

UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Telecomunicación



**Propuesta de una nueva
metodología basada en la imagen
médica para la cuantificación de la
afectación de la médula ósea**

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Eva Milara Hernando

MSc in Biomedical Engineering

Madrid, 2025



UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de
Telecomunicación

Doctoral Degree in Biomedical Engineering

**Propuesta de una nueva
metodología basada en la imagen
médica para la cuantificación de la
afectación de la médula ósea**

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Eva Milara Hernando

MSc in Biomedical Engineering

Under the supervision of:

Dr. Patricia Sánchez González

Madrid, 2025

Title: Propuesta de una nueva metodología basada en la imagen médica para la
cuantificación de la afectación de la médula ósea

Author: Eva Milara Hernando

Doctoral Programme: Doctoral Degree in Biomedical Engineering

Thesis Supervision:

Dr. Patricia Sánchez González, Universidad Politécnica de Madrid (Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been partially supported by a predoctoral Fellowship (ayuda del Programa Propio de I+D+i 2020) from Universidad Politécnica de Madrid

A mi madre, por ser mi heroína.

A David, por un futuro juntos.

Acknowledgement

Hace cuatro años veía este día como algo muy lejano, un reto al que se enfrentaría la Eva del futuro. Hace diez años, mientras empezaba la universidad, jamás pensé que acabaría haciendo una tesis. Pero la vida da muchas vueltas y lo importante no es el destino, es la compañía durante el camino.

Me gustaría dar las gracias a todas las personas con las que he tenido el placer de coincidir en el Grupo de Bioingeniería y Telemedicina (GBT) a lo largo de estos años. En primer lugar, gracias a Paty por tu confianza y por acompañarme hasta el final. Gracias a Enrique por hacer del GBT lo que es hoy. Gracias a mis compañeros y compañeras, Ana, Jose, Paloma, Carlos, Gema, Elena, Javier, Manu, Paula, Ainara y Nacho, por hacer que no todo sea trabajar, y haya cada día un poquito de alegría. Gracias a Diego por las risas y todas las enseñanzas que me llevo, y gracias a Paco por escucharme, ayudarme y ser el descubrimiento que jamás pensé que vendría con la tesis. Pero especialmente, gracias a Alex, gracias por estar siempre, por preocuparte y por dedicarle más tiempo del que te podía pedir.

Este trabajo no habría sido posible sin todas aquellas personas del Hospital Universitario 12 de Octubre de Madrid que han participado en esta investigación. Gracias a Adolfo y Pilar del Servicio de Medicina Nuclear, y gracias a Rafa y Ana, del Servicio de Hematología. Gracias por vuestra paciencia conmigo, por vuestro amor a la medicina y por ayudarme a intentar hacer del mundo un lugar un poquito mejor.

Gracias a mis amigos, por hacer de la vida *un viaje que merezca la pena*. Gracias a todos aquellos amigos que hice en este largo camino y aún conservo, sobre todo Paula por acompañarme en la aventura que fue Gante. Pero especialmente quiero agradecerse a mi Equipo Z, Andrea, María y Bea, por el amor que me dais cada día con una amistad verdadera, esa que sufre cuando sufres, te apoya cuando lo necesitas, y se alegra por tus logros como si fueran suyos. Os quiero, no me faltéis nunca.

Gracias a toda mi familia, que siempre ha confiado en que yo era capaz de hacer lo que me propusiera. Gracias a mi yaya, por ese querer tan suyo. Gracias a Silvia, Ana y Mario, por recordarme lo afortunada que soy por tener esta familia. Gracias a María, por ser como una hermana. Gracias a Alberto, por quererme aún con nuestras diferencias, por tus lentejas y por traer a mi vida a mi vieja gruñona y a mi bebé, Trapi y Kupa. Y gracias a mi tita Jose, ojalá hubieras visto donde hemos llegado.

Gracias a David, mi persona favorita, *¡qué casualidad, coincidir en estos tiempos, en el mismo lugar!* Gracias por hacer de esa casualidad mi hogar. Gracias por tu apoyo incondicional durante estos largos nueve años, por ser mi compañero de vida. Gracias

por enseñarme lo que es el amor. Gracias por demostrarme que alguien que te quiere bien, puede llegar a convertirse en tu familia. Gracias por darme más felicidad y risas de la que creía que podía traerme una sola persona, solo deseo *que los mejores momentos estén por llegar*.

Todos ellos han sido importantes, pero nada de esto habría sido posible sin mi mamá, Eloisa, mi otra persona favorita. Gracias por demostrarme que una madre puede ser tu mejor amiga y compañera de aventuras. Gracias por hacer que sea *mucho más fácil amarte que odiarte*. Gracias por enseñarme tanto, por ser mi mayor fan y confiar en mi incluso cuando yo no podía. Espero haber hecho que te sientas orgullosa, porque todo esto es gracias a ti, y espero poder descubrir el mundo entero a tu lado.

Abstract

The bone marrow (BM) is the main organ responsible for the production of blood and immune cells. Its function can be affected mainly by: plasma cell or lymphoid neoplasms. From the first group, Multiple Myeloma (MM), which directly affects the status of the BM, stands out. A high percentage of treated MM patients suffer from persistent tumour burden, a condition known as Residual Measurable Disease (MRD), which is associated with a higher risk of relapse and worse prognosis. Prior to MM, there is a condition known as Smouldering MM (SMM), characterised by progression, over an uncertain period of time, to MM. To improve personalisation of treatment, there is growing interest in defining the risk of progression to MM. From the second group, Non-Hodgkin's Lymphoma (NHL) comprises 90% of lymphoma cases, with Follicular Lymphoma (FL) and Diffuse Large B-cell Lymphoma (DLBCL) being the most common subtypes. Although not the primary lesion, BM infiltration (BMI) occurs in 20% to 40% of patients with NHL, with important therapeutic and prognostic implications.

Although BM biopsy is the gold standard for the diagnosis of BM states, its combination with visual assessment of [^{18}F]FDG PET/CT images by experts in nuclear medicine is recommended. The main reason is that, despite its high sensitivity, the sample taken may not be representative of the entire BM, whereas imaging assesses the whole organ. However, the assessment of these images is purely visual and therefore observer-dependent, as well as doubtful in many cases.

For this reason, the present thesis focuses on the proposal and development of methodologies for quantification of BM from [^{18}F]FDG PET/CT images to aid visual interpretation by nuclear medicine experts. In addition, for the development of a suitable quantification methodology, a methodology of BM segmentation is previously developed.

The main results of the present thesis are grouped in five areas:

- Development of an automatic BM segmentation methodology from CT images. Thanks to its performance, this segmentation allows a more accurate quantification. Moreover, its inclusion in a platform that allows not only automatic segmentation, but also manual editing, generates a useful tool for this line of research.
- Implementation of quantification methodologies for the detection of MRD. The potential of the quantification methodologies developed for the detection of MRD from the images used for the evaluation of response to treatment in MM patients is demonstrated. In addition, an experiment using only patients assessed as negative by imaging is performed, demonstrating the potential of

these methodologies in the detection of patterns useful for the identification of this condition that are not detectable to the human eye.

- Implementation of quantification methodologies for the prediction of MRD. Using MM diagnostic images, without considering any treatment-dependent factors, the potential of radiomics feature-based methodology for the prediction of MRD is demonstrated.
- Identification of potential biomarkers based on diagnostic imaging of SMM for the assessment of risk of progression to MM. Some of the radiomics features extracted from [^{18}F]FDG PET/CT images of SMM diagnosis are identified as potential biomarkers in differentiating the risk of progression to MM.
- Detection of BMI and its differentiation from other BM stages in NHL staging. We demonstrate the potential of the methodologies developed for the differentiation of possible BM status during NHL diagnosis, improving staging and, consequently, improving the personalisation of treatment.

The results presented lay the foundation for the identification of the possible status of BM in the main pathologies by which it is affected, assisting clinical experts. In this way, they provide the bases for tools to improve the personalisation of medicine and the clinical management of patients.

Resumen

La médula ósea (MO) es el principal órgano encargado de la producción de células sanguíneas e inmunes. Su función puede verse afectada principalmente por: neoplasmas de células plasmáticas o linfoides. Del primer grupo, destaca el Mieloma Múltiple (MM), que afecta directamente al estado de la MO. Un alto porcentaje de pacientes de MM tratados sufren la persistencia de carga tumoral, condición conocida como Enfermedad Residual Medible (ERM), relacionada con un mayor riesgo de recaída y un peor pronóstico. Previo al MM, existe una condición conocida como *Smouldering* MM (SMM), caracterizada por progresar, en un periodo de tiempo incierto, a MM. Para mejorar la personalización del tratamiento, está creciendo el interés en definir el riesgo de progresión a MM. Del segundo grupo, el Linfoma No-Hodgkin (LNH) abarca el 90% de casos de linfoma, siendo sus subtipos más habituales el Linfoma Folicular (LF) y el Linfoma Difuso de Células-B Grandes (LDCBG). A pesar de no ser la lesión principal, la infiltración de la MO (IMO) se produce entre un 20% y un 40% de pacientes con LNH, con importantes implicaciones terapéuticas y pronósticas.

A pesar de ser la biopsia de MO el *gold standard* para el diagnóstico de los estados de la MO, se recomienda su combinación con la valoración visual de imágenes $[^{18}\text{F}]\text{FDG}$ PET/CT por parte de expertos en Medicina Nuclear. El principal motivo es que, a pesar de su alta sensibilidad, la muestra tomada puede no ser representativa de toda la MO, mientras que las imágenes evalúan el órgano completo. Sin embargo, la evaluación de estas imágenes es meramente visual, por lo que es observador-dependiente, además de dudosa en muchos casos.

Por este motivo, la presente tesis se centra en la propuesta y desarrollo de metodologías de cuantificación de la MO a partir de imágenes $[^{18}\text{F}]\text{FDG}$ PET/CT para ayudar a su interpretación visual a los expertos en medicina nuclear. Además, para el desarrollo de una metodología de cuantificación adecuada, se desarrolla previamente una metodología de segmentación de MO.

Los principales resultados de la presente tesis se agrupan en cinco áreas:

- Desarrollo de una metodología de segmentación automática de MO a partir de imágenes de CT. Gracias a su rendimiento, esta segmentación permite una cuantificación más precisa. Además, su inclusión en una plataforma que permite, no solo la segmentación automática, sino también una edición manual, genera una herramienta útil para esta línea de investigación.
- Implementación de las metodologías de cuantificación para la detección de la ERM. Se demuestra el potencial de las metodologías de cuantificación desarrolladas para la detección de la ERM a partir de las imágenes utilizadas

para la evaluación de respuesta al tratamiento en pacientes de MM. Además, se realiza un experimento utilizando exclusivamente los pacientes valorados como negativos mediante las imágenes, demostrando el potencial de estas metodologías en la detección de patrones útiles para la identificación de esta condición no detectables para el ojo humano.

- Implementación de las metodologías de cuantificación para la predicción de la ERM. Utilizando las imágenes de diagnóstico de MM, sin tener en cuenta ningún factor dependiente del tratamiento, se demuestra el potencial de la metodología basada en características de radiómica para la predicción de la ERM.
- Identificación de potenciales biomarcadores basados en imagen de diagnóstico de SMM para la evaluación del riesgo de progresión a MM. Algunas de las características de radiómica extraídas de las imágenes [^{18}F]FDG PET/CT del diagnóstico de SMM se identifican como potenciales biomarcadores en la diferenciación del riesgo de progresión a MM.
- Detección de la IMO y su diferenciación de otros estados de MO en la estadificación de LNH. Se demuestra el potencial de las metodologías desarrolladas para la diferenciación de los posibles estados de la MO durante el diagnóstico de LNH, mejorando la estadificación y, a consecuencia, la mejora de la personalización del tratamiento.

Los resultados presentados sientan las líneas futuras dirigidas a la identificación de los posibles estados de la MO en las principales patologías por las que se ve afectada, ayudando a los expertos clínicos. De esta manera, se establecen las bases para conseguir herramientas que permitan la mejora de la personalización de la medicina y la gestión clínica de los pacientes.

Table of Contents

1. Introduction	1
1.1. Bone marrow bases	1
1.1.1. Haematopoiesis	2
1.2. Bone marrow pathological affectations.....	2
1.2.1. Plasma cell neoplasms	2
1.2.1.1. Multiple Myeloma.....	3
1.2.1.2. Monoclonal Gammopathy of Undetermined Significance	4
1.2.1.3. Smouldering Multiple Myeloma.....	4
1.2.1.4. Measurable Residual Disease	5
1.2.2. Lymphoid neoplasms	6
1.2.2.1. Follicular lymphoma.....	7
1.2.2.2. Diffuse large B-cell lymphoma.....	8
1.2.2.3. Bone marrow infiltration in lymphoproliferative disorders.....	8
1.2.2.4. Reactive bone marrow	10
1.3. Bone marrow assessment.....	10
1.3.1. Haematologic assessment.....	10
1.3.2. Bone marrow assessment in Multiple Myeloma	11
1.3.2.1. [¹⁸ F]FDG PET/CT assessment	12
1.3.3. Bone marrow involvement assessment in NHL.....	15
1.3.3.1. [¹⁸ F]FDG PET/CT assessment	15
1.4. Problem statement and justification	17
1.5. Thesis Outline.....	18
2. Research hypotheses and objectives.....	19
2.1. Research hypotheses	19
2.2. Objectives	23
3. State of art	25
3.1. Bone and bone marrow segmentation	25
3.1.1. Bone segmentation.....	25
3.1.2. Bone marrow segmentation	32
3.2. Bone marrow quantification.....	34
3.2.1. Quantification of bone marrow for Multiple Myeloma	35
3.2.2. Quantification of bone marrow for Non-Hodgkin Lymphoma	41
4. Materials and methods.....	46
4.1. Materials	47
4.2. Bone marrow segmentation.....	48
4.2.1. Semi-automatic segmentation based on traditional algorithms.....	49
4.2.1.1. Stretcher removal	49

4.2.1.2.	Bone segmentation.....	50
4.2.1.3.	Spinal canal subtraction	53
4.2.1.4.	Compact bone removal	54
4.2.1.5.	Manual revision	55
4.2.2.	Automatic segmentation based on Deep Learning.....	55
4.2.2.1.	Bone segmentation.....	56
4.2.2.2.	Bone marrow segmentation	67
4.2.3.	Evaluation of segmentation models.....	72
4.3.	Bone marrow quantification in [¹⁸ F]FDG PET/CT.....	74
4.3.1.	Pseudo-ground truth labelling.....	75
4.3.2.	PET processing	76
4.3.3.	Quantification by Machine Learning models based on radiomics.....	77
4.3.3.1.	Radiomics features extraction.....	77
4.3.3.2.	Feature groups creation	85
4.3.3.3.	Machine Learning implementation	90
4.3.4.	Quantification using Convolutional Neural Networks	94
4.3.4.1.	Inputs creation.....	95
4.3.4.2.	Deep Learning models.....	97
4.3.5.	Performance assessment	105
5.	Results	109
5.1.	Bone marrow segmentation.....	109
5.1.1.	Platform for bone and bone marrow segmentation.....	110
5.1.1.1.	Automatic segmentation tabs	111
5.1.1.2.	Editing tab.....	117
5.1.2.	Semi-automatic segmentation based on traditional algorithms	129
5.1.2.1.	Bone segmentation.....	130
5.1.2.2.	Bone marrow segmentation	132
5.1.3.	Automatic segmentation using Deep Learning models	134
5.1.3.1.	Bone segmentation.....	134
5.1.3.2.	Bone marrow segmentation	137
5.1.4.	Discussion	141
5.1.5.	Conclusions	148
5.2.	Bone marrow Multiple Myeloma quantification in [¹⁸ F]FDG PET/CT.....	148
5.2.1.	Measurable Residual Disease detection	148
5.2.1.1.	Quantification by Machine Learning models	149
5.2.1.2.	Quantification using Convolutional Neural Networks	151
5.2.1.3.	Discussion	154
5.2.1.4.	Conclusions	158
5.2.2.	Measurable Residual Disease prediction.....	158
5.2.2.1.	Quantification results	159

5.2.2.2.	Discussion	160
5.2.2.3.	Conclusions	161
5.2.3.	Risk assessment in Multiple Myeloma Smouldering	161
5.2.3.1.	Quantification by univariate analysis	162
5.2.3.2.	Discussion	166
5.2.3.3.	Conclusions	168
5.3.	Bone marrow Non-Hodgkin's Lymphoma quantification in [¹⁸ F]FDG PET/CT	168
5.3.1.	Quantification by radiomics extraction.....	170
5.3.2.	Quantification using Deep or Machine Learning models	175
5.3.3.	Discussion	179
5.3.4.	Conclusions	183
6.	Conclusions and future works	185
6.1.	Discussion of research hypotheses	185
6.2.	Contributions of the thesis.....	190
6.2.1.	Main contributions	190
6.2.2.	Scientific publications	197
6.2.2.1.	Journal articles	197
6.2.2.2.	National congresses.....	198
6.3.	Future works.....	200
6.3.1.	Future works based on limitations.....	200
6.3.2.	New research lines	203
References.....		206
Annex A.	Hyperparameters grid.....	235
Annex B.	Instructions of the automatic segmentation tab	239
Annex C.	Extra information of the Editing tab.....	242
Annex D.	Instructions of the Editing tab	244
Annex E.	Unprocessed bone segmentation	246
Annex F.	Stretcher removal model results	247
Annex G.	Performance of ML models MRD detection.....	248
Annex H.	Performance of CNNs in MRD detection.....	251
Annex I.	Performance of ML models in MRD prediction.....	258
Annex J.	Univariate analysis results for risk assessment in SMM.....	260
Annex K.	BM quantification in NHL patients by ML models.....	267
Annex L.	BM quantification in NHL patients by CNNs	279

