. R. Fink, M. Muzi, M. Peck, and K. A. Krohn, “Multimodality Brain Tumor Imaging: MR Imaging, PET, and PET/MR Imaging,” *J. Nucl. Med.*, vol. 56, no. 10, pp. 1554–1561, Oct. 2015.

[199] B. Hakyemez, C. Erdogan, G. Gokalp, A. Dusak, and M. Parlak, “Solitary metastases and high-grade gliomas: radiological differentiation by morphometric analysis and perfusion-weighted MRI,” *Clin. Radiol.*, vol. 65, no. 1, pp. 15–20, Jan. 2010.

[200] T. J. D. Byrnes, T. R. Barrick, B. A. Bell, and C. A. Clark, “Diffusion tensor imaging discriminates between glioblastoma and cerebral metastases in vivo,” *NMR Biomed.*, vol. 24, no. 1, pp. 54–60, Jan. 2011.

[201] N. Mouthuy, G. Cosnard, J. Abarca-Quinones, and N. Michoux, “Multiparametric magnetic resonance imaging to differentiate high-grade gliomas and brain metastases,” *J. Neuroradiol.*, vol. 39, no. 5, pp. 301–307, Dec. 2012.

[202] P. Svolos, E. Kousi, E. Kapsalaki, K. Theodorou, I. Fezoulidis, C. Kappas, and I. Tsougos, “The role of diffusion and perfusion weighted imaging in the differential diagnosis of cerebral tumors: a review and future perspectives.,” *Cancer Imaging*, vol. 14, p. 20, Apr. 2014.

[203] D. O. Kamson, S. Mittal, A. Buth, O. Muzik, W. J. Kupsky, N. L. Robinette, G. R. Barger, and C. Juhász, “Differentiation of glioblastomas from metastatic brain tumors by tryptophan uptake and kinetic analysis: a positron emission tomographic study with magnetic resonance imaging comparison.,” *Mol. Imaging*, vol. 12, no. 5, pp. 327–37, 2013.

[204] J. Sachdeva, V. Kumar, I. Gupta, N. Khandelwal, and C. K. Ahuja, “A package-SFERCB-‘Segmentation, feature extraction, reduction and classification analysis by both SVM and ANN for brain tumors,’” *Appl. Soft Comput.*, vol. 47, pp. 151–167, Oct. 2016.

[205] M. Kuhn and K. Johnson, “Data pre-processing,” in *Applied predictive modeling*, 1st ed., New York, NY, USA: Springer-Verlag, 2013, pp. 27–59.

[206] M. Gupta, V. Rajagopalan, E. P. Pioro, and B. V. V. S. N. P. Rao, “Volumetric analysis of MR images for glioma classification and their effect on brain tissues,” *Signal, Image Video Process.*, vol. 11, no. 7, pp. 1337–1345, Oct. 2017.

[207] M. Soltaninejad, X. Ye, G. Yang, N. Allinson, and T. Lambrou, “An image analysis

approach to MRI brain tumour grading," *Oncol. News*, vol. 9, no. 6, pp. 204–207, 2015.

[208] P. Tiwari, J. Sachdeva, C. K. Ahuja, and N. Khandelwal, "Computer Aided Diagnosis System - A Decision Support System for Clinical Diagnosis of Brain Tumours," *Int. J. Comput. Intell. Syst.*, vol. 10, no. 1, pp. 104–119, 2017.

[209] Y. Gao, Z. Shi, Y. Wang, J. Yu, L. Chen, Y. Guo, Q. Zhang, and Y. Mao, "Histological grade and type classification of glioma using Magnetic Resonance Imaging," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2016, pp. 1808–1813.

[210] M. P. Arakeri and G. R. M. Reddy, "Computer-aided diagnosis system for tissue characterization of brain tumor on magnetic resonance images," *Signal, Image Video Process.*, vol. 9, no. 2, pp. 409–425, Feb. 2015.

[211] M. Sasikala and N. Kumaravel, "A wavelet-based optimal texture feature set for classification of brain tumours," *J. Med. Eng. Technol.*, vol. 32, no. 3, pp. 198–205, Jan. 2008.