List of Figures

Figure 1. Bone anatomy representation [4].	1
Figure 2. Distribution of B-cell, T-cell, and Hodgkin lymphomas among the nodal or extranodal lymphomas. Adapted from [39].	6
Figure 3. Relative frequencies of B-cell lymphomas. Adapted from [39].	7
Figure 4. Patterns of BMI in lymphoma disorders [1].	9
Figure 5. Kaplan-Meier curves with MFC and ^{18}F FDG PET assessments for (a) Progression-Free Survival (PFS) and (b) Overall Survival (OS). Adapted from [53].	12
Figure 6. MIP of ^{18}F FDG PET images of newly diagnosed MM patients, representing examples of different pathologic patterns: (a) multiple skeleton focal lesions; (b) diffuse uptake in the BM without focal lesions; (c) diffuse uptake pattern and focal BM lesions. Adapted from [59].	13
Figure 7. MIP of ^{18}F FDG PET post-treatment images of MM patients, representing: (a) complete response; (b) MRD BM uptake pattern.	14
Figure 8. MIP of ^{18}F FDG PET images of FL patients representing different patterns of FDG uptake: (a) pure diffuse uptake pattern, considered BMI-; (b) multifocal uptake pattern without diffuse uptake, considered BMI+; (c) multifocal uptake with additional diffuse uptake greater than liver reference, considered BMI+. The arrows indicate some of the most evident areas of increased uptake. Adapted from [47].	16
Figure 9. Illustration of hypothesis definition.	20
Figure 10. Results of the vertebral column segmentation with the method of Banik <i>et al.</i> [94]: (a) good and (b) acceptable segmentation. (Black) contour of the vertebrae drawn by the radiologist; (yellow) contour produced by the Banik <i>et al.</i> method.	26
Figure 11. Results of the vertebrae segmentation for (a) Kratky <i>et al.</i> [95] and (b) Zhang <i>et al.</i> [96]. In (a) different colours represent different level of smoothness: (red) low smoothness with segmentation artefacts; (green) excessive smoothing; (blue) properly for fitting the boundary correctly.	27
Figure 12. Results of Yousefi <i>et al.</i> [98] methodology: (a) segmentation delineated in red over the CT image; (b): (white) true positive pixels; (pink) false positive; (green) false negative; and (black) true negative.	28

Figure 13. Results of the segmentation for (a) Ababneh <i>et al.</i> [99] and (b) Younes <i>et al.</i> [100] methods. In (a/b) different colours represent: (yellow/blue) GT and (white/yellow) segmentation obtained with the proposed algorithm.....	29
Figure 14. Examples of the segmentation obtained with the methodology proposed by Pérez-Carrasco <i>et al.</i> [77] for (a) pelvic, (b) femurs, and (c) shoulders and ribs structures. (Red) GT; (blue) obtained segmentation.	29
Figure 15. Results of the segmentation for (a) Korez <i>et al.</i> [101] showing in red the vertebrae segmentation results; (b) Klein <i>et al.</i> [78] showing (superior) the GT and (inferior) the result of the segmentation methodology; and (c) Noguchi <i>et al.</i> [79] showing (superior) the original CT and (inferior) the result of the segmentation methodology.....	30
Figure 16. Examples of the (a) Li <i>et al.</i> [107] and (b) Fiandra <i>et al.</i> [108] segmentations of active BM. In (b): (yellow) [^{18}F]FDG PET and (red) Decision Tree Classifier.....	32
Figure 17. Segmentation of Sambuceti <i>et al.</i> [83]. (a) division of the histogram in the different regions; (b) result of the division of the histogram overlayed in the CT slice.	33
Figure 18. Results for (a) compact bone and (b) BM segmentation in Martínez-Martínez <i>et al.</i> [84] and (c) Nishida <i>et al.</i> [85] BM segmentation. For (c), red rectangle represents horizontal and vertical ranges selected for the segmentation, and the rest of colours are according to the HU: (yellow-red) high, and (blue-purple) low.	33
Figure 19. Workflow for the BM quantification. GTseg: GT for the segmentation. DDBB: database of a concrete clinical context.	47
Figure 20. Pipeline of the automatic segmentation based on traditional image processing algorithms. Axial skeleton represents the bone segmentation from the torso, i.e. from neck to pelvis, excluding limbs.....	49
Figure 21. Stretcher removal process represented over a CT thresholded between 0 and 300 HU for a better visualization. The mask of the stretcher correspondent to each step is shown in blue. (a) stretcher mask as a result of a threshold of 98 HU in the CT; (b) mask of (a) after the filling operation; (c) mask after selection of the greatest volume; (d) correction of the volume selected based on the position.	50
Figure 22. Representation of the whole bone segmentation, with different colours for each segmented structure: (blue) stretcher, (purple) right femur, (yellow) left femur, (orange) right humerus, (pink) left humerus, (green) axial skeleton.	

Examples of slices at different heights (a) femurs region; (b and c) ribs region. (d) complete volume over thresholded CT differing by the same colours.....	52
Figure 23. (a) Coronal, (b) sagittal and (c) axial planes of the bone mask (green) over the CT.	53
Figure 24. (a) Coronal, (b) sagittal and (c) axial planes of the spinal canal mask (green) over the CT. (d) Coronal, (e) sagittal and (f) axial planes of the bone mask (green) after the spinal canal subtraction over the CT.	54
Figure 25. (a) Coronal, (b) sagittal and (c) axial planes of the BM mask (green) over the CT.	55
Figure 26. Simplified pipeline for obtaining the BM segmentation model. ① BM segmentation based on an untrained U-Net model; ② BM segmentation based on retraining the bone segmentation model.	56
Figure 27. Pipeline for the pre-processing steps of the bone segmentation model. X: unprocessed data for train set; Y: unprocessed data for test set; X': pre-processed data for train set; and Y': pre-processed data for test set.	57
Figure 28. U-Net architecture from Ronneberger <i>et al.</i> [102]. Blue boxes: multi-channel feature map; white boxes: copied feature maps; arrows: operations; number on the top of the box: number of channels of each feature map; number at the lower left edge: x-y-size of the image.	58
Figure 29. U-net architecture adapted for input image of 512x512 resolution. Blue boxes: multi-channel feature map; arrows: operations; number inside the box: number of channels of each feature map (F, 2F, 4F, 8F and 16F); number at the top of the box: x-y-size of the image. The left half of the diagram represents the contractive path and the right half the expansive path.	58
Figure 30. Dataset creation. Example of dataset creation with patient and slice randomization.	59
Figure 31. Bone segmentation mistakes for unprocessed database: (a) lack of BM in skeleton segmentation; (b) Exclusion of cortical bone regions of low grey intensity levels i.e. reduced thickness; (c) organs with higher grey intensity levels than bone inclusion; (d) stretcher inclusion due to its high grey intensity level.	60
Figure 32. Zoom of Figure 31 for a better visualization of the mistakes for the bone segmentation model trained with unprocessed database. Each subfigure is the correspondent zoom of the Figure 31, i.e. the (a) is the zoom of Figure 31a, (b) is the zoom of Figure 31b, etc.	60
Figure 33. Pipeline for obtaining the bone segmentation model. X: train set data; Y: test set data.	61

Figure 34. Stretcher removal I: margin location I. (a) Axial slice of the torso area extracted from the original CT scan; (b) slice shown in (a) threshold between 0 and 300 HU for a better visualization; (c) location of the stretcher points over the CT thresholded slice: closest to the left image margin in the horizontal central zone (i0_1, red), furthest from the left image margin in the central zone (i0_2, yellow), the point representing where most of the elevation of the stretcher ends (i0, green). 62

Figure 35. Stretcher removal I: margin location II. All images are shown over Figure 34b. (a) Location of the vertical lines that separate the stretcher from the body of the patient: in the central zone (i1, green), in the curve that goes from the central zone to the ends of the stretcher (i2, yellow), and at these ends (i3, red); (b) horizontal division of the slice in 8 equidistant regions (R1-R8, blue); (c) intersection points described as Pe (white) or Pi (pink), exterior or interior, shown over the correspondent vertical and horizontal lines (CT scan is not included for getting an easier visualization).....64

Figure 36. Stretcher removal I: margin location III. All images are shown over Figure 34b. (a) Combination of (Figure 35b) and (Figure 35c) to visualise the intersection points over the thresholded CT scan; (b) generation of the margin (blue) of separation between body and stretcher based on the previously calculated equidistant regions (green) and vertical lines of the indices (yellow), going through every intersection point; (c) margin (blue) separating the stretcher from the skeletal area shown in the thresholded CT slice.....64

Figure 37. Stretcher removal II: stretcher filling. Blue mask represents the stretcher mask. All images are shown over Figure 34b. (a) Threshold slice of the CT scan highlighting the pixels with intensity level higher or equal to 98 HU in the region defined by the left border and the calculated margin; (b) stretcher mask after filling holes; (c) stretcher mask after filling the area between stretcher and superior and inferior edge of the image; (d) stretcher mask after filling holes by morphological close operation. 65

Figure 38. Stretcher removal III: stretcher adaptation I. Blue mask represents the stretcher mask. All images are shown over Figure 34b. (a) Stretcher mask for a CT slice of the torso area without adaptation; (b) stretcher mask for a CT slice of the end of the femur with adaptation by means of dilatation, the green mask represents the adaptation with respect to the original stretcher; (c) zoom of (b) for a better visualization of the dilatation.....65

Figure 39. Stretcher removal III: stretcher adaptation II. Blue mask represents the stretcher mask over the thresholded CT. (a) Stretcher mask for a CT slice of the neck region with adaptation by means of stretcher accessory searching; (b) zoom

of (a) for a better visualization of the dilatation; (c) original CT scan slice substituting the values of the stretcher region by 0.	66
Figure 40. Thresholding process. (a) CT scan slice in (Figure 39c) thresholding pixels with values lower than or equal to 0 HU; (b) CT scan slice in (a) thresholding pixels with values greater than or equal to 300 HU.....	67
Figure 41. General pipeline for obtaining the BM segmentation model. For approaches I and II the U-Net from Ronneberger is the original architecture. For approach III, the model is the one obtained as the bone segmentation model. Although it is not marked for an easy visualization, in the approach II, the bone mask is also multiplied with the CT scan for creating the input of the model.	68
Figure 42. Examples of the pGT created for BM segmentation in: (a) and (c) pelvis regions, and (b) torso area. (Green) edges removed as a result of the step (2); (yellow) pixels over 300 HU in the original CT removed in the step (3); (blue) resulting BM mask after applied the four steps.	69
Figure 43. Zoom of the Figure 42 for a better visualization. Each subfigure is the correspondent zoom of the Figure 42, i.e. the (a) is the zoom of Figure 42a, (b) is the zoom of Figure 42b, and (c) is the zoom of Figure 42c.	70
Figure 44. Examples of stretcher masks used as GT for the stretcher removal model: (a) usual stretcher shape; (b) usual stretcher shape with some metallic material inside for a better support of the patient weight; (c) accessory metallic object for supporting the patient head.....	71
Figure 45. Model training pipeline for BM segmentation.	71
Figure 46. Example of two masks to visualise the IoU and Dice indexes. (Blue or A) GT mask; (yellow or B) predicted mask. (a) the example is explained with set theory; (b) the example is expressed as confusion matrix metrics.....	73
Figure 47. General pipeline of the BM quantification methodology. (Purple dashed box) approach corresponding to ML models; (yellow dashed box) approach corresponding to DL models. DB: database.	74
Figure 48. Pipeline of the quantification based on ML models. ORIG: original group of radiomics features; RED: reduced group of radiomics features; OS: oversampled group maintaining the original radiomics features; OS_R: oversampled group reducing the number of radiomics features. SMOTE: Synthetic Minority Over-sampling TEchnique. BW: Bin width.....	77
Figure 49. Pipeline of the radiomics image processing scheme for computing radiomics features. IH: intensity histogram; IVH: intensity-volume histogram; GLDZM: Grey Level Distance Zone Matrix [167].	78

Figure 50. Example of both discretization methods in two $[^{18}\text{F}]$ FDG PET images with a SUVmax of 27.8 and 6.6 for image A and B, respectively [170].	80
Figure 51. Example of the GLCM matrix calculation. (a) Representation of the different intensity values of an image, rounded in red the combination of intensity values 2 and 3, with a direction of 0° and distance of 1 pixel and in green the combination of the same intensity values with a direction of 135° and distance of 1 pixel; (b) GLCM for the pair of direction 0° and distance of 1 pixel, marked in red the number of repetitions for the intensity values 2 and 3, rounded in red in (a); (c) GLCM for the pair of direction 135° and distance of 1 pixel, marked in green the number of repetitions for the intensity values 2 and 3, rounded in green in (a). Adapted from [179].	85
Figure 52. Pipeline of univariate analysis for clinical contexts with only two classes. BH: Benjamini-Hochberg.	86
Figure 53. Pipeline of univariate analysis for clinical contexts with three or more classes. BH: Benjamini-Hochberg.	87
Figure 54. Summary of groups creation from each of the radiomics extraction, i.e. for each BW. The square for reduced features is obtained only once (from the ORIG Train Set), and then, it appears in other parts of the scheme representing that, although no other univariate analysis is applied, the same features removal is applied.	89
Figure 55. Cross-validation illustration adapted from [211]. (Grey) unassigned to any process (training, validation or test) data; (green) training data; (yellow) validation data; and (blue) test data.	93
Figure 56. Pipeline of the quantification based on DL models. DL models square content three variants of the same model, based only on its input.	95
Figure 57. Model of a neuron [213]. X: inputs of the neuron; w: synaptic weights for each input; v_k : linear combination of the inputs; y_k : output; b_k : bias added to the v_k .	98
Figure 58. Overall architecture of the CNN with its respective layers: input, convolutional, max-pooling and classification [213].	98
Figure 59. Representation of (a) sigmoid and (b) ReLu activation functions [213].	99
Figure 60. Representation of the average (top) and max (right) pooling operations [213].	99
Figure 61. AlexNet architecture [213,217]. Conv.: convolutional layer; MXP: max pooling layer; LRN: local response normalization; FC: fully connected. In the bottom part, the number of filters of each of the feature maps are shown.	100

Figure 62. VGG architecture of the features extractor [213,218]. Conv.: convolutional layer; FC: fully connected.	101
Figure 63. Residual block representation [219].	101
Figure 64. Example of a ResNet architecture [216].	102
Figure 65. Dense block representation [221]. BN: Batch Normalization.	103
Figure 66. Example of DenseNet architecture representation with 3 dense blocks [221].	103
Figure 67. 3D_RPET_NET architecture [222].	103
Figure 68. Representation of the confusion matrix for (a) a binary classification; and (b) a classification of three classes, evaluating the class 1.	106
Figure 69. ‘Pre-processing’ tab of the semi-automatic segmentation platform: initiation.	112
Figure 70. ‘Pre-processing’ tab of the semi-automatic segmentation platform: CT processing.	113
Figure 71. Slices location I: first approximation for (a) the lower slice, i.e. close to the femurs, and (b) the upper slice, i.e. close to the head, showing the binarised CT (blue) over the thresholded CT scan between 0 and 300 HU for a better visualization.	114
Figure 72. Slices location II: last slice location. Two examples of (a and c) location of the last slice including the chin, and (b and d) slice directly below to (a) and (c) respectively, showing the binarised CT (blue) over the thresholded CT scan between 0 and 300 HU for a better visualization.	114
Figure 73. Slices location III: first slice location. Examples of (a) pelvis location as the largest area, (b) first slice for CT starting from femur and (c) first slice for CT of whole-body protocol, showing the binarised CT (blue) over the thresholded CT scan between 0 and 300 HU for a better visualization.	115
Figure 74. Slices location IV: pelvis slice wrongly located. Examples of (a) pelvis location as the greatest area mistaken by a high HU in some internal organs, (b) slice defined as first slice due to the mistake in (a), showing the binarised CT (blue) over the CT thresholded between 0 and 300 HU for a better visualization.	116
Figure 75. ‘Pre-processing’ tab of the semi-automatic segmentation platform: slices selection.	116
Figure 76. ‘Segmentation’ tab of the semi-automatic segmentation platform: initiation.	117

Figure 77. ‘Editing’ tab of the semi-automatic segmentation platform: initiation.....	118
Figure 78. ‘Editing’ tab of the semi-automatic segmentation platform: saving information.	119
Figure 79. ‘Editing’ tab of the semi-automatic segmentation platform: Removing masks emergent window.	119
Figure 80. ‘Editing’ tab of the semi-automatic segmentation platform: including SUVmap.	120
Figure 81. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Visualization settings’ window.	121
Figure 82. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Zoom’ emergent window.....	121
Figure 83. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Volume’ emergent window.....	122
Figure 84. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Draw ROI’ and ‘Edit ROI’ functions.	123
Figure 85. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Extra Functions Window’.....	123
Figure 86. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Volume’ window for a concrete region selected in the ‘Extra Functions Window’.	124
Figure 87. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Thresholding’ window.....	125
Figure 88. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Hide superior panels’ button.	126
Figure 89. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Editing tab instructions’.....	127
Figure 90. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Keyboard functions instructions’.....	127
Figure 91. ‘Editing’ tab of the semi-automatic segmentation platform: thresholding example. Purple: mask; orange: ROI created in the thresholding process.	129
Figure 92. Examples of a satisfactory bone segmentation based on traditional algorithms: (a and b) pelvis region and (c) ribs region, (d), (e) and (f) are zoom of (a), (b) and (c), respectively. (Green) true positive pixels; and (yellow) false positive pixels.	130

Figure 93. Examples of a satisfactory bone segmentation based on traditional algorithms: (a) lower torso, (b) mid-chest torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.	131
Figure 94. Examples of a satisfactory BM segmentation based on traditional algorithms: (a) pelvis, (b) lower torso and (c) shoulders region, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.	132
Figure 95. Examples of an unacceptable BM segmentation based on traditional algorithms: (a) pelvis, (b) lower torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.	133
Figure 96. Comparison of IoU and Dice metrics for different combination of parameters F, E and training order. F: number of channels in each feature map; E: number of epochs in the training process.	134
Figure 97. Comparison of training orders for each bone segmentation model, i.e., different combination of F parameter and number of epochs (E). NC: Non-Convergent. In lighting green are shown the best result for each pair of feature models: the best Dice Index; if same value of Dice Index, the best Dice (mask) Index; if same value of Dice (mask) Index, the model with higher value of TP. ..	135
Figure 98. Examples of a satisfactory bone segmentation based on DL model: (a) pelvis, (b) mid-chest torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.....	136
Figure 99. Examples of main mistakes of bone segmentation based on DL model: (a) misleading areas in pelvis region, (b) missing areas inside correctly segmented structures in upper torso and (c) missing structures in upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.	137
Figure 100. Examples of a satisfactory BM segmentation based on DL model: (a) pelvis, (b) mid-chest torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.....	140
Figure 101. Examples of the main mistakes of BM segmentations based on DL model: (a) pelvis, (b) mid-chest torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.	141

Figure 102. Boxplot for Low Grey Level Zone Emphasis for IMWG labelling of SMM patients, differing between low-, intermediate- and high-risk of progression to MM.	163
Figure 103. Boxplot for PETHEMA labelling of SMM patients, differing between low-, intermediate- and high-risk of progression to MM, for radiomics features: (a) DependenceNonUniformityNormalized, (b) LowGrayLevelZoneEmphasis and (c) SmallAreaLowGrayLevelEmphasis.....	166
Figure 104. Outline of possible future lines of action based on the limitations encountered.	200
Figure 105. Instructions of the ‘Pre-processing’ tab: I.....	239
Figure 106. Instructions of the ‘Pre-processing’ tab: II.	240
Figure 107. Instructions of the ‘Pre-processing’ tab: III.	240
Figure 108. Instructions of the ‘Segmentation’ tab: I.....	241
Figure 109. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Volume’ window with PET overlaid.....	242
Figure 110. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Information’ window with a warning message, result of selecting (a) a lower than 1 for the slices and (b) a higher than the maximum CT and SUVmap slices.....	243
Figure 111. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Extra Functions Window’, visualization of slices function.	243
Figure 112. ‘Editing’ tab of the semi-automatic segmentation platform: main window instructions.	244
Figure 113. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Visualization settings instructions’ window.....	245
Figure 114. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Extra Functions Window’ instructions.	245
Figure 115. Comparison of training orders for each model, i.e., different pairs of F and E in unprocessed database. NC: Non-Convergent. In lighting green are shown the best result for each pair of feature models: the best Dice Index; a same value of Dice Index, the best Dice (mask) Index; a same value of Dice (mask) Index, the model with higher value of TP.	246
Figure 116. Results of U-Net architecture trained for the stretcher location, varying the number of channels or filters in the first feature map (F) and the number of epochs. In green is marked the model with the best performance.	247

List of Tables

Table 1. Proposed definitions for normo-cellularity for patient age groups [8].	11
Table 2. Resume of all studies for bone segmentation. (-) not quantitatively evaluated. Studies marked with * obtained an automatic segmentation.	31
Table 3. Resume of all studies for bone segmentation. (-) not quantitatively evaluated; (--) not quantitatively evaluated and mean for another objective.	34
Table 4. Resume of all studies for BM quantification in MM field.	39
Table 5. Resume of the features used in the studies of Table 4. Martínez-Martínez <i>et al.</i> [84] is not included since it used bone density as the only feature. Wennmann <i>et al.</i> [113], is not included since they mentioned only radiomics, no groups. A: SMM to MM; B: MM vs. Metastases; C: MM prognosis (survival); D: BM infiltration; E: Pre and post treatment comparison. GLSM: Grey Level Scale Matrix.	40
Table 6. Resume of all studies for quantification in NHL field. The first column indicates the aim of the study, the second one indicated the main methodology applied. No image modality is mentioned since all the studies use [¹⁸ F]FDG PET/CT imaging. S: sensitivity, Sp: specificity, PPV: positive predictive value, NPV: negative predictive value.	45
Table 7. Confusion matrix for the pGT labelling definition. X: no MFC test available.	75
Table 8. Example of discretised intensity values for fixed bin size discretization. N: intensity values.	81
Table 9. Representation of the layers of the convolutional block for the pRGB adaptation.	96
Table 10. Division of train and test sets in each training step for both, bone and BM segmentations. N: number of patients, i.e. CT volumes; S: number of samples, i.e. slices of the volumes.	110
Table 11. Results of BM segmentation models, varying the approaches, image collections and training phase. Approaches based on: (I and II) untrained U-Net model using as input (I) the pre-processed CT scan; and (II) the pre-processed CT scan in the regions of the bone mask; (III) U-Net model trained for the bone segmentation using as input the CT pre-processed scan, as (I). Image collections: (A and B) same cohort used for the bone segmentation model and (C) larger cohort than the one used for the bone segmentation model, using as bone masks (A) those manually segmented for being the GT of the bone segmentation model, and (B and C) those obtained from the bone segmentation model. NC: non-	

convergent, X: not applicable since the previous model is NC. The model with the best performance is marked in green.....	139
Table 12. Summary of the segmentation results for both, bone and BM. TRAD: segmentation based on traditional algorithms; BM I: test set for phase I of DL training, with the 13 manually segmented masks; BM II: test set for phase I of DL training, with 7 out of the 13 manually segmented masks; Bg: background...	143
Table 13. Confusion matrix for PET visual assessment and MFC assessment of MRD. X: no biopsy available.....	149
Table 14. Performance metrics for the ML model with the best F1_macro in MRD detection. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted. Outstanding values, i.e. above 0.85, are marked in bold.	150
Table 15. BW and model type for the ML model with the best F1_macro in MRD detection.....	151
Table 16. Performance metrics for CNNs architectures with the best F1_macro in MRD detection. DA: Data Augmentation; S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted. Outstanding values, i.e. above 0.85, are marked in bold.	152
Table 17. Architecture, F value and class weights for the CNN with the best F1_macro in MRD detection. DA: data augmentation. DA column is marked in green for combinations which obtains an acceptable value of F1_macro.	154
Table 18. Confusion matrix for PET visual assessment and MFC assessment of MRD. X: no biopsy available.....	158
Table 19. Performance metrics for the ML model with the best F1_macro in MRD prediction. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted. Outstanding values, i.e. above 0.85, are marked in bold.	159
Table 20. Bin Width (BW) and model type for the ML model with the best F1_macro in MRD detection.....	160
Table 21. Confusion matrix for IMWG assessment and PETHEMA assessment of risk of progression to MM in SMM patients.	162
Table 22. P-value lower than 0.05 for ANOVA or Kruskal-Wallis tests for univariate analysis to assess the risk of progression to MM in patients with SMM in IMWG labelling. P-value corrected by Benjamini-Hochberg correction is not shown, as its values is higher than 0.05. P-values of post hoc analysis for each pair of classes are also included, shown in bold those below 0.05.	162
Table 23. P-values lower than 0.05 (shown in bold type), for any of the bin widths, obtained in ANOVA or Kruskal-Wallis tests for univariate analysis to assess the	

risk of progression to MM in patients with SMM in PETHEMA labelling. P-values corrected by Benjamini-Hochberg correction are not shown, as none obtain values which means significant differences between classes.....164

Table 24. P-values of post hoc analysis for each pair of classes for variables shown in Table 23. P-values below 0.05 are shown in bold. Gr: group of features. 165

Table 25. Confusion matrix comparing PET and MFC labelling. X: no MFC test available. 169

Table 26. Sensitivity, precision, F1(BMI-): F1_score for BMI- class; F1(BMI+): F1_score for BMI+ class; F1(rBM): F1_score for rBM class; F1_m: F1_score_macro, calculated as the mean F1_score of the three previous metrics; F1_m*: F1_score_macro*, calculated as the mean F1_score considered as binary problem for any class; F1_w: F1_score_weighted; all of them using PET labelling as predicted and MFC or pGT labelling as GT. 169

Table 27. Performance metrics for the ML model with the best F1_macro (upper) and F1_weighted (lower) for BM assessment in NHL patients. Green cells represent coincidence between the model with the best F1_macro* and the best F1_weighted, excluding the metrics from F1_w set to avoid redundant information. Outstanding values, i.e. above 0.85, are marked in bold. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted; X: indeterminate number resulting from dividing by 0.....171

Table 28. Performance metrics for the ML model with the best F1_macro (upper) and F1_weighted (lower) for BM assessment in FL patients. Green cells represent coincidence between the model with the best F1_macro* and the best F1_weighted, excluding the metrics from F1_w set to avoid redundant information. Outstanding values, i.e. above 0.85, are marked in bold. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted; X: indeterminate number resulting from dividing by 0.....172

Table 29. Performance metrics for the ML model with the best F1_macro (upper) and F1_weighted (lower) for BM assessment in DLBCL patients. Green cells represent coincidence between the model with the best F1_macro* and the best F1_weighted, excluding the metrics from F1_w set to avoid redundant information. Outstanding values, i.e. above 0.85, are marked in bold. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted; X: indeterminate number resulting from dividing by 0.....173

Table 30. Bin Width (BW) and model type for the ML model with the best F1_macro (upper) and F1_weighted (lower) for BM assessment in NHL, FL and DLBCL. Grey cells represent the lack of those groups, i.e., no significant features

after reduction. Green cells represent coincidence between the model with the best F1_macro* and the best F1_weighed. 175

Table 31. Performance metrics for CNNs architectures with the best F1_macro* (upper) and F1_weighted (lower) in BM status classification. DA: Data Augmentation; S: sensitivity; Pr: precision; F1_m: F1_macro; F1_m*: F1_macro* F1_w: F1_weighted; X: indeterminate number resulting from dividing by 0. Outstanding values, i.e. above 0.85, are marked in bold. Green cells represent coincidence between the model with the best F1_m* and the best F1_w. 177

Table 32. Architecture, F value and class weights for the CNN with the best F1_macro in NHL patients for BM status classification. DA: data augmentation. DA column is marked in green for combinations which obtains an acceptable value of F1_macro. 179

Table 33. Performance metrics for the ML model for the PET labelling in MRD detection. Outstanding values, i.e. higher than 0.85, are marked in green. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 249

Table 34. Performance metrics for the ML model for the MFC labelling in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 250

Table 35. Performance metrics for the ML model for the pGT labelling in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 250

Table 36. Performance metrics for CNNs architectures for PET labelling and 3D input in MRD detection. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted. 251

Table 37. Performance metrics for CNNs architectures for PET labelling and pRGB input in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 252

Table 38. Performance metrics for CNNs architectures for PET labelling and MIP input in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked

in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.....	253
Table 39. Performance metrics for CNNs architectures for MFC labelling and 3D input in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.....	253
Table 40. Performance metrics for CNNs architectures for MFC labelling and pRGB input in MRD detection. Outstanding values, i.e. above 0.85, are marked in green. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.....	254
Table 41. Performance metrics for CNNs architectures for MFC labelling and MIP input in MRD detection. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.	255
Table 42. Performance metrics for CNNs architectures for pGT labelling and 3D input in MRD detection. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.....	255
Table 43. Performance metrics for CNNs architectures for pGT labelling and pRGB input in MRD detection. Outstanding values, i.e. above 0.85, are marked in green. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.....	256
Table 44. Performance metrics for CNNs architectures for pGT labelling and MIP input in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.....	257
Table 45. Performance metrics for the ML model for the PET labelling in MRD prediction. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.....	258
Table 46. Performance metrics for the ML model for the PET labelling in MRD prediction. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.....	259

Table 47. P-values for ANOVA or Kruskal-Wallis tests for univariate analysis to assess the risk of progression to MM in patients with SMM in IMGW labelling. Variables with a p-value lower than 0.05, i.e. with significant differences between classes, for any of the bin widths are marked in green and bold type. P-values corrected by Benjamini-Hochberg correction are not shown, as none obtain values which means significant differences between classes. Gr: group of features. 263

Table 48. P-values for ANOVA or Kruskal-Wallis tests for univariate analysis to assess the risk of progression to MM in patients with SMM in PETHEMA labelling. Variables with a p-value lower than 0.05, i.e. with significant differences between classes, for any of the bin widths are marked in green and bold type. P-values corrected by Benjamini-Hochberg correction are not shown, as none obtain values which means significant differences between classes. Gr: group of features. 266

Table 49. Performance metrics for the ML model for the PET labelling in NHL patients for BM classification. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 268

Table 50. Performance metrics for the ML model for the MFC labelling in NHL patients for BM classification. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 269

Table 51. Performance metrics for the ML model for the pGT labelling in NHL patients for BM classification. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 270

Table 52. Performance metrics for the ML model for the PET labelling in FL patients for BM classification. Outstanding values, i.e. higher than 0.85, are marked in green. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 271

Table 53. Performance metrics for the ML model for the MFC labelling in FL patients for BM classification. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 273

Table 54. Performance metrics for the ML model for the pGT labelling in FL patients for BM classification. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 274

Table 55. Performance metrics for the ML model for the PET labelling in DLBCL patients for BM classification. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 276

Table 56. Performance metrics for the ML model for the MFC labelling in DLBCL patients for BM classification. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 276

Table 57. Performance metrics for the ML model for the pGT labelling in DLBCL patients for BM classification. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. 278

Table 58. Performance metrics for CNNs architectures for PET labelling and 3D input for BM status classification in NHL patients. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted. 279

Table 59. Performance metrics for CNNs architectures for PET labelling and pRGB input for BM status classification in NHL patients. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted. 280

Table 60. Performance metrics for CNNs architectures for PET labelling and MIP input for BM status classification in NHL patients. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted. 281

Table 61. Performance metrics for CNNs architectures for MFC labelling and 3D input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted. 281

Table 62. Performance metrics for CNNs architectures for MFC labelling and pRGB input for BM status classification in NHL patients. The maximum value

for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.....282

Table 63. Performance metrics for CNNs architectures for MFC labelling and MIP input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted. 283

Table 64. Performance metrics for CNNs architectures for pGT labelling and PET input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted. 284

Table 65. Performance metrics for CNNs architectures for pGT labelling and pRGB input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.....284

Table 66. Performance metrics for CNNs architectures for pGT labelling and MIP input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted. 285

Abbreviations and Acronyms

A	[¹⁸ F]FDG	Fluorine-18 fluorodeoxyglucose
	2D	Two-dimensional
	3D	Three-dimensional
	AUC	Area Under the Curve
B	BM	Bone Marrow
	BMI	Bone Marrow Infiltration
	BW	Bin Width
C	CNN	Convolutional Neural Network
	CR	Complete response
	CT	Computerized Tomography
	CTB	Centre for Biomedical Technology
D	DenseNet	Densely Connected Network
	DL	Deep Learning
	DLBCL	Diffuse Large B-cell Lymphoma
	DT	Decision Tree
E	EANM	European Association Nuclear Medicine
	ERM	<i>Enfermedad Residual Medible</i>
	ETSIT	<i>Escuela Técnica Superior de Ingenieros de Telecomunicación</i>
F	F1 _i	F1_score for class i, being i = 0,1,2...
	F1 _{macro}	F1_score_macro
	F1 _{weighted}	F1_score_weighted
	FL	Follicular Lymphoma

G	FN	False Negative
	FP	False Positive
	GB	Gradient Boosting
	GBT	<i>Grupo de Bioingeniería y Telemedicina</i>
	GEM	<i>Grupo Español de Mieloma</i>
	GLCM	Grey Level Co-occurrence Matrix
	GLDM	Grey-Level Dependence Matrix
	GLRLM	Grey Level Run Length Matrix
	GLSZM	Grey Level Size Zone Matrix
H	GT	Ground truth
	HL	Hodgkin's Lymphoma
	HU	Hounsfield Units
I	IBI	Intensity Bone Involvement
	IBSI	Image Biomarker Standardization Initiative
	ICML	International Conference on Malignant Lymphomas
	Ig	Immunoglobulin
	IMO	<i>Infiltración de médula ósea</i>
	IMWG	International Myeloma Working Group
	IoU	Intersection over Union
K	k-NN	k-Nearest Neighbours
L	LDCBG	<i>Linfoma Difuso de Células-B Grandes</i>
	LF	<i>Linfoma Folicular</i>
	LIN	Linear kernel
	LNH	<i>Linfoma No-Hodgkin</i>
	LR	Logistic Regression
M	MFC	Multiparametric Flow Cytometry

N	MGUS	Monoclonal Gammopathy of Undetermined Significance
	MIP	Maximum Intensity Projection
	ML	Machine Learning
	MLP	Multi-Layer Perceptron
	MM	Multiple Myeloma/ <i>Mieloma múltiple</i>
	MO	<i>Médula ósea</i>
	MPCD	Monoclonal Plasma Cell Disorder
	MRD	Measurable Residual Disease
	MRI	Magnetic Resonance Imaging
	MTV	Metabolic Tumour Volume
O	NB	Naïve Bayes
	NDC	Nucleated differential cell count
	NGS	Next-generation sequencing
	NGTDM	Neighbouring Gray Tone Difference Matrix
P	NHL	Non-Hodgkin's Lymphoma
	ORIG	Original radiomics features
	OS	Oversampled radiomics features
R	OS_R	Oversampled and reduced radiomics features
	PBI	Percentage of Bone Involvement
	PCA	Principal Component Analysis
	PET	Positron Emission Tomography
	pGT	pseudo Ground Truth
	Pr	Precision
	pRGB	pseudo RGB
R	RBF	Radial basis function kernel
	rBM	Reactive bone marrow
	REAL	Revised European-American Lymphoma
	RECIST1.1	Response Evaluation Criteria In Solid Tumors 1.1

S	RED	Reduced radiomics features
	ReLU	Rectified Linear Unit
	ResNet	Residual Network
	RF	Random Forest
	ROC	Receiver Operating Curve
	ROI	Region of interest
	S	Sensitivity/Recall
	SIG	Sigmoidal kernel
	SMM	Smouldering Multiple Myeloma
	SMOTE	Synthetic Minority Over-sampling TEchnique
T	SPM12	Statistical Parametric Mapping
	SUV	Standardized Uptake Value
	SUVmax	Maximum SUV
	SUVmean	Mean value of SUV
U	SVM	Support Vector Machines
	TN	True Negatives
	TP	True Positive
V	UPM	Universidad Politécnica de Madrid
	VGG	Visual Geometry Group
	VOI	Volume of interest

1. Introduction

1.1. Bone marrow bases

Cortex and medulla are the main components of bones. Cortex, also known as compact bone, is a strong layer which protects the medulla, cancellous bone or spongy bone, and the medullary cavity which contains the **bone marrow (BM)**. BM is divided in red marrow, composed by hematopoietic cells; and yellow marrow, which is formed by adipose tissue. The anatomical distribution of these tissues is shown in Figure 1. The different proportion of both subtypes of marrow defines the BM cellularity, calculated as the relative percentage of red marrow compared to yellow marrow. The cellularity is strongly related to the process of blood and immune cells production, known as **haematopoiesis**, and the higher the cellularity, the higher the hematopoietic activity. This way, red marrow is the main organ in charge of the haematopoiesis process during extra-uterine life. Then the hematopoietic marrow distribution changes, being reduced with age: in newborns, the red marrow is the most extended, including even the phalanges; in children, hematopoietic marrow starts to be replaced by adipose marrow; in early adults, it is confined to skull, vertebrae, ribs, clavicles, sternum, pelvis and the proximal half of the humeri and femurs. This reduction is characterised by the reduction of hematopoietic tissue and bone volume, and the increment of adipose tissue [1–3].