[212] R. Rudà, M. Borgognone, F. Benech, E. Vasario, and R. Soffietti, "Brain metastases from unknown primary tumour: a prospective study.," *J. Neurol.*, vol. 248, no. 5, pp. 394–8, 2001.

[213] M. Pekmezci and A. Perry, "Neuropathology of brain metastases," *Surg. Neurol. Int.*, vol. 4, no. 5, p. 245, 2013.

[214] L. Bekaert, E. Emery, G. Levallet, and E. Lechapt-Zalcman, "Histopathologic diagnosis of brain metastases: current trends in management and future considerations," *Brain Tumor Pathol.*, vol. 34, no. 1, pp. 8–19, Jan. 2017.

[215] S. Bartelt and J. Lutterbach, "Brain metastases in patients with cancer of unknown primary.," *J. Neurooncol.*, vol. 64, no. 3, pp. 249–53, Sep. 2003.

[216] S. Maesawa, D. Kondziolka, T. P. Thompson, J. C. Flickinger, and L. Dade Lunsford, "Brain metastases in patients with no known primary tumor: The role of stereotactic radiosurgery," *Cancer*, vol. 89, no. 5, pp. 1095–1101, 2000.

[217] S. Agazzi, S. Pampallona, A. Pica, O. Vernet, L. Regli, F. Porchet, J. G. Villemure, S. Leyvraz, *et al.*, "The origin of brain metastases in patients with an undiagnosed primary tumour," *Acta Neurochir. (Wien).*, vol. 146, no. 2, pp. 153–157, 2004.

[218] B. Klee, I. Law, L. Højgaard, and M. Kosteljanetz, "Detection of unknown primary tumours in patients with cerebral metastases using whole-body 18F-flouorodeoxyglucose positron emission tomography.," *Eur. J. Neurol.*, vol. 9, no. 6, pp. 657–62, Nov. 2002.

[219] G. Duygulu, G. Y. Ovali, C. Çalli, Ö. Kitis, N. Yünten, T. Akalin, and S. Islekel, "Intracerebral metastasis showing restricted diffusion: Correlation with histopathologic findings," *Eur. J. Radiol.*, vol. 74, no. 1, pp. 117–120, Apr. 2010.

[220] Y. Hayashida, T. Hirai, S. Morishita, M. Kitajima, R. Murakami, Y. Korogi, K. Makino, H. Nakamura, *et al.*, "Diffusion-weighted imaging of metastatic brain tumors: comparison with histologic type and tumor cellularity.," *AJNR. Am. J. Neuroradiol.*, vol. 27, no. 7, pp. 1419–25, Aug. 2006.

[221] B. Y. Huang, L. Kwock, M. Castillo, and J. K. Smith, "Association of Choline Levels and Tumor Perfusion in Brain Metastases Assessed with Proton MR Spectroscopy and Dynamic Susceptibility Contrast-enhanced Perfusion Weighted MRI," *Technol. Cancer Res. Treat.*, vol. 9, no. 4, pp. 327–337, Aug. 2010.

[222] R. Zakaria, K. Das, M. Bhojak, M. Radon, C. Walker, and M. D. Jenkinson, "The role of magnetic resonance imaging in the management of brain metastases: diagnosis to prognosis," *Cancer Imaging*, vol. 14, no. 1, pp. 1–8, 2014.

[223] C. C. Quattrocchi, Y. Errante, C. Gaudino, C. A. Mallio, A. Giona, D. Santini, G. Tonini, and B. B. Zobel, "Spatial brain distribution of intra-axial metastatic lesions in breast and lung cancer patients," *J. Neurooncol.*, vol. 110, no. 1, pp. 79–87, Oct. 2012.

[224] W. Feng, P. Zhang, X. Zheng, G. Shan, M. Chen, and W. Mao, "Neuroimaging and clinical characteristics of brain metastases from esophageal carcinoma in Chinese patients," *J. Cancer Res. Ther.*, vol. 10, no. 8, p. 296, 2014.

[225] R.-H. Yeh, J.-C. Yu, C.-H. Chu, C.-L. Ho, H.-W. Kao, G.-S. Liao, H.-W. Chen, W.-Y. Kao, *et al.*, "Distinct MR Imaging Features of Triple-Negative Breast Cancer with Brain Metastasis," *J. Neuroimaging*, vol. 25, no. 3, pp. 474–481, May 2015.