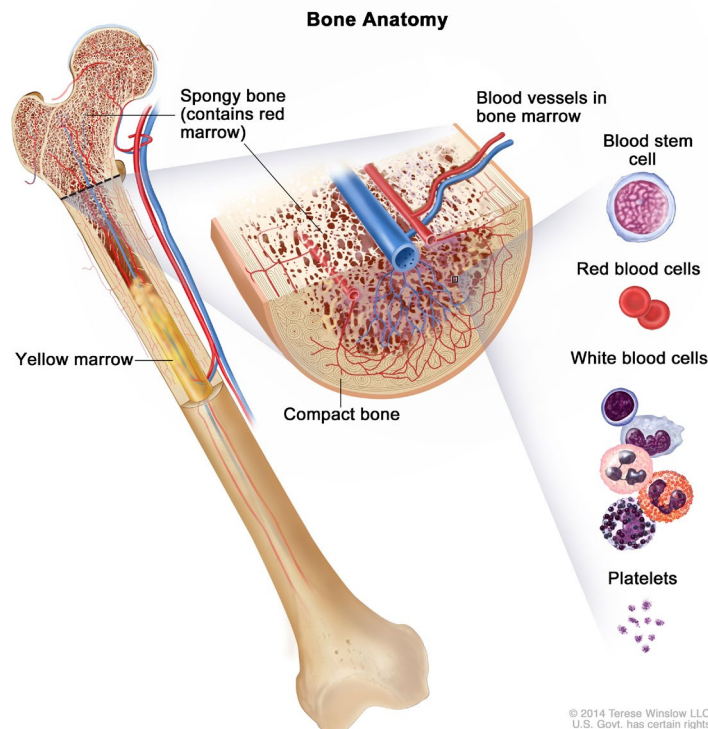


Figure 1. Bone anatomy representation [4].

1.1.1. Haematopoiesis

As mentioned before, haematopoiesis is a process characterised by the continuous production of blood and immune cells in the BM. Only a small proportion of nucleated BM cells are hematopoietic stem cells ($<0.1\%$) [2,5]. During the haematopoiesis, stem or self-renewing cells are converted into the different types of blood cells. This way, for each stem cell, a mature differentiated cell and another self-renewed stem cell are produced [6]. The mature cell produced can be differentiated in **myeloid** (ascendant of red blood cells, monocyte lineages and neutrophils) and **lymphoid** (ascendant of T-cell lineages and B-cell lineages) progenitor cells, being finally differentiated as a blood or immune cell [3,5,7,8].

Despite the BM is the main organ in the hematopoietic process, the immune system is composed of more organs: lymph nodes, spleen, thymus, liver, gastrointestinal tract, and the upper and lower respiratory tracts. All these organs and tissues are communicated by lymphatic channels and by the blood stream [5,9].

Thus, the spleen is the largest lymphoid organ in charge of blood filtering. Its functions include: bacteria and abnormal cells phagocytosis, protection against micro-organisms, and antibodies and red cells production. On the other hand, the liver plays an important role in the first-line immune defence, due to its lymphocyte population enriched in natural killers [9,10].

1.2. Bone marrow pathological affectations

When BM is affected, the haematopoiesis process is compromised, affecting the blood and immune cells production. This cells reduction directly increase the possibilities of suffering other spontaneous diseases such as infections or inflammations. In addition, the neurological problems consequence of the skeletal pain and instability are a cause of morbidity [11]. The worst-case scenario is end-organ damage, and even death. Some plasma cell and lymphoid neoplasms are two of the main pathologies which can affect the BM at these levels.

1.2.1. Plasma cell neoplasms

In spite of deriving from terminally differentiated B-cells, plasma cell neoplasms are characterised by the presence of plasma cells without a B-cell component. In plasma cell disorders, three main groups are differentiated: **Multiple Myeloma (MM)**, the most common disease with BM affectation; localised plasma cell tumours, characterised by the lack of BM infiltration; and disorders characterised by an abnormal immunoglobulin deposition [12,13].

1.2.1.1. Multiple Myeloma

The proliferation of neoplastic plasma cells in the BM is a heterogeneous disease known as MM [14]. MM is a hematologic cancer which corresponds to the 10% of all hematologic tumours, with 187,952 new cases (0.9% of all cancers) and 121,388 new deaths (1.2% of all cancers) in 2020 [13,15–20]. This disease appears predominantly in the middle-aged and elderly people, with a median age of 69 years. Clinical consequences of MM appear as a direct effect of the neoplastic proliferation or indirectly, from the effects caused by the myeloma cells secreted paraprotein. The most typical secreted paraproteins are immunoglobulin (Ig) in 60% of cases; IgA in about 20%; and monoclonal light chain in 15-20% of cases. IgM, IgD or IgE are produced in a minority of patients [1,16]. The indirect and direct effects are caused by four factors [2]:

- I. osteoclast activation and osteoblast suppression with the consequent **bone destruction**;
- II. blood cell deficiencies due to the replacement of hematopoietic marrow by neoplastic plasma cells and the reduction of erythropoietin caused by **renal damage**;
- III. suppression of normal polyclonal plasma cells which causes an **immunoglobulin deficiency**;
- IV. and, the overproduction of monoclonal immunoglobulins and its consequence, **the renal damage**.

The most common symptoms presented by these factors are anaemia, bone pain, fractures or lesions (osteoporosis, osteolytic lesions or osteosclerosis), hypercalcaemia, renal failure and BM involvement. Less common symptoms are hepatomegaly, lymphadenopathy, or splenomegaly [1,3]. The **International Myeloma Working Group (IMWG)** stated that MM is certainly diagnosed if two of the following criteria are present [1,13,16,21–23]:

- I. more than 10% plasma cells in a BM biopsy;
- II. a paraprotein in serum or urine;
- III. evidence of **end-organ damage** attributed to the underlying plasma cell proliferative disorder (at least one): hypercalcaemia, renal insufficiency, anaemia, or osteolytic bone lesions or osteoporosis (one or more sites of osteolytic destruction).

1.2.1.2. Monoclonal Gammopathy of Undetermined Significance

Although observation of two of the mentioned symptoms is necessary for MM diagnosis, it is typically preceded by an asymptomatic stage. This condition is known as **Monoclonal Gammopathy of Undetermined Significance (MGUS)**, and it is presented in 3-4% of the population over the age of 50 years [2,13]. MGUS patients are characterised by being **asymptomatic** and lacking laboratory and radiological features of MM, except for the presence of paraproteins in serum and/or urine [2,9]. In spite of being different conditions, 27% of patients with MGUS develop MM and 80% of MM patients are originated from MGUS. Consequently, MGUS is defined as a precursor, since it increases the **risk of progression to MM** [2,3,13,22]. The progression to MM or other related conditions (such as lymphoma) is about 12%, 21% and 26% by 10, 20 and 25 years respectively [17]. The IMWG stated that MGUS is certainly diagnosed if all three of the following criteria are met [13,22]:

- I. less than 10% plasma cells in a BM;
- II. less than 3gm/dl of monoclonal protein in serum;
- III. absence of end-organ damage which can be attributed to a plasma cell proliferative disorder.

1.2.1.3. Smouldering Multiple Myeloma

An intermediate clinical stage between MGUS and MM is known as **Smouldering MM (SMM)**. The main difference between both states is the risk of progression to MM per year in the first 5 years after diagnosis, being much higher in SMM (10%) compared to MGUS (0.5-1%) [3,22]. SMM is a plasma cell disorder characterised by MM symptoms, monoclonal protein in serum and/or BM plasma cells, but with no evidence of other MM symptoms: high serum calcium concentration, renal failure or insufficiency, anaemia, or bone lesions [1,24]. The IMWG [13] considers SMM as an intermediate clinical stage between MGUS and MM, accounting a 14% of all patients with MM. Concretely, SMM is certainly diagnosed if both criteria are met [22]:

- I. more than 3gm/dl of monoclonal protein IgG or IgA in serum and/or more than 10% plasma cells in a BM;
- II. absence of end-organ damage which can be attributed to a plasma cell proliferative disorder.

SMM patients are those with the best prognosis in the MM group [1]. However, the prognosis depends on the rate of progression. In order to differentiate between different risk groups, the Spanish Myeloma Working Group (GEM, *Grupo Español de Mieloma*) [25,26] and the IMWG [13,24] define two high-risk factors: the percentage of BM plasma cells and immunoparesis (reduction of 1 or 2

immunoglobulins), with the cut-off values differing for each group. Patients can be divided into three different groups depending on these factors [13,24,25]:

- I. **Low risk**, 4% risk of progression at 5 years. Patients are characterised by absence of high-risk factors, i.e., they behave similarly to MGUS. Annual revisions are recommended. This group accounts for 50% of SMM patients.
- II. **Intermediate risk**, 46% risk of progression at 5 years. Patients are characterised by only some of the high-risk factors. Six-month revisions are recommended.
- III. **High risk**, 50% risk of progression in less than 2 years from diagnosis (median 15 months) or 72% at 5 years. Patients are characterised by a 60% or greater plasma cells involvement in BM examination [27–29]. Two- or three-month revisions are recommended.

No test or feature has been demonstrated to be useful in distinguishing those with clonal premalignant plasma cell (similar to MGUS, low risk of progression) from those with clonal malignant myeloma cells (high risk of progression). The lack of this measurement obstructs the treatment strategy, since early therapy for SMM can improve the overall survival but can lead to overtreatment in high-risk patients [13].

1.2.1.4. Measurable Residual Disease

Taking into account these three states: MM, MGUS and SMM, the optimal therapeutic strategy is defined in each case. Thanks to the new and advanced therapies, in the last decades, the number of MM patients that achieve a deep response has increased, even achieving a complete response (CR) [30]. However, and although a promising initial response, most of MM patients with CR are incapable of experiencing a survival benefit due to a persistent **Measurable Residual Disease (MRD)**, also known as Minimal Residual Disease [31,32].

MRD is a MM state characterised by sub-clinical levels of **tumour burden**. These remnant cells continue to proliferate after treatment, causing relapse [33]. Even though this condition is usually not detected in serum or urine protein evaluation, it is detectable using techniques with higher sensitivity, such as BM biopsy or aspiration (posteriorly described in subsection 1.3.1. *Haematologic assessment*) [34]. These sensitive techniques are used for estimating the level of MRD, which is considered a progression-free and overall survival marker since it is inversely related with the depth of response and directly associated with the risk of relapse [31]. Consequently, the efficacy of the treatment at different stages can be evaluated using MRD as a biomarker [31,32]. For these reasons, the IMWG [35,36] defines MRD as a **prognostic biomarker** for identifying groups of MM patients with better survival outcomes.

1.2.2. Lymphoid neoplasms

Neoplasms of lymphoid cells (T, B or natural killers) and their precursors are known as **lymphomas**. These neoplasms are the consequence of multiple mutations in a single lymphoid cell. Lymphoma cells not only are characterised by the features of lymphoid cells at a particular stage differentiation, but they also share the characteristics of lymphocytes resident of a particular organ or tissue [9]. They are divided into two main groups: Hodgkin's Lymphoma (HL), accounting for about 10% of lymphoma cases, and **Non-Hodgkin's Lymphoma (NHL)**, accounting for the remaining 90% [37]. These groups are distinguished mainly by the presence and absence of Reed-Sternberg cells, respectively [38].

HL is a clonal B-cell origin lymphoproliferative disorder characterised by a small number of malignant cells in a reactive benign inflammatory background. It is commonly nodal, and the most frequent sites are cervical, axillary and inguinal nodes. BM is involved in only 5% of patients diagnosed with HL [3,38].

NHL is a pathology characterised for including a heterogeneous group of cancers. The vast majority of cases occurs at lymph nodes, although any tissue can be affected. Subgroups encompass from indolent subtypes to extremely aggressive subtypes. For this reason, an early and precise diagnosis is important to decide the most adequate therapy for each patient [37]. Figure 2 shows the imperative prevalence of B-NHL (with B-cells origin) against HL or T-NHL (with T-cells origin) [39]. HL accounts for 0.4% of new cases of all cancers (82,469) and 0.2% of new deaths of all cancers (23,376) in 2020 [18,19,40], while NHL accounts for 2.8% of new cases of all cancers (544,352) and 2.6% of new deaths of all cancers (259,793) also in 2020 [18,19,41].

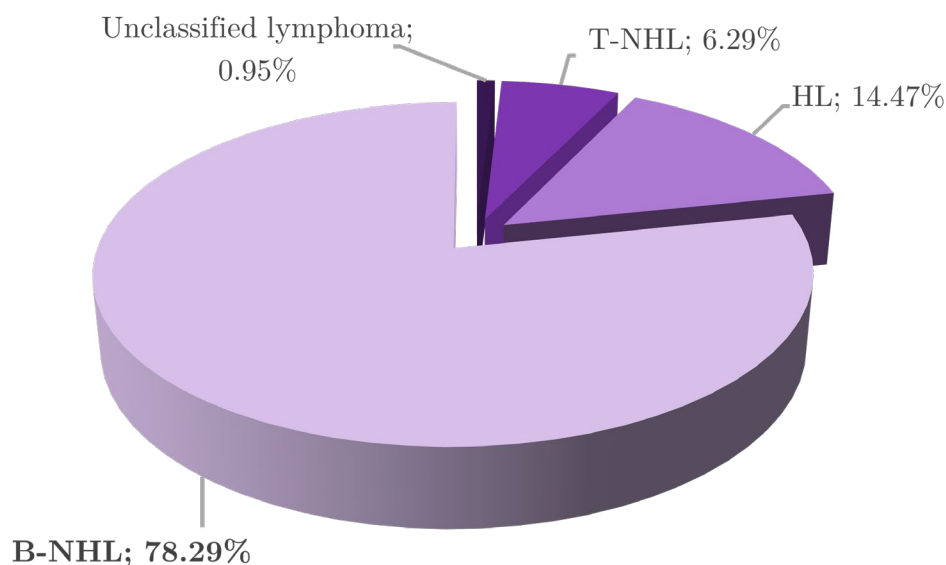


Figure 2. Distribution of B-cell, T-cell, and Hodgkin lymphomas among the nodal or extranodal lymphomas. Adapted from [39].

Figure 3 shows the relative frequency of prevalence for each subtype of B-cell lymphoma [1,42]. Considering nodal and extranodal lymphomas, **Follicular Lymphoma (FL)** and **Diffuse Large B-cell Lymphoma (DLBCL)** stand as the most common B-NHLs in the Western world, accounting for about 19.61% and 41.4% of B-NHL. Moreover, in adults, DLBCL is the most common lymphoid malignancy, conforming 31% of all NHL in developed countries [43].

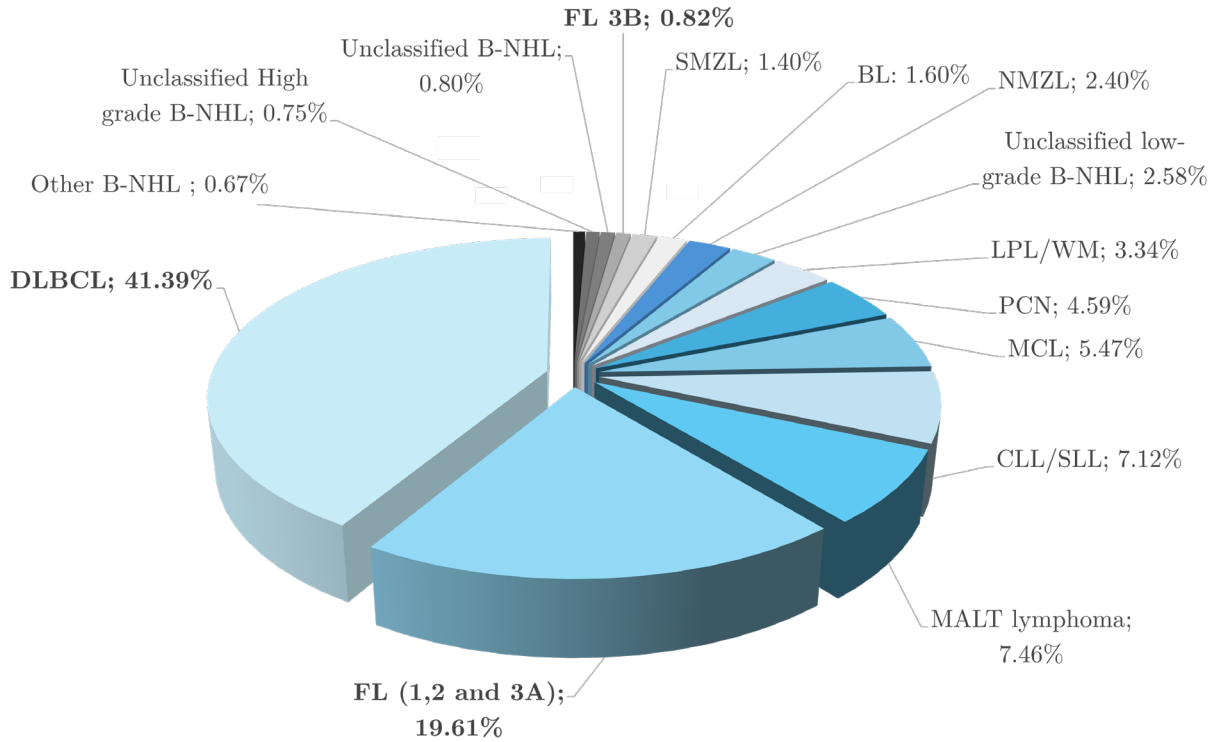


Figure 3. Relative frequencies of B-cell lymphomas. Adapted from [39].

1.2.2.1. Follicular lymphoma

FL, also known as follicular B-cell neoplasm, is an indolent disease characterised for being a tumour of germinal centre B-cells, centroblasts and centrocytes, with a follicular growth pattern arising in lymphatic tissues. Occasionally, the follicular growth pattern can be substituted by a diffuse pattern or a diffuse-follicular pattern. The clinical course is prolonged and, in some cases, even asymptomatic. However, FL is characterised for being a very variable disease. This variability can be observed in the form of multiple relapses or fast apparition of advanced stage. FL is unlikely during adolescence and childhood, and its incidence does not depend on gender [1,3,9,42].

Clinically, it presents localised or generalised lymphadenopathy. It involves multiple nodes and, unfrequently, the spleen. In spite of being often asymptomatic, patients with advanced disease can show B symptoms (fever, weight loss and night sweats), anaemia and thrombocytopenia [1,9,42]. FL staging is based on the relative number

of centroblasts, creating three grades: grade 1, 2 and 3, with <5 , 6-15 and >15 centroblasts per sample respectively. Grades 1 and 2 are considered low-grade, whereas grade 3 is divided in two different subgrades: 3A when not only centroblasts, but also centrocytes are present; and 3B when only centroblasts are present [3].

Due to its variability, **FL tends to transform into more aggressive lymphomas** (30% of cases, with a cumulative risk of 3% per year). If FL shows large blastic cells in diffuse areas, it can transform into DLBCL, typically reducing the prognosis [1,3,9,42].

1.2.2.2. Diffuse large B-cell lymphoma

In the REAL (Revised European-American Lymphoma) classification [44], DLBCL is defined as an entity with several morphological variants, all of them characterised by diffuse infiltration and large B-lineage lymphoma cells [1,9]. Regularly, quick growing tumour masses in single or multiple nodal or extranodal sites are presented, being extranodal representation in 40% of patients. In addition, this B-NHL subtype is extremely heterogeneous in response to therapy and prognosis, with a 60-70% of 5-year overall survival. Patients present B-symptoms such as fever or weight loss. Likewise, DLBCL is also divided into different subgroups depending on the cell-of-origin: Germinal Centre B-cell, characterised by a better prognosis; Activated B-cell; and Unclassified subtypes, accounting the 10-15% of DLBCL [45].

1.2.2.3. Bone marrow infiltration in lymphoproliferative disorders

Lymphoma disorders frequently trigger **BM Infiltration (BMI)** in 5-15% of HL and 20-40% of NHL patients. BMI in lymphoma indicates a stage IV, which has **therapeutic and prognostic implications**. For this reason, an accurate BM assessment is crucial in staging lymphoma, especially those patients considered early-stage before BM evaluation [46,47]. This involvement can be observed in three major patterns [1]:

- I. **Interstitial infiltration** characterised by the presence of individual neoplastic cells among haemopoietic and fat cells.
- II. **Focal infiltration** characterised by focal points of neoplastic cells separated by residual haemopoietic marrow. Focal infiltration can be observed in two different sub-patterns depending on its way of aggregation: trabecular infiltration, where the neoplastic cells are adjacent to the trabecular bone; and, random infiltration, where the aggregation is not characterised for its relationship with the trabecular bone.
- III. **Diffuse infiltration** characterised by a huge replacement of regular marrow, affecting the marrow architecture.

All these patterns and sub-pattern can be observed in Figure 4. In addition, mixed patterns, such as interstitial-diffuse, can be also observed in lymphoma patients [1].

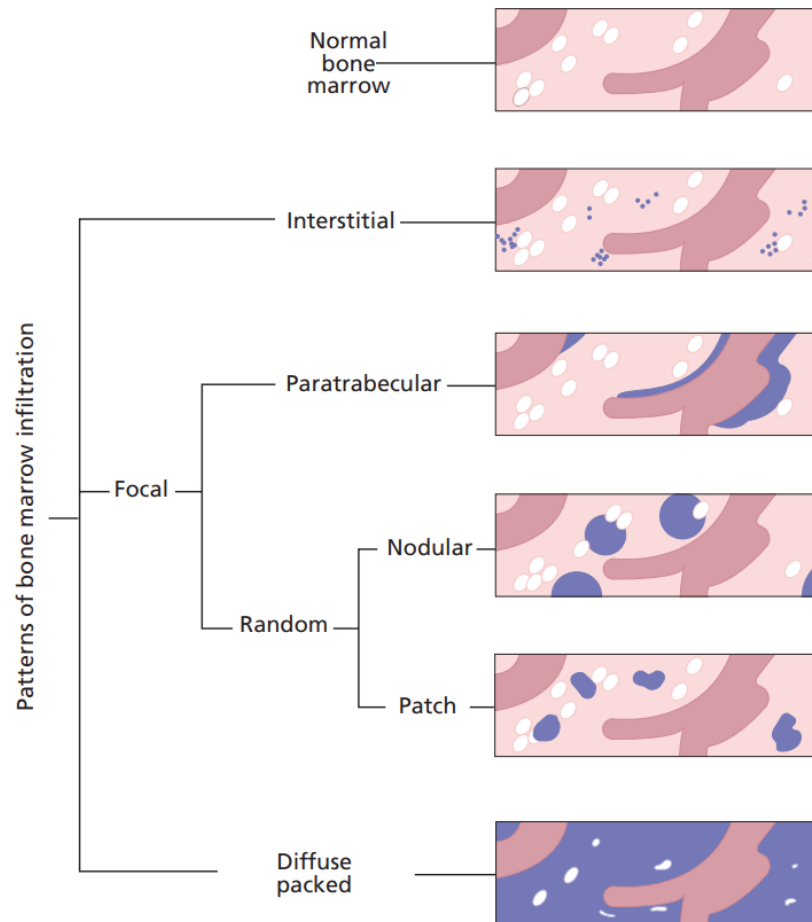


Figure 4. Patterns of BMI in lymphoma disorders [1].

Despite FL being a nodal disease, BMI (25-60% of patients), spleen or liver are usually encountered. The BMI in FL is normally focal and infrequently diffuse. A heavy BMI produces a reduction of the haemoglobin concentration and counts of platelet and neutrophils. In addition, it can combine focal lesions and replace large areas of marrow [1,9,42].

In large B-cell lymphomas, BMI is infrequently presented (10-20%) compared to the rest of B-cells neoplasm. However, patients with BMI shows a diffuse pattern of infiltration, related to a poorer prognosis [1,45]. In addition to BM diffuse uptake, many of lymphoma patients show an increment of spleen activity. The association of this increment of activity with inflammatory or lymphoma activation is difficult to determine [10].

1.2.2.4. **Reactive bone marrow**

Taking into account that BM is the main hematopoietic organ, it has multiple responses during infections, inflammations and similar conditions: (1) emergency myelopoiesis to produce large numbers of myeloid cells; (2) mobilization of the mature myeloid and hematopoietic stem cells and progenitors into the circulation, producing the migration of progenitor to the spleen and other differentiating organs; (3) hematopoietic injury, causing an inefficient haematopoiesis; (4) damage and remodelling of the BM. Due to the necessity of the immune cells overproduction, other immune organs, such as spleen and liver, increase their functionality [10,48].

1.3. Bone marrow assessment

1.3.1. **Haematologic assessment**

The most common techniques for BM evaluation are aspiration, trepanning and **biopsy**. The discomfort caused to the patient by the BM aspiration is mild compared to the discomfort caused by the trephine. Furthermore, more sedation is needed for trephine than for aspiration. For these reasons, aspiration biopsy is used more frequently, but always complementary to the BM biopsy. Consequently, not only cytological details are obtained from the aspiration, but also BM architecture from the biopsy, allowing the location of focal skeletal lesions [1,2,49].

Aspiration biopsy is commonly obtained from flat bones, such as sternum or the ilium, mainly in the posterior superior iliac crest. Even though sternal aspirations are demonstrated to be technically easier and to contain more haematopoietically active tissue, they are more painful and dangerous than iliac crest aspirations [1,49,50]. In addition, posterior iliac crest aspiration can be used for any age, while sternal aspirations are impossible to perform in young children [1].

Once both samples are acquired, aspiration and biopsy, they are stored at 18-25°C and transferred immediately to the **Multiparametric Flow Cytometry (MFC)** laboratory, as the period time between acquisition and analysis must be less than 48h. The collected BM or blood cells are incubated with a mix of antibodies conjugated with fluorochromes. In this way, the flow cytometer works on two phases [2]:

- I. Emission of laser lights, which is absorbed by the fluorochromes.
- II. Measurement of the light emitted by the fluorochromes to estimate the quantity of each cell type.

The result of this test is compared to the normal BM. Firstly, the **BM cellularity** is estimated as the relative percentage of haematopoietic tissue (red marrow)

compared to adipose tissue (yellow marrow). For this evaluation, the normal BM cellularity for the patient age (shown in Table 1) has to be considered, to avoid a mistaken hypo- or hyper-cellularity (below or above the normo-cellularity range) and its consequent misdiagnoses [2,8].

Age group	Normo-cellularity (%)
<20	45-85
20-40	40-70
40-60	35-65
>60	30-60

Table 1. Proposed definitions for normo-cellularity for patient age groups [8].

Secondly, cellular elements counts are measured in relative and absolute terms. To this end, BM nucleated differential cell count (NDC) is performed. In addition, NDC allows the comparison of cellular elements counts with reference ranges and the quantification of abnormal cells. Thirdly, observations such as the presence of metastatic cells, lymphoid aggregates or the myeloid-to-erythroid ratio are considered and calculated if necessary. With this process, the three main cell lineages present at the BM are described for differentiating an infiltrative lesion or involvement [2,49].

1.3.2. Bone marrow assessment in Multiple Myeloma

BM biopsy and aspiration are commonly used in clinical practice not only in MM diagnosis, but also in prognosis estimation or MRD evaluation, i.e. after MM treatment [1,2,35]. Once the sample is extracted: monoclonal plasma cells, in MM diagnosis, and residual tumour burden, in MRD assessment, are measured using MFC or next-generation sequencing (NGS) with a recommended minimum sensitivity of 10^{-5} [35]. Discrepancies can exist between biopsy and aspiration for the BM plasma cells. In these cases, the highest value is used [13].

Both, percentage of plasma cells and their degree of dysplasia, are directly related to a poorer prognosis [51]. In addition, the presence of clusters of plasma cells is likely to be a malignancy [2]. However, the most important measure is the cell count of BM plasma cells, since this is the key distinguishing feature between MGUS and SMM from MM [13].

5% of aspirations result in undiagnosed patients due to the characteristic BM affectation: patchy and with not excessive abnormally morphological cells. In addition, the principles of the aspiration imply not having a sample containing a large number of plasma cells [1,2]. Despite the high sensitivity of these techniques, a

number of MFC- patients, i.e. patients with a negative biopsy result, still experience relapse due to the small representation or poor quality of the taken sample, the heterogeneity of the BM affectation, the existence of extramedullary disease or the insufficient sensitivity of the applied technique [34,52]. For this reason, the combined use of MFC or NGS with visual evaluation of imaging by low-dose whole body Computerized Tomography (CT) or Positron Emission Tomography (PET)/CT with fluorine-18 fluorodeoxyglucose (^{18}F]FDG) is becoming increasingly important in the MRD detection [53–55].

In fact, the predictive value of visual analysis of ^{18}F]FDG PET in combination with MFC was demonstrated by Alonso *et al.* [53]. As shown in Figure 5, patients with negative ^{18}F]FDG PET scans (no ^{18}F]FDG-avid lesions) with MFC- showed longer survival rates than with MFC+. In contrast, PET+ scans showed significantly lower survival rates than PET- scans independently of the MFC results for MRD.

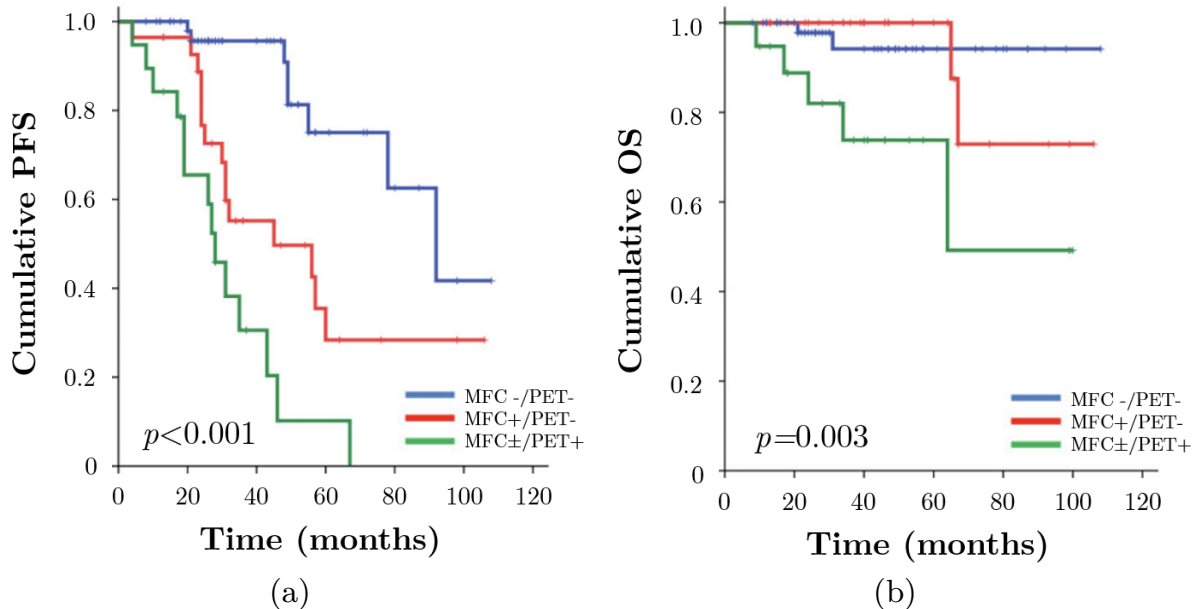


Figure 5. Kaplan-Meier curves with MFC and ^{18}F]FDG PET assessments for (a) Progression-Free Survival (PFS) and (b) Overall Survival (OS). Adapted from [53].

1.3.2.1. ^{18}F]FDG PET/CT assessment

The 2003 IMWG criteria defines Magnetic Resonance Imaging (MRI) and CT as the main techniques to evaluate the presence of bone disease [56]. Over the last decade, advances in these imaging techniques have extended their use, especially for detecting bone lesions (greater sensitivity than X-ray scans, detecting 80% of bone lesions). However, to assess BM infiltration, PET imaging has become increasingly important thanks to its usefulness to evaluate disease activity, to detect extra-bone disease involvement and to direct local therapy. Due to its functional nature, **^{18}F]FDG uptake is correlated with the percentage of BM plasma cells.** Individually CT and PET, both techniques provide an acceptable assessment of

treatments outcomes, but combined more accurate results are obtained. This technique is becoming more recommended than MRI thanks to its potential to detect medullary and extramedullary lesions and the possibility of distinguishing between new active and old diseases [53,57–62]. Moreover, although this technique detects less focal lesions, it is equally effective for bone involvement in MM patients [63–67].

Furthermore, this imaging technique has also multiple advantages in the MM evaluation: extended field of view; absence of possible collateral or adverse effects; possible to perform for renal failure patients; short acquisition time; soft tissue and organs evaluation simultaneous to extramedullary disease; semi-quantify disease activity using the **Standardized Uptake Value (SUV)**; and no limitations in relation to metallic bone implants [11]. Consequently, the IMWG recommends the use of [^{18}F]FDG PET/CT for detecting bone and extramedullary disease (Figure 6) and to evaluate the response to the therapy for MM patients, i.e. the presence of MRD (Figure 7) [58,68,69]. As shown in both figures, Figure 6 and Figure 7, which are examples of Maximum Intensity Projection (MIP) of [^{18}F]FDG PET images, different uptake patterns can be observed using this technique. If the diagnosis is not clear enough, the study should be repeated in 3-6 months and/or a biopsy of one of the bone lesions should be performed [13].

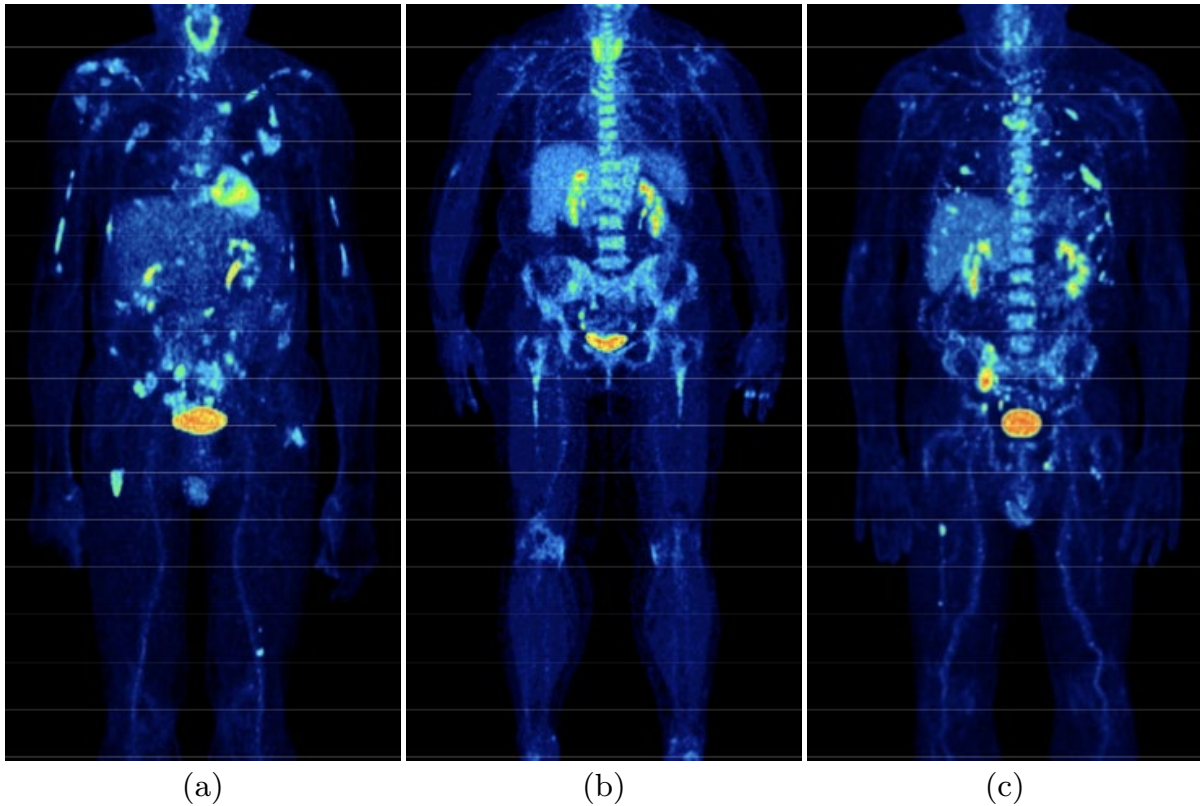


Figure 6. MIP of [^{18}F]FDG PET images of newly diagnosed MM patients, representing examples of different pathologic patterns: (a) multiple skeleton focal lesions; (b) diffuse uptake in the BM without focal lesions; (c) diffuse uptake pattern and focal BM lesions. Adapted from [59].