[226] K. R. Fink and J. R. Fink, "Imaging of brain metastases.," *Surg. Neurol. Int.*, vol. 4, no. Suppl 4, pp. S209-19, 2013.

[227] B. M. Ellingson, M. Bendszus, J. Boxerman, D. Barboriak, B. J. Erickson, M. Smits, S. J. Nelson, E. Gerstner, *et al.*, "Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials," *Neuro. Oncol.*, vol. 17, no. 9, pp. 1188–1198, Aug. 2015.

[228] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, 2001.

[229] F. Provost and P. Domingos, "Tree induction for probability-based ranking," *Mach. Learn.*, vol. 52, no. 3, pp. 199–215, 2003.

[230] M. Béresová, A. Larroza, E. Arana, J. Varga, L. Balkay, and D. Moratal, "2D and 3D texture analysis to differentiate brain metastases on MR images: proceed with caution," *Magn Reson Mater Phy*, Springer Berlin Heidelberg, pp. 1–10, Sep-2017.

[231] M. Prince, A. Comas-Herrera, M. Knapp, and M. Guerchet, *World Alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future*. London, UK: Alzheimer's Disease International (ADI), 2016.

[232] R. C. Petersen, "Mild cognitive impairment as a diagnostic entity," *J. Intern. Med.*, vol. 256, no. 3, pp. 183–194, Sep. 2004.

[233] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 270–279, May 2011.

[234] K. E. Pike, G. Savage, V. L. Villemagne, S. Ng, S. A. Moss, P. Maruff, C. A. Mathis, W. E. Klunk, *et al.*, "β-amyloid imaging and memory in non-demented individuals: evidence for preclinical Alzheimer's disease," *Brain*, vol. 130, no. 11, pp. 2837–2844, Apr. 2007.

[235] E. M. Reiman, Y. T. Quiroz, A. S. Fleisher, K. Chen, C. Velez-Pardo, M. Jimenez-Del-Rio, A. M. Fagan, A. R. Shah, *et al.*, "Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: a case-control study," *Lancet Neurol.*, vol. 11, no. 12, pp. 1048–1056, Dec. 2012.

[236] J. R. Petrella, R. E. Coleman, and P. M. Doraiswamy, "Neuroimaging and Early Diagnosis of Alzheimer Disease: A Look to the Future," *Radiology*, vol. 226, no. 2, pp. 315–336, Feb. 2003.

[237] C. R. Jack, M. M. Shiung, J. L. Gunter, P. C. O'Brien, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. J. Ivnik, *et al.*, "Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD," *Neurology*, vol. 62, no. 4, pp. 591–600, Feb. 2004.

[238] R. I. Scahill, J. M. Schott, J. M. Stevens, M. N. Rossor, and N. C. Fox, "Mapping the evolution of regional atrophy in Alzheimer's disease: Unbiased analysis of fluid-registered serial MRI," *Proc. Natl. Acad. Sci.*, vol. 99, no. 7, pp. 4703–4707, Apr. 2002.

[239] L.-O. Wahlund, O. Almkvist, K. Blennow, K. Engedahl, A. Johansson, G. Waldemar, and H. Wolf, "Evidence-based Evaluation of Magnetic Resonance Imaging as a Diagnostic Tool in Dementia Workup," *Top. Magn. Reson. Imaging*, vol. 16, no. 6, pp. 427–437, Dec. 2005.

[240] S. Leandrou, I. Mamais, S. Petroudi, P. A. Kyriacou, C. C. Reyes-Aldasoro, and C. S. Pattichis, "Hippocampal and entorhinal cortex volume changes in Alzheimer's disease patients and mild cognitive impairment subjects," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2018, pp. 235–238.

[241] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, *et al.*, "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI

database," *Neuroimage*, vol. 56, no. 2, pp. 766–781, May 2011.

[242] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nat. Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, Feb. 2010.

[243] C. Bauer, H. Cabral, and R. Killiany, "Multimodal Discrimination between Normal Aging, Mild Cognitive Impairment and Alzheimer's Disease and Prediction of Cognitive Decline," *Diagnostics*, vol. 8, no. 1, p. 14, Feb. 2018.