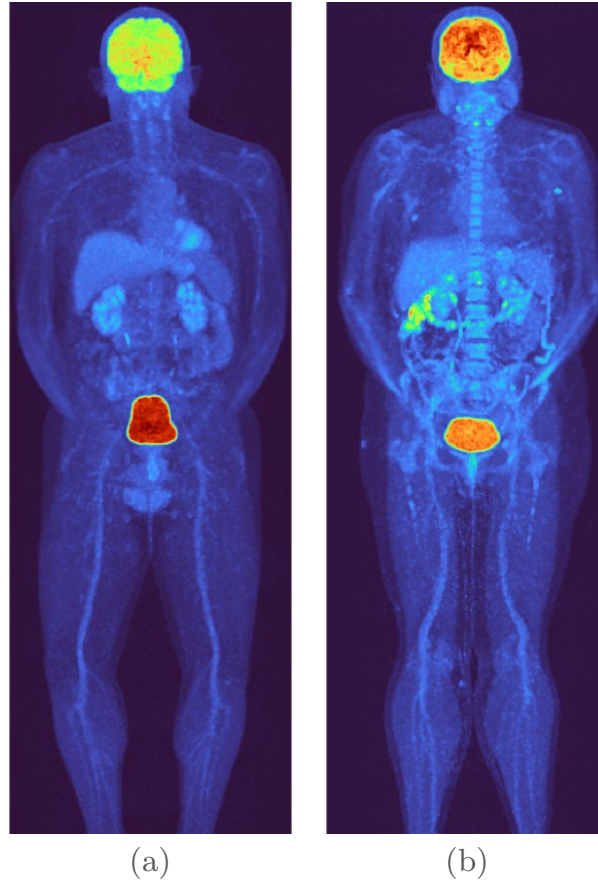


Figure 7. MIP of $[^{18}\text{F}]\text{FDG}$ PET post-treatment images of MM patients, representing: (a) complete response; (b) MRD BM uptake pattern.

The interpretation of $[^{18}\text{F}]\text{FDG}$ PET studies is currently based on visual analysis. **IMPETUs**, proposed by C. Nanni and E. Zamagni [70], is a standardised reading model for MM $[^{18}\text{F}]\text{FDG}$ PET studies based on visual analysis. The final version of these criteria is based on the following points: (1) metabolic state of BM; (2) number, sites and metabolic state of focal PET positive lesions with or without osteolytic characteristics; (3) presence, site and metabolic state of extramedullary disease; (4) presence of paramedullary disease; and (5) presence of fractures [11]. In addition, the IMPETUs criteria exclude SUV metrics as a MM marker [70] due to its variability in clinical practice, since these metrics could be affected by inflammation processes, noise and the reconstruction method [71]. The first three points are measured based on the **Deauville Criteria**. Deauville Criteria [72] is based on the use of the liver uptake threshold as a cut-off point to differentiate positive from negative lesions in MM. This criterion establishes five different metabolic states: (1) no uptake; (2) equal or less uptake than SUVmax of mediastinal blood; (3) more uptake than mediastinal blood but equal or less than SUVmax of liver; (4) more than liver SUVmax plus its 10%; and (5) at least twice more than the maximum liver uptake [11]. Deauville Criteria is useful for characterization of focal bone lesions and extramedullary disease, but insufficient for BM evaluation in difficult-to-assess cases [72].

On the other hand, in clinical practice, MRI is the main technique used in SMM patients as part of the initial assessment for detecting abnormalities such as focal lesions. Focal lesions are present in 28% of SMM patients, with only one lesion in 55% of cases. 70% of patients with more than one focal lesion progress in less than 2 years, i.e. they are categorised as high-risk SMM. Despite the importance of focal lesions detection, in all patients with suspected SMM, the IMWG [24] recommends [^{18}F]FDG PET/CT or low-dose whole-body CT for bone disease evaluation due to the diffuse marrow infiltration pattern caused by the BM plasma cells, difficult to assess by MRI [13].

1.3.3. Bone marrow involvement assessment in NHL

BM aspirations or trephine biopsy are routine first-line methods to detect BMI in NHL, with a posterior analysis of the biopsy content. In addition to being a painful procedure with high risk of excessive bleeding, the taken sample can be not large enough for acquiring conclusive results since is inadequate sampling of the whole BM. Additionally, due to its patchy and heterogeneous pattern, the BM aspiration can be normal when BMI is histologically detectable or focal lesions are presented. Indeed, focal infiltration is commonly undetected by BM aspiration [1,9,42,46].

Although these techniques are the gold standard for assessing BMI in lymphoma staging, due to their limitations, there is growing interest in the use of [^{18}F]FDG PET/CT imaging. Some studies have demonstrated the power of biopsy and [^{18}F]FDG PET/CT combination for lymphoma staging, leading to its increased use in clinical practice [73,74]. Despite discrepancies between both techniques results, their combination is considered the optimal methodology for an accurate staging [47].

1.3.3.1. [^{18}F]FDG PET/CT assessment

For aggressive NHL staging, CT and [^{18}F]FDG PET has demonstrate to be useful tools, with a sensitivity of 78% and 86% respectively. However, the combination of both techniques stands as **the most useful** technique for staging lymphoma, since [^{18}F]FDG PET/CT images have higher sensitivity and specificity, 91.6% and 90.3%, respectively, than MRI, 90.3% and 75.9%, in detecting extramedullary lesions [46].

This technique is included in the clinical practice not only for NHL diagnosis, but also BMI assessment and, consequently, staging. In some subtypes of NHL, such as DLBCL, this image technique has almost replaced the need of biopsies. However, in patients with FL, the biopsy evaluation is still needed [47].

Not only [^{18}F]FDG PET/CT is demonstrated to be a useful technique for BMI evaluation, as it can be seen in Figure 8 where different BM uptake patterns are shown, but also to be a prognostic factor for both, overall and progression-free

survival [47]. For these reasons, [^{18}F]FDG PET/CT scan is recommended by the International Conference on Malignant Lymphomas (ICML) [75] for lymphoma staging.

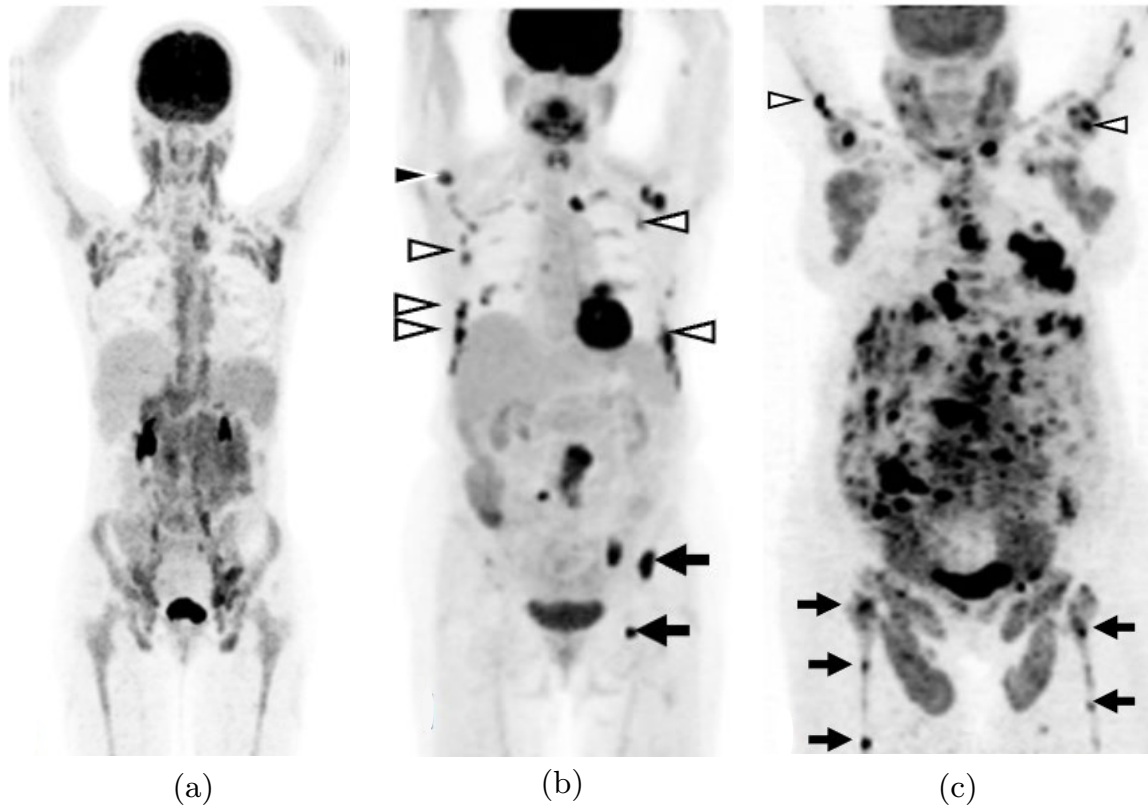


Figure 8. MIP of [^{18}F]FDG PET images of FL patients representing different patterns of FDG uptake: (a) pure diffuse uptake pattern, considered BMI-; (b) multifocal uptake pattern without diffuse uptake, considered BMI+; (c) multifocal uptake with additional diffuse uptake greater than liver reference, considered BMI+. The arrows indicate some of the most evident areas of increased uptake. Adapted from [47].

In addition to BM evaluation, [^{18}F]FDG PET allows to evaluate other organs belonging to immune system. In healthy subjects, the [^{18}F]FDG uptake in the spleen is less than in the liver. However, in both, patients with lymphoma and patients affected by inflammation or infection, there is a greater uptake in the spleen than in the liver, since the affinity of glucose transporters for deoxyglucose is increased during conditions which requires an overproduction of immune cells. For this reason, the [^{18}F]FDG PET/CT acquired during the lymphoma diagnosis is also useful to determine the diffuse high uptake of the spleen and liver. However, the spleen diffuse increment of uptake can reflect lymphoma infiltration, but also proliferation and activation of macrophages [10].

1.4. Problem statement and justification

Multiple pathologies can affect the function of the BM. Its clinical assessment is generally performed by biopsy extraction of a sample of BM from flat bones and its subsequent analysis by techniques such as MFC or NGS [35]. However, due to the pain caused during the process and the lack of representativeness of the small sample taken, in recent years there has been a growing tendency to use functional medical images, especially **[¹⁸F]FDG PET/CT images**, and their corresponding **visual assessment** by experts in nuclear medicine for many clinical contexts such as MM [58,68] and lymphoma [76]. Despite being a widely used assessment, it still has some **subjectivity** because of the observers and their years of experience, visualising systems, subtle image intensity changes, etc. Due to these factors, some challenging-to-assess cases cause **discrepancies** between experts and even between visual evaluation and biopsy results.

The present thesis research focuses on **the proposal and development of methodologies for quantifying BM in [¹⁸F]FDG PET/CT images with the aim of assisting the visual interpretation by nuclear medicine experts**, seeking to improve personalised medicine. Furthermore, these quantification methodologies are meant to be applied in different clinical contexts. On the one hand, they are applied in the detection and prediction of MRD in MM patients, which implies a worse prognosis and survival. In the MM field, they are also utilised to estimate the risk of progression to MM in SMM patients, to personalise a therapeutic strategy. On the other hand, they are employed in the differentiation of BM status between healthy, infiltrated, or reactive, in patients diagnosed with NHL for their staging, improving the treatment adequation.

The present PhD thesis is developed in the Biomedical Engineering and Telemedicine Centre (Grupo de Bioingeniería y Telemedicina, GBT) of the Escuela Técnica Superior de Ingenieros de Telecomunicación (ETSIT), Centre for Biomedical Technology (CTB) of the Universidad Politécnica de Madrid, Madrid, Spain. This thesis is specifically performed within the “Medical imaging, surgical training and image-guided surgery” research laboratory. This laboratory has more than 45 years of experience in medical imaging analysis, thanks to its participation in numerous national and international projects in this field. In addition, it has actively collaborated with numerous internationally renowned hospitals and research centres.

As the project developed in this thesis is a **multidisciplinary biomedical engineering project**, it is necessary to collaborate with clinical specialists. In this way, a close collaboration with the clinical experts of the Nuclear Medicine Department and the Haematology Department of the Hospital Universitario 12 de Octubre (Madrid) is performed. Thanks to this collaboration, a definition of the

clinical objectives, databases obtention, especially images, and results validation from the clinical point of view are achieved.

1.5. Thesis Outline

The thesis outline is structured in 6 chapters: Introduction (Chapter 1), Research hypotheses and objectives (Chapter 2), State of art (Chapter 3), Materials and methods (Chapter 4), Results (Chapter 5) and Conclusions and future works (Chapter 6).

In **Chapter 1**, the clinical context of the present thesis is introduced. With this aim, a basic introduction of the BM is described. Then, different pathological affectations of this organ together with the detection techniques are detailed. Then the justification of the work is described. After the justification statement, the **Chapter 2** includes the hypotheses and objectives of the present thesis.

Once the objectives are defined, the **Chapter 3** is defined as the state of art for evaluating the current existing techniques for both main objectives, segmentation and quantification of BM in different pathological contexts. **Chapter 4** covers the materials available for testing the methodology also proposed in this chapter. This methodology is divided in the two main objectives. In terms of segmentation, two different approaches are proposed: a classical one, based on traditional image processing techniques; and an approach based on Deep Learning (DL). On the other hand, for BM quantification, two approaches are also proposed, focused on BM region obtained from the optimal segmentation approach: one using Machine Learning (ML) models based on radiomics features extracted from $[^{18}\text{F}]\text{FDG}$ PET; and another approach based on DL architectures. Then, **Chapter 5** includes the results of the methodology proposed in the previous chapter, following the same structure.

Finally, **Chapter 6** includes the discussion of the research hypotheses. Once the proposed methodology and results are discussed, the conclusions of the complete thesis and proposals for future works following the research described are included.

2. Research hypotheses and objectives

2.1. Research hypotheses

The main hypothesis of this PhD thesis is defined as follows:

H1. Quantitative analysis of bone marrow, using [^{18}F]FDG PET image processing methodologies, allows to quantify the differences in uptake patterns for the study of bone marrow involvement or affectation. The integration of these methodologies in a platform supports the clinical decision-making process in the evaluation of bone marrow.

Multiple pathologies can affect the function of the BM such as MM and Lymphoma. Its clinical assessment is generally performed by biopsy extraction of a sample of BM from flat bones, usually sternum and iliac crest, and its subsequent analysis by techniques such as MFC or NGS [35]. However, in recent years there has been a growing tendency to use functional medical images, especially PET images combined with CT, and their corresponding visual assessment by experts in nuclear medicine to perform this evaluation for two reasons:

- I. BM assessment by biopsy is an invasive and painful process for the patient. In the case of [^{18}F]FDG PET/CT imaging, the patient only suffers the radiotracer administration, and the ionizing radiation received during image acquisition, which is not an alarming amount of radiation.
- II. The biopsy taken is a small sample relative to the full extent of the BM, so it may not be representative of its overall status. In the case of [^{18}F]FDG PET/CT imaging, nuclear medicine experts can visually assess the radiotracer uptake for the full extent of the BM and can locate focal lesions that are not assessable in most cases where a biopsy is performed.

In addition, [^{18}F]FDG PET has been recommended by the IMWG [58,68] as the main tool for the assessment of BM affectation in MRD. On the other hand, the ICML [75] also suggest this image technique for staging in Lymphoma patients.

This PhD thesis is focused **on the proposal and development of [^{18}F]FDG PET/CT image processing methodologies and the implementation of a platform that incorporates them**. First, a methodology for the **segmentation** of the BM from anatomical images, specifically CT images, followed by the **quantification** of this segmentation by analysing the BM uptake in functional images, in this case, [^{18}F]FDG PET images. By this, obtaining a **clinical decision support system** in the form of an easy-to-use platform for nuclear medicine experts

to objectify the quantification of the state of BM involvement or infiltration is the main aim. Thus, the tools developed and included in the platform are divided into two main phases: (1) BM segmentation and (2) BM quantification for evaluating its status in different clinical contexts. As soon in Figure 9, the development of both phases and the application of these quantification methodologies in the different clinical contexts represent the six hypotheses of this thesis.