[244] X. Li, H. Xia, Z. Zhou, and L. Tong, "3D texture analysis of hippocampus based on MR images in patients with alzheimer disease and mild cognitive impairment," in *2010 3rd International Conference on Biomedical Engineering and Informatics*, 2010, pp. 1–4.

[245] L. Sørensen, C. Igel, N. Liv Hansen, M. Osler, M. Lauritzen, E. Rostrup, and M. Nielsen, "Early detection of Alzheimer's disease using MRI hippocampal texture," *Hum. Brain Mapp.*, vol. 37, no. 3, pp. 1148–1161, Mar. 2016.

[246] K. Oppedal, T. Eftestøl, K. Engan, M. K. Beyer, and D. Aarsland, "Classifying Dementia Using Local Binary Patterns from Different Regions in Magnetic Resonance Images," *Int. J. Biomed. Imaging*, vol. 2015, pp. 1–14, 2015.

[247] F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, F. Segovia, D. Salas-Gonzalez, D. Castillo-Barnes, I. A. Illán, and A. Ortiz, "Evaluating Alzheimer's Disease Diagnosis Using Texture Analysis," in *Medical Image Understanding and Analysis 2017. MIUA 2017. Communications in Computer and Information Science, vol 723*, 2017, pp. 470–481.

[248] P. A. Freeborough and N. C. Fox, "MR image texture analysis applied to the diagnosis and tracking of Alzheimer's disease," *IEEE Trans. Med. Imaging*, vol. 17, no. 3, pp. 475–478, Jun. 1998.

[249] R. Simões, A.-M. van Cappellen van Walsum, and C. H. Slump, "Classification and localization of early-stage Alzheimer's disease in magnetic resonance images using a patch-based classifier ensemble.," *Neuroradiology*, vol. 56, no. 9, pp. 709–21, Sep. 2014.

[250] R. Simões, C. Slump, and A. M. van Cappellen van Walsum, "Using local texture maps of brain MR images to detect Mild Cognitive Impairment," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 153–156.

[251] E.-J. Hwang, H.-G. Kim, D. Kim, H. Y. Rhee, C.-W. Ryu, T. Liu, Y. Wang, and G.-H. Jahng, "Texture analyses of quantitative susceptibility maps to differentiate Alzheimer's disease from cognitive normal and mild cognitive impairment," *Med. Phys.*, vol. 43, no. 8, pp. 4718–4728, Jul. 2016.

[252] E. Pellegrini, L. Ballerini, M. del C. Valdés-Hernández, F. M. Chappell, V.

González-Castro, D. Anblagan, S. Danso, S. M. Maniega, *et al.*, "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, 2018.

[253] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, *et al.*, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010.

[254] P. S. Aisen, R. C. Petersen, M. C. Donohue, A. Gamst, R. Raman, R. G. Thomas, S. Walter, J. Q. Trojanowski, *et al.*, "Clinical core of the Alzheimer's disease neuroimaging initiative: Progress and plans," *Alzheimer's Dement.*, vol. 6, no. 3, pp. 239–246, May 2010.

[255] F. Jessen, S. Wolfsgruber, B. Wiese, H. Bickel, E. Mösch, H. Kaduszkiewicz, M. Pentzek, S. G. Riedel-Heller, *et al.*, "AD dementia risk in late MCI, in early MCI, and in subjective memory impairment," *Alzheimer's Dement.*, vol. 10, no. 1, pp. 76–83, Jan. 2014.

[256] C. Konrad, T. Ukas, C. Nebel, V. Arolt, A. W. Toga, and K. L. Narr, "Defining the human hippocampus in cerebral magnetic resonance images—An overview of current segmentation protocols," *Neuroimage*, vol. 47, no. 4, pp. 1185–1195, Oct. 2009.

[257] S. Gourtsoyianni, G. Doumou, D. Prezzi, B. Taylor, J. J. Stirling, N. J. Taylor, M. Siddique, G. J. R. Cook, *et al.*, "Primary Rectal Cancer: Repeatability of Global and Local-Regional MR Imaging Texture Features," *Radiology*, vol. 284, no. 2, pp. 552–561, Aug. 2017.

[258] J. H. McDonald, *Handbook of Biological Statistics*, 3rd ed. Baltimore, Maryland, U.S.A.: Sparky House Publishing, 2014.

[259] N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor, "Presymptomatic hippocampal atrophy in Alzheimer's disease," *Brain*, vol. 119, no. 6, pp. 2001–2007, 1996.

[260] H. Wolf, M. Grunwald, F. Kruggel, S. . Riedel-Heller, S. Angerhöfer, A. Hojjatoleslami, A. Hensel, T. Arendt, *et al.*, "Hippocampal volume discriminates between normal cognition; questionable and mild dementia in the elderly," *Neurobiol. Aging*, vol. 22, no. 2, pp. 177–186, Mar. 2001.

[261] V. A. Kovalev, F. Kruggel, H. J. Gertz, and D. Y. Von Cramon, "Three-dimensional texture analysis of MRI brain datasets," *IEEE Trans. Med. Imaging*, vol. 20, no. 5, pp. 424–433, 2001.

[262] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M. del C. Valdés-Hernández, *et al.*, "White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks," *NeuroImage Clin.*, vol. 17, pp. 918–934, 2018.

[263] C. Qin, R. Guerrero, C. Bowles, L. Chen, D. A. Dickie, M. del C. Valdés-Hernández, J. Wardlaw, and D. Rueckert, "A large margin algorithm for automated segmentation of white matter hyperintensity," *Pattern Recognit.*, vol. 77, pp. 150–159, May 2018.

[264] C. Bowles, C. Qin, C. Ledig, R. Guerrero, R. Gunn, A. Hammers, E. Sakka, D. A. Dickie, *et al.*, "Pseudo-healthy Image Synthesis for White Matter Lesion Segmentation," in *Simulation and Synthesis in Medical Imaging. SASHIMI 2016. Lecture Notes in Computer Science*, vol. 9968, S. A. Tsaftaris, A. Gooya, A. F. Frangi, and J. L. Prince, Eds. Athens, Greece: Springer, Cham, 2016, pp. 87–96.

[265] J. M. Wardlaw, M. del C. Valdés-Hernández, and S. Muñoz-Maniega, "What are White Matter Hyperintensities Made of?," *J. Am. Heart Assoc.*, vol. 4, no. 6, Jun. 2015.

[266] L. Puy, M. Barbay, M. Roussel, S. Canaple, C. Lamy, A. Arnoux, C. Leclercq, J.-L. Mas, *et al.*, "Neuroimaging Determinants of Poststroke Cognitive Performance," *Stroke*, vol. 49, no. 11, pp. 2666–2673, Nov. 2018.

[267] M. Habes, G. Erus, J. B. Toledo, T. Zhang, N. Bryan, L. J. Launer, Y. Rosseel, D. Janowitz, *et al.*, "White matter hyperintensities and imaging patterns of brain ageing in the general population," *Brain*, vol. 139, no. 4, pp. 1164–1179, Apr. 2016.

[268] M. del C. Valdés-Hernández, T. Booth, C. Murray, A. J. Gow, L. Penke, Z. Morris, S. M. Maniega, N. A. Royle, *et al.*, "Brain white matter damage in aging and cognitive ability in youth and older age," *Neurobiol. Aging*, vol. 34, no. 12, pp. 2740–2747, Dec. 2013.

[269] S. D. J. Makin, F. N. Doubal, M. S. Dennis, and J. M. Wardlaw, "Clinically Confirmed Stroke With Negative Diffusion-Weighted Imaging Magnetic Resonance Imaging," *Stroke*, vol. 46, no. 11, pp. 3142–3148, Nov. 2015.

[270] L. Viksne, M. del C. Valdés-Hernández, K. Hoban, A. K. Heye, V. González-Castro, and J. Wardlaw, "Textural Characterisation on Regions of Interest : A Useful Tool for the Study of Small Vessel Disease," *Proc. 19th Conf. Med. Image Underst. Anal.*, pp. 66–71, 2015.

[271] E. L. Bailey, C. Smith, C. L. M. Sudlow, and J. M. Wardlaw, "Pathology of Lacunar Ischemic Stroke in Humans-A Systematic Review," *Brain Pathol.*, vol. 22, no. 5, pp. 583–591, Sep. 2012.