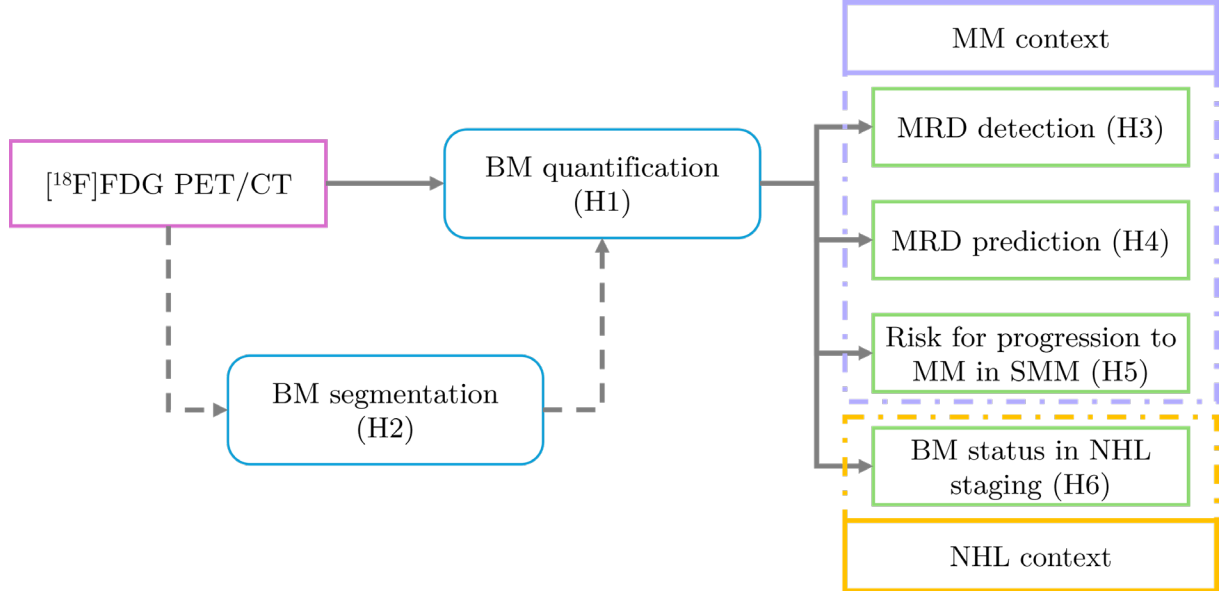


Figure 9. Illustration of hypothesis definition.

For the **phase (1)**, three different segmentation methodologies are proposed, from the highest to the lowest time-consuming process: (a) manual segmentation; (b) semi-automatic segmentation consisting in an automatic segmentation based on traditional medical image processing algorithms followed by a manual revision and edition of the segmentation; and (c) automatic segmentation based on DL models.

For the **phase (2)**, the use of two different quantification techniques for classifying BM in different status depending on the uptake of $[^{18}\text{F}]\text{FDG}$ in the PET are proposed: (a) quantification by **ML models** based on extracting **radiomics features** from $[^{18}\text{F}]\text{FDG}$ PET images in regions classified as BM; (b) **DL models** by using $[^{18}\text{F}]\text{FDG}$ PET images in BM region as the input. The methodologies of the second phase, corresponding to the quantification of BM status and its possible classification, are evaluated in different clinical contexts: MM, especially for the detection of MRD, but also for MRD prediction and risk of progression to MM evaluation in SMM patients; and NHL, for the evaluation of BM status.

Both the hypothesis related to phase 1, i.e. BM segmentation for correct quantification, and the hypotheses related to the application of phase 2, i.e. quantification methodologies to each of the proposed clinical contexts, are defined

below. The relationship between these secondary hypotheses and the main hypothesis can be seen in Figure 9.

H2. Automatic segmentation models from CT images enable the generation of bone marrow masks to perform a subsequent quantification and/or evaluation of its condition, providing a less time-consuming alternative compared to manual segmentation.

Currently, there are multiple studies that perform automatic segmentation of the entire skeleton from CT images [77–80]. However, there are no known studies that perform automatic segmentation of the entire BM and evaluate its performance quantitatively [81–85].

Manual BM segmentation from a three-dimensional (3D) CT image requires a large amount of time to obtain an accurate segmentation. Additionally, even with visual assessment, it involves a subjective segmentation due to the visual relationship between the region corresponding to cortical bone and BM, and also the subjectivity of the user who makes use of the tool. For these reasons, the present investigation aims to obtain an automatic segmentation of sufficient quality to perform an adequate quantification of the BM.

H3. [^{18}F]FDG PET images used in the assessment of treatment response in patients with Multiple Myeloma can be quantified to detect Measurable Residual Disease through analysis of bone marrow.

MM is a disease characterised by BM affection due to the production of immunoglobulins in this region [14]. Even though in recent decades the number of patients achieving a complete response after treatment has increased considerably, some patients show BM affection defined as MRD by IMWG, identified as a biomarker of worse prognosis and poorer survival [35,36]. As discussed above, [^{18}F]FDG PET has been recommended by the IMWG [58,68] as the main tool for the assessment of BM affection in MRD in order to assess the extent of bone disease and the presence of extramedullary disease, to identify lesions before treatment, and to evaluate the response to the therapy.

The present thesis research focuses on quantifying [^{18}F]FDG PET/CT images visually evaluated by nuclear medicine experts in an attempt to **detect possible MRD** with the aim of assisting the work of the clinical expert, reducing the discrepancies with biopsy evaluation. For this purpose, the previously described quantification methodologies are applied to treatment response assessment imaging of treated MM patients, trying to detect MRD using as ground truth (GT) individually: (1) visual assessment of [^{18}F]FDG PET/CT by nuclear medicine experts, (2) biopsy evaluation by MFC, and (3) biopsy evaluation by MFC only in

patients assessed as MRD- by visual assessment of ^{18}F FDG PET/CT imaging. In this way it is not only desired to quantify the PET-based assessment, but also to try to find uptake patterns to differentiate the biopsy result, especially in the most doubtful cases.

H4. Progression to Measurable Residual Disease can be predicted by quantifying ^{18}F FDG uptake in bone marrow on Multiple Myeloma diagnostic imaging.

In addition to the importance of MRD detection as an aid to clinical therapeutic decision making, a possible identification at the time of MM diagnosis of cases with a higher risk of developing MRD could help to personalise the therapeutic strategy to be followed. Thus, based on the premise of the ability of ^{18}F FDG PET imaging to detect extramedullary disease during the diagnosis of MM [58,68,69], the present hypothesis focuses on the possibility of **identifying patterns** of BM uptake at the time of diagnosis that would allow assessment of an increased risk of developing MRD. For this purpose, the quantification methodology based on radiomics features and ML models is applied, in order to predict MRD. Different GTs based on MRD diagnosis techniques are used: (1) visual assessment of ^{18}F FDG PET/CT by nuclear medicine experts, (2) biopsy evaluation by MFC.

H5. Potential biomarkers of risk for progression to Multiple Myeloma can be identified by quantitative analysis of diagnostic ^{18}F FDG PET images of Smouldering Multiple Myeloma.

In order to guide the clinician throughout the management of the MM patient, it is proposed to divide patients diagnosed with SMM according to their risk of progression to MM. In this way, it is intended to **predict earlier or later progression**, with the aim of helping clinicians in defining the specific therapeutic strategy for each patient.

Currently, this division is made on the basis of clinical data, but the present hypothesis focuses on identifying patterns of BM uptake in ^{18}F FDG PET imaging used for the diagnosis of SMM which allow discerning a higher or lower risk of progression. To do this, radiomics features are extracted from PET images in regions of the BM and analysed individually, trying to find patterns or potential biomarkers that distinguish different levels of risk of progression, using two clinical assessments as GT.

H6. [^{18}F]FDG PET images used in the Non-Hodgkin's Lymphoma diagnosis can be quantified to detect bone marrow infiltration or reaction to inflammation or infection.

In the case of NHL, BMI is not the main lesion, but it can produce anaemia, thrombocytopenia and neutropenia, therefore adequate detection is key for the staging of Lymphoma, prediction of its **prognosis and therapeutic consequences**. In the same way as for MRD, the evaluation of BMI is performed by biopsy analysed by MFC together with a visual assessment of [^{18}F]FDG PET performed by nuclear medicine experts, as recommended by the ICML [75].

The present thesis research focuses on quantifying [^{18}F]FDG PET/CT images visually interpreted by nuclear medicine experts in an attempt to **detect the existence of BMI**, and its differentiation from a **reactive state**, in diagnostic images of NHL patients with the aim of assisting the work of the clinical expert. For this purpose, two different quantification techniques are used. In the same way as for MRD, the aim is to differentiate BM status at the time of NHL staging, using as GT individually: (1) visual assessment of [^{18}F]FDG PET/CT by nuclear medicine experts, (2) biopsy evaluation by MFC, and (3) a combination of both techniques based on the sensitivity of each for the detection of the possible states. In addition, this differentiation is also performed individually for each of the NHL subtypes, with the aim of evaluating the ability of the quantification methodologies in reference to the aggressiveness of the subtype evaluated.

2.2. Objectives

Based on the hypotheses proposed for this thesis, the main objective is defined as follows:

Proposal and development of [^{18}F]FDG PET/CT image processing methodologies to quantify bone marrow in diseases such as Multiple Myeloma or non-Hodgkin's lymphoma as an aid to clinical decision making.

The specific objectives are as follows:

- I. The definition and preparation of a database of whole-body [^{18}F]FDG PET/CT images of patients diagnosed with one of the mentioned clinical contexts
- II. The development of optimal BM segmentation methodologies from CT images for subsequent quantification, based on:
 - a. Traditional medical image processing algorithms

- b. DL models
- III. The development of [^{18}F]FDG PET quantification methodologies to assess the BM involvement:
 - a. Extracting radiomics features to obtain potential prognostic biomarkers
 - b. Developing automatic classification models based on the extracted radiomics features
 - c. Implementing automatic classification models based on [^{18}F]FDG PET images, instead of radiomics features
- IV. The design and implementation of a platform that includes the methodologies developed for BM segmentation and quantification that can be used as a clinical decision support system by nuclear medicine experts
- V. The application and validation of the proposed segmentation and quantification methodology to different clinical contexts:
 - a. MRD detection in MM patients
 - b. MRD prediction in MM patients
 - c. Risk for progression to MM in SMM patients
 - d. BMI and rBM differentiation in NHL patients

3.State of art

3.1. Bone and bone marrow segmentation

For a precise BM quantification, an accurate segmentation is essential. Compact bone and BM are neighbouring components of bones. Due to the difficulty of segmenting the BM from the CT, the majority of authors who have tried to segment this region draw from the compact bone segmentation, to extract the BM afterwards [86].

3.1.1. Bone segmentation

Bone segmentation is considered of great usefulness, not only for BM segmentation, but also for multiple clinical contexts. One example is the evaluation and detection of bone morphological changes, since it is the most common and prevalent metastasis site, mainly in breast and prostate cancer [87,88]; especially when the visual assessment of morphological images is not adequate according to the Response Evaluation Criteria In Solid Tumors 1.1 (RECIST1.1) [89]. Other important examples where the bone segmentation can be beneficial are the assessment of new oncological therapies or the evaluation of osteoporosis in order to evaluate the fracture or weakness examination [90–92].

Routinely, bone segmentation is manually obtained, implying a time-consuming and subject to inter-operator variability process. In the last two decades, several approaches for this segmentation have been proposed. In 2003, Kang *et al.* [93] proposed a semi-automatic method for the segmentation of 3D skeleton structures from CT images. The segmentation was based on the use of local adaptative thresholds estimated from a sphere manually selected by the user for developing a **region growing** algorithm, followed by multiple morphological operations for obtaining a smoother surface. Once the structures were segmented, the periosteal and endosteal surfaces were separated. Despite the semi-automatic nature of the method, the results were visually satisfactory for knee, femur and skull images. Furthermore, the precision error, calculated as the standard deviation of the differences between segmentations, was lower than 2% in the pelvic CTs that were tested.

Several authors combined the use of **thresholding along with morphological operations** for skeleton segmentation. In 2007, Buie *et al.* [86] developed a methodology for tibia and radius segmentation based on applying two manually selected thresholds followed by median filter, multiple morphological operations, such

as dilation and erosion, connectivity filter and gaussian filter for edge smoothing. With the complete methodology, they obtained not only the periosteal surface but also the endosteal surface, i.e. the BM. Their cohort was composed of 45 mouse, rat, human and cadaver CT images. Although the high correlation obtained for the majority of segmentation results ($R^2 \sim 0.95$), it was considered not good enough for measuring cortical thickness. Then in 2010, Banik *et al.* [94] implemented a fully automatic segmentation method for three different structures: ribs, vertebral column and spinal canal; which was tested with 39 CTs of 13 different patients. Although the first step for the three segmentation methods was the same, the removal of peripheral artefacts, skin and fat, so each structure has its own algorithm for being segmented. All of them were based on the combination of thresholding and morphological operations. Despite not segmenting the whole-body, this is the first completely automatic method found in this field. An example of the resulting segmentation for the vertebral column structure is shown in Figure 10.

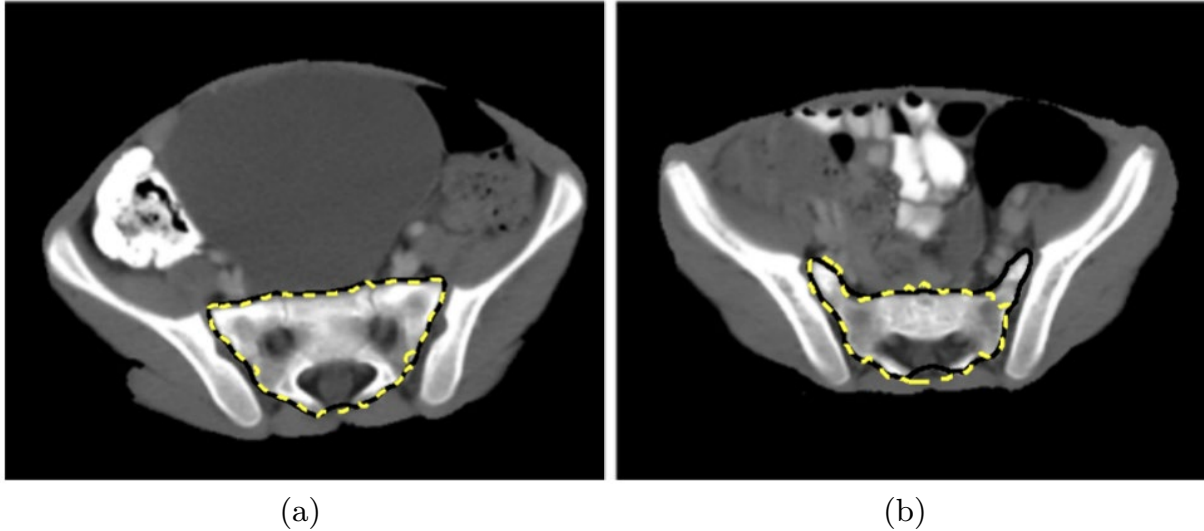


Figure 10. Results of the vertebral column segmentation with the method of Banik *et al.* [94]: (a) good and (b) acceptable segmentation. (Black) contour of the vertebrae drawn by the radiologist; (yellow) contour produced by the Banik *et al.* method.

Following with traditional image analysis methods, in 2008 Kratky *et al.* [95] and Zhang *et al.* [96] proposed methodologies based on **level-sets** algorithm. Kratky *et al.* used a fast level-set method based on moving the pixels corresponding to the external boundary of the object, while Zhang *et al.* focused on integrating boundary and region information of the object to the level-set method. Both methods started from an initial seed manually positioned, although Kratky *et al.* method also required a manual parameters selection. Finally, both methods were tested in 3D CT images, for thorax and for trachea segmentation, respectively, without quantitative metrics to evaluate the performance. The results of both methods can be seen in Figure 11.

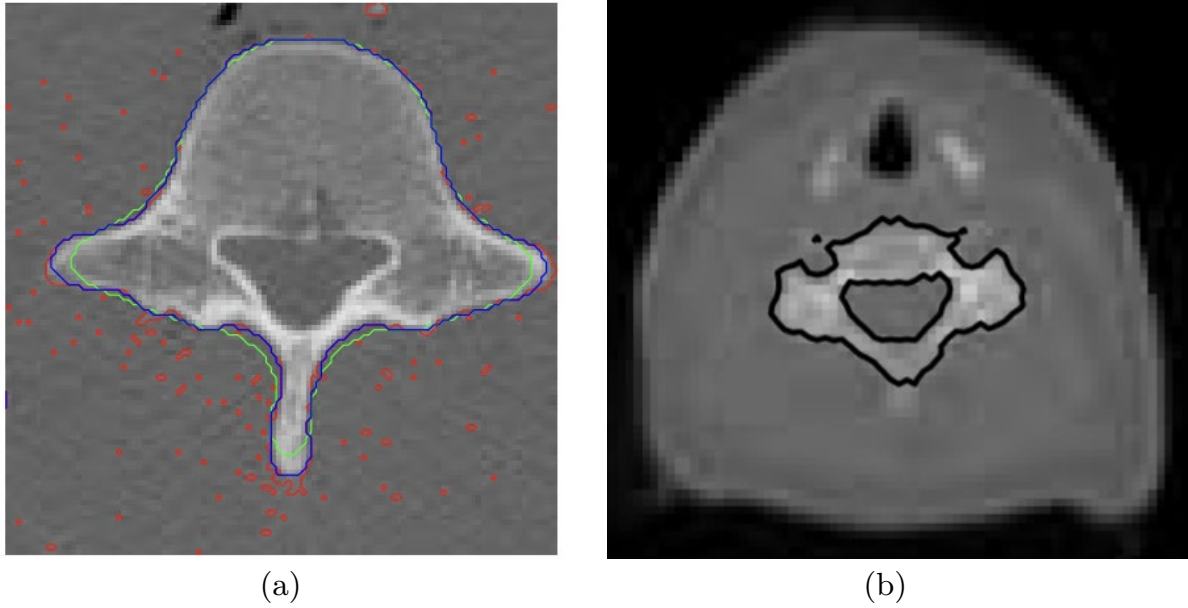


Figure 11. Results of the vertebrae segmentation for (a) Kratky *et al.* [95] and (b) Zhang *et al.* [96]. In (a) different colours represent different level of smoothness: (red) low smoothness with segmentation artefacts; (green) excessive smoothing; (blue) properly for fitting the boundary correctly.