[272] J. M. Wardlaw, C. Smith, and M. Dichgans, "Mechanisms of sporadic cerebral small vessel disease: insights from neuroimaging," *Lancet Neurol.*, vol. 12, no. 5, pp. 483–497, May 2013.

[273] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O'Brien, *et al.*, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," *Lancet Neurol.*, vol. 12, no. 8, pp. 822–838, Aug. 2013.

[274] M. del C. Valdés-Hernández, P. A. Armitage, M. J. Thrippleton, F. Chappell, E. Sandeman, S. Muñoz Maniega, K. Shuler, and J. M. Wardlaw, "Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke," *Brain Behav.*, vol. 5, no. 12, p. e00415, Dec. 2015.

[275] J. M. Wardlaw, S. J. Makin, M. del C. Valdés-Hernández, P. A. Armitage, A. K. Heye, F. M. Chappell, S. Muñoz-Maniega, E. Sakka, *et al.*, "Blood-brain barrier failure as a core mechanism in cerebral small vessel disease and dementia: evidence from a cohort study," *Alzheimer's Dement.*, vol. 13, no. 6, pp. 634–643, Jun. 2017.

[276] I. J. Deary, A. J. Gow, M. D. Taylor, J. Corley, C. Brett, V. Wilson, H. Campbell, L. J. Whalley, *et al.*, "The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond," *BMC Geriatr.*, vol. 7, no. 1, p. 28, Dec. 2007.

[277] I. J. Deary, A. J. Gow, A. Pattie, and J. M. Starr, "Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936," *Int. J. Epidemiol.*, vol. 41, no. 6, pp. 1576–1584, Dec. 2012.

[278] M. del C. Valdés-Hernández, K. J. Ferguson, F. M. Chappell, and J. M. Wardlaw, "New multispectral MRI data fusion technique for white matter lesion segmentation: method and comparison with thresholding in FLAIR images," *Eur. Radiol.*, vol. 20, no. 7, pp. 1684–1691, Jul. 2010.

[279] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *Neuroimage*, vol. 62, no. 2, pp. 782–790, Aug. 2012.

[280] W. H. Xu, "Large artery: an important target for cerebral small vessel diseases," *Ann. Transl. Med.*, vol. 2, no. 8, 2014.

[281] M. del C. Valdés-Hernández, L. C. Maconick, S. Muñoz Maniega, X. Wang, S. Wiseman, P. A. Armitage, F. N. Doubal, S. Makin, *et al.*, "A Comparison of Location of Acute Symptomatic vs. 'Silent' Small Vessel Lesions," *Int. J. Stroke*, vol. 10, no. 7, pp. 1044–1050, Oct. 2015.

# Chapter 11.
# Publications

## 11.1. Publications from the PhD Thesis

In this section, the contributions published in the context of this thesis are shown. The works published in journals are indicated with black triangles (►), while the works presented in international conferences are indicated with black squares (■).

**Chapter 5. Differentiation between Brain Metastases and Glioblastomas**

► R. Ortiz-Ramón, S. Ruiz-España, E. Mollá-Olmos, and D. Moratal, "Texture analysis on MRI to differentiate between Glioblastomas and Brain Metastases following a Radiomics approach," *Radiology*.
(Submitted)

**Chapter 6. Classification of Brain Metastases by their Primary Site of Origin**

► R. Ortiz-Ramón, A. Larroza, S. Ruiz-España, E. Arana, and D. Moratal, "Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study," *European Radiology*, vol. 28, no. 11, pp. 4514–4523, Nov. 2018.
DOI: 10.1007/s00330-018-5463-6

- R. Ortiz-Ramon, A. Larroza, E. Arana, and D. Moratal, "A radiomics evaluation of 2D and 3D MRI texture features to classify brain metastases from lung cancer and melanoma," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'17)*, Jeju Island, Korea; Jul. 2017.
  DOI: 10.1109/EMBC.2017.8036869

- R. Ortiz-Ramon, A. Larroza, E. Arana, and D. Moratal, "Identifying the primary site of origin of MRI brain metastases from lung and breast cancer following a 2D radiomics approach," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Melbourne, Australia; Apr. 2017.
  DOI: 10.1109/ISBI.2017.7950735

**Chapter 7. Evaluation of New Biomarkers for Alzheimer's disease**

- C. López-Gómez, R. Ortiz-Ramón, E. Mollá-Olmos, and D. Moratal, "ALTEA: A Software Tool for the Evaluation of New Biomarkers for Alzheimer's Disease by Means of Textures Analysis on Magnetic Resonance Images," *Diagnostics*, vol. 8, no. 3, p. 47, Jul. 2018.
  DOI: 10.3390/diagnostics8030047

- C. López-Gómez, R. Ortiz-Ramón, and D. Moratal, "ALTEA (Alzheimer TExture Analyzer)," in *21st Annual Conference on Medical Image Understanding and Analysis (MIUA 2017)*, Edinburgh, UK; Jul. 2017.

**Chapter 8. Characterization of Ischemic Stroke**

- R. Ortiz-Ramón, M. del C. Valdés-Hernández, V. González-Castro, S. Makin, P. A. Armitage, B. S. Aribisala, M. E. Bastin, I. J. Deary, J. M. Wardlaw and David Moratal, "Identification of the presence of ischaemic stroke lesions by means of texture analysis on brain magnetic resonance images," *Computerized Medical Imaging and Graphics*.
  (Submitted)

## 11.2. Other publications

In this section, additional contributions published in the context of other projects not included in this thesis are presented. The works published in journals are indicated with black triangles (►), while the works presented in international conferences are indicated with black squares (■).

► E. Ozturk-Isik, I. Marshall, P. Filipiak, A. J. V. Benjamin, V. G. Ones, R. Ortiz-Ramón, and M. del C. Valdés-Hernández, "Workshop on reconstruction schemes for magnetic resonance data: summary of findings and recommendations," *R. Soc. Open Sci.*, vol. 4, no. 2, p. 160731, Feb. 2017.
DOI: 10.1098/rsos.160731

■ S. Ruiz-España, R. Ortiz-Ramón, U. Perez-Ramirez, A. Diaz-Parra, R. Ciccocioppo, S. Canals, and D. Moratal, "3D Texture Analysis on fMRI to Detect Alterations in the Striatal Network of an Alcohol-Preferring Rat Model," in *27th Joint Annual Meeting ISMRM - ESMRMB 2018*, Paris, France; Jun. 2018.

■ R. Ortiz-Ramón, S. Ruiz-España, U. Perez-Ramirez, A. Diaz-Parra, R. Ciccocioppo, S. Canals, and D. Moratal, "Evaluation of texture features on resting-state networks of a rat model of alcohol use disorders," in *34th Annual Scientific Meeting of the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB 2017)*, Barcelona, Spain; Oct. 2017.
DOI: 10.1007/s10334-017-0635-y

■ S. Ruiz-España, R. Ortiz-Ramón, U. Perez-Ramirez, A. Diaz-Parra, R. Ciccocioppo, S. Canals, and D. Moratal, "Evaluation of 2D texture analysis on fMRI data to identify changes in the striatal network induced by alcohol drinking" in *34th Annual Scientific Meeting of the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB 2017)*, Barcelona, Spain; Oct. 2017.
DOI: 10.1007/s10334-017-0635-y

- R. Ortiz-Ramón, A. Llorca, Ú. Pérez-Ramírez, A. Diaz-Parra, O. Marín, and D. Moratal, "Registration of mouse brain microscopy images to a MR mouse brain atlas for locating interneuron cells: a preliminary study," in *34th Annual Scientific Meeting of the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB 2017)*, Barcelona, Spain; Oct. 2017.
  DOI: 10.1007/s10334-017-0635-y

- R. Ortiz-Ramón, J. M. Morales, S. Ruiz-España, V. Bodi, D. Monleon, and D. Moratal, "Magnetic resonance microimaging of a swine infarcted heart: Performing cardiac virtual histologies," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'15)*, Milano, Italy; Aug. 2015.
  DOI: 10.1109/EMBC.2015.7318676

*"Para empezar, diré que es el final.*
*No es un final feliz, tan solo es un final.*
*Pero parece ser que ya no hay vuelta atrás"*

M-Clan, Miedo, 2004