Another method used for bone segmentation and based on finding object edges like level-sets, is the **active contour** method. In 2009, Li *et al.* [97] tried the active contour method based on snake model for visually evaluating its usefulness in the wrist segmentation from X-ray images. Then in 2013, Yousefi *et al.* [98] progressed in the used of active contours. They developed a fully automatic method for segmenting radius bone from MRI, drawing from a cohort of 23 subjects with 600 slices. First, denoising and contrast enhancement algorithms were applied as a pre-processing step. Once the 3D image was prepared, a Canny filter was applied to find the medium slice of the radius, in this case, the slice with the biggest radius mask, i.e. more pixels. From this medium slice, the algorithm was extended into the rest of slices, obtaining an automatic segmentation with a Dice index of 0.95. Thus, although in 2008 Zhang justified the use of level set because active contours are limited in the detection of irregular edges and can converge to erroneous solutions, Yousefi *et al.* demonstrated the usefulness of active contours for the proposed segmentation.

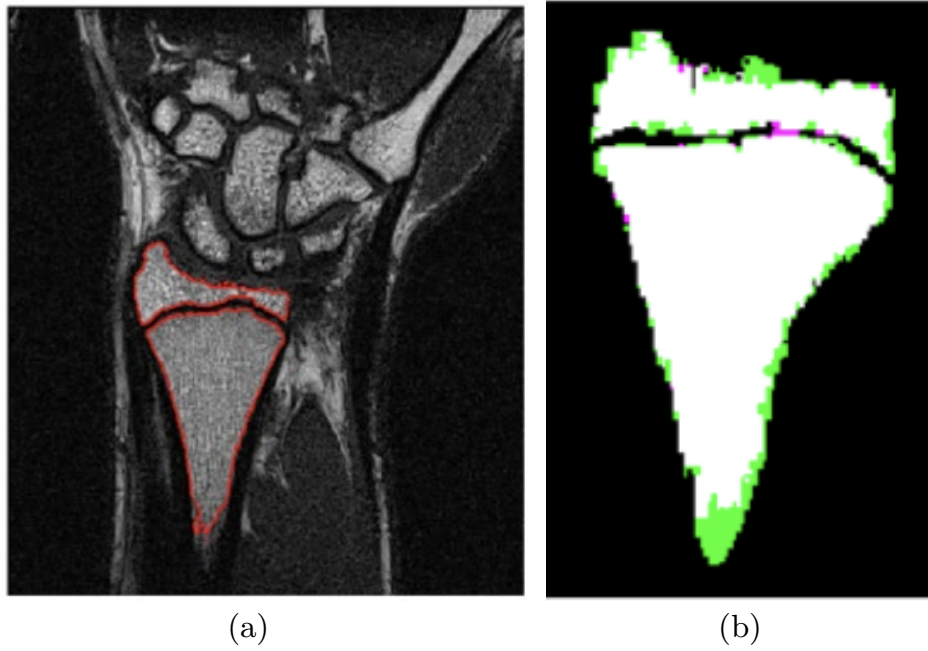


Figure 12. Results of Yousefi *et al.* [98] methodology: (a) segmentation delineated in red over the CT image; (b): (white) true positive pixels; (pink) false positive; (green) false negative; and (black) true negative.

In those years, more automatic methodologies were proposed. In 2011, Ababneh *et al.* [99] introduced the use of **textural metrics** for knee bones segmentation. In this methodology, a significant block discovery was twice applied for distinguishing between bone and background. This method was based on pixel intensity distribution, Grey Level Co-occurrence Matrix (GLCM) metrics and a developed metric which measure the dissimilarity between bone and background. Moreover, the method applied some morphological operations after the two blocks. The developed system was tested in 376 MRI of tibia and femur of 200 patients from the Osteoarthritis Initiative database. In addition, it was also tested in 10 3D MRI of 14 slices each volume. This way Ababneh *et al.* obtained a 0.95 Dice index using GLCM metrics (Figure 13a). In 2014, Younes *et al.* [100] proposed an innovative methodology based on 3D **deformable models**. Using 20 CT images of the human femur, they developed an automatic method based on three steps: the detection of the femoral head and shaft as a sphere and a cylinder, respectively; the registration between the shape models and the CT images; and the adjustment of the model to the CT image. They obtained satisfactory results, with a Dice index of 0.87 (Figure 13b).

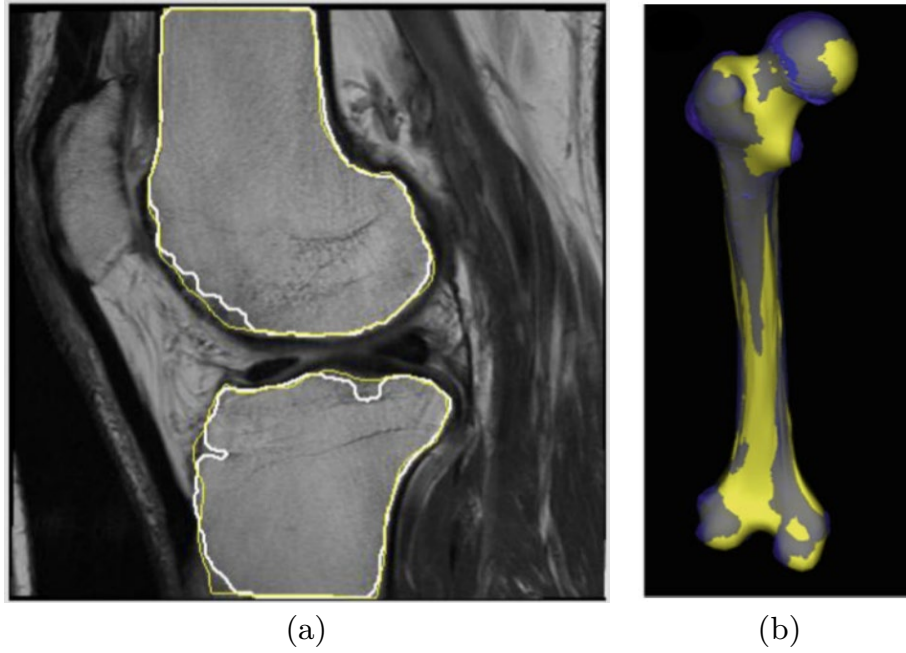


Figure 13. Results of the segmentation for (a) Ababneh *et al.* [99] and (b) Younes *et al.* [100] methods. In (a/b) different colours represent: (yellow/blue) GT and (white/yellow) segmentation obtained with the proposed algorithm.

All of the mentioned algorithms were developed for the segmentation of only some concrete structures of the skeleton. However, none of the mentioned studies apply the algorithms for whole-body segmentation. In 2018, Pérez-Carrasco *et al.* [77] implemented a segmentation based on energy minimization algorithm by means of measuring **histogram distances**, after region growing algorithm for skin delimitation as pre-processing step. With 27 CT images of different parts of the body of 20 patients and a 3-folds cross-validation method, they obtained a Jaccard index (Intersection over Union, IoU) of 0.80 and a Dice index of 0.88. Although the cohort used for this study was composed of images of several parts of the body (Figure 14), none of the images represent a whole-body CT.

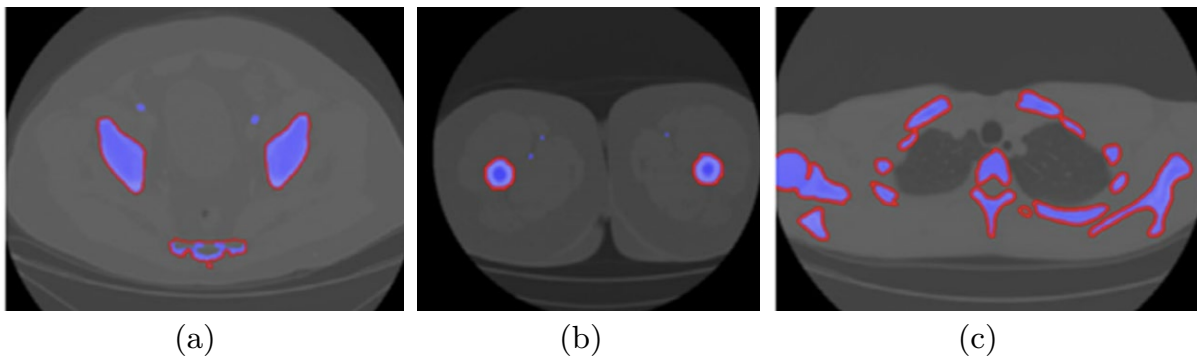


Figure 14. Examples of the segmentation obtained with the methodology proposed by Pérez-Carrasco *et al.* [77] for (a) pelvic, (b) femurs, and (c) shoulders and ribs structures. (Red) GT; (blue) obtained segmentation.

Not only traditional image processing algorithms are applied for bone segmentation, but also **DL models**, which are currently widely extended. Korez *et al.* [101] developed in 2016 an automated method for the supervised segmentation of vertebral bodies. From a cohort of 23 T2-weighted 3D MRI of the lower spine, they designed a Convolutional Neural Network (CNN) for obtaining a probability map to guide deformable models into the shape of the structure with a Dice index of 0.93. Beyond using CNNs as an intermediate step in segmentation, some authors use them as the main method. In 2018, Klein *et al.* [78] implemented a CNN with an architecture based on U-Net [102] for whole-body skeleton automatic segmentation. With 64 channels for the first feature map and a dataset of 18 CT volumes, the resulting model obtains a segmentation of the skeleton from the two-dimensional (2D) slices with a IoU and Dice Index of 0.91 and 0.95, respectively. In line with this study, Noguchi *et al.* [79] implemented in 2020 a very similar model with the difference of using a dataset of 32 CT volumes and 16 channels for the first feature map. In addition, they trained the model with robust data augmentation methodology. Despite the simpler model due to the reduction of channels, the performance exceeds the previous one, reaching values of 0.97 and 0.98 for the same indexes. Examples of these three segmentations can be found in Figure 15.

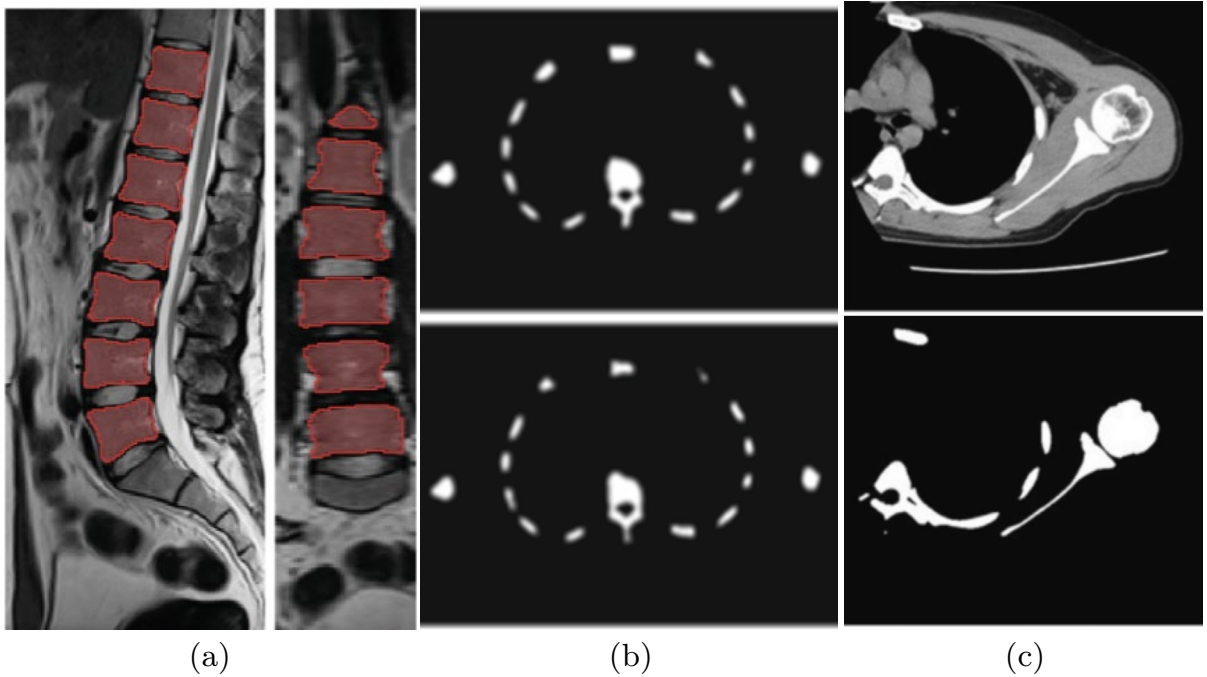


Figure 15. Results of the segmentation for (a) Korez *et al.* [101] showing in red the vertebrae segmentation results; (b) Klein *et al.* [78] showing (superior) the GT and (inferior) the result of the segmentation methodology; and (c) Noguchi *et al.* [79] showing (superior) the original CT and (inferior) the result of the segmentation methodology.

In Table 2, all of the mentioned studies for bone segmentation are included, classified by: method of segmentation, author and year, imaging modality, segmented anatomical regions and results.

Method	Authors, year	Imaging modality	Anatomical regions	Results
Region Growing	Kang <i>et al.</i> [93], 2003	CT	Knee, femur and skull	Precision error < 2%
Thresholding	Buie <i>et al.</i> [86], 2007	CT	Tibia and radius	$R^2 \sim 0.95$
+ Morphological Operations	Banik <i>et al.</i> [94], 2010	CT	Ribs, vertebral column and spinal canal	*Fully automatic segmentation
Level sets	Kratky <i>et al.</i> [95], 2008	CT	Thorax	-
	Zhang <i>et al.</i> [96], 2008	CT	Trachea	-
Active contours	Li <i>et al.</i> [97], 2009	X-ray	Wrist	-
	Yousefi <i>et al.</i> [98], 2013	MRI	Radius	* Dice = 0.95
GLCM	Ababneh <i>et al.</i> [99], 2011	MRI	Tibia and femur	* Dice = 0.95
Deformable models	Younes <i>et al.</i> [100], 2014	CT	Femur	* Dice = 0.87
Histogram distances	Pérez-Carrasco <i>et al.</i> [77], 2018	CT	Several bones	* Dice = 0.88
CNN	Korez <i>et al.</i> [101], 2016	MRI	Vertebral bodies	* Dice = 0.93
	Klein <i>et al.</i> [78], 2018	CT	Whole-body	* Dice = 0.95
	Noguchi <i>et al.</i> [79], 2020	CT	Whole-body	* Dice = 0.98

Table 2. Resume of all studies for bone segmentation. (-) not quantitatively evaluated. Studies marked with * obtained an automatic segmentation.

3.1.2. Bone marrow segmentation

As it has been mentioned, some of the bone segmentation studies also extract the BM mask or isolate the compact bone [86,93], but not automatically. In addition, although many studies focused on the segmentation of spinal cord or canal [103–106], which are also surrounded by compact bone, the studies focused on BM are very limited.

In 2015, Perasso *et al.* [82] evaluated the use of the **Hough Transform** to the spinal marrow segmentation. This commonly used method is a technique based on the classical pattern (lines, circles, ellipses, etc.) recognition to define different profiles of interest from the images. In this study, they recognised the spinal marrow from a CT image to visually evaluate the result, but no quantitative tests or trials in other CT images were done.

The studies focused on BM segmentation are usually centred in optimizing radiation in radiotherapy treatments. Li *et al.* [107] and Fiandra *et al.* [108], in 2017 and 2023 respectively, developed algorithms for active BM segmentation from CT images of the pelvic region in cervical and anal cancer, respectively. Both algorithms were based on segmenting as active BM all pixels with a value higher than the mean SUV (**SUVmean**) of the pelvic region, being this region manually delineated or thresholded from a previous manual location (see Figure 16). In the case of Fiandra *et al.*, they also developed a decision tree as a k-means model based on 36 **radiomics features** extracted from the active BM for distinguishing between compact bone and spongy bone. However, even in Figure 16b, the absence of BM segmentation can be observed in two of the three bones shown, as only the active BM is segmented.

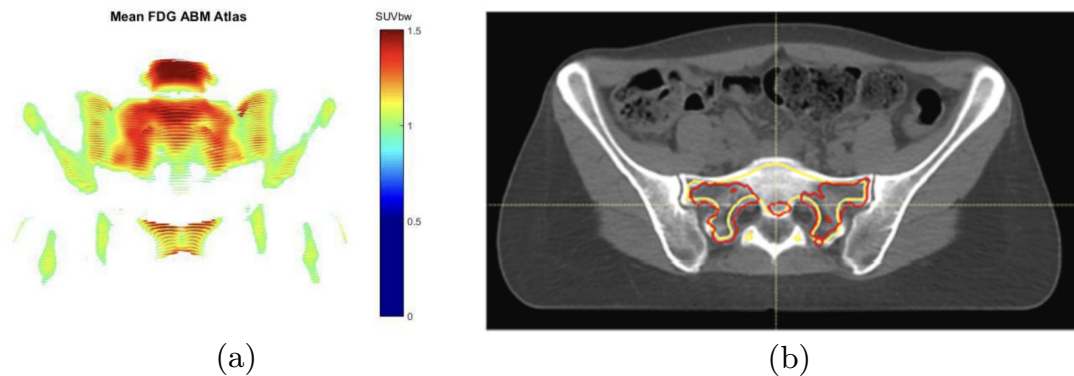


Figure 16. Examples of the (a) Li *et al.* [107] and (b) Fiandra *et al.* [108] segmentations of active BM. In (b): (yellow) ^{18}F FDG PET and (red) Decision Tree Classifier.

Other studies focused on infiltration or affectation of BM, also developed some estimation of its segmentation. In 2012, Sambuceti *et al.* [83], developed an algorithm for BM location in 102 CTs. First, they defined the highest Hounsfield Units (HU) of the image as the compact bone. Then, the algorithm defined a **threshold**, which samples a 2-voxel-thick structure and calculated its average HU. All voxels with a

HU value higher than this average value were considered compact bone, while the rest were considered trabecular bone (as shown in Figure 17). Despite the possible inaccuracies of the segmentation, the methodology developed by Sambuceti *et al.* has subsequently been used by multiple authors [109–111].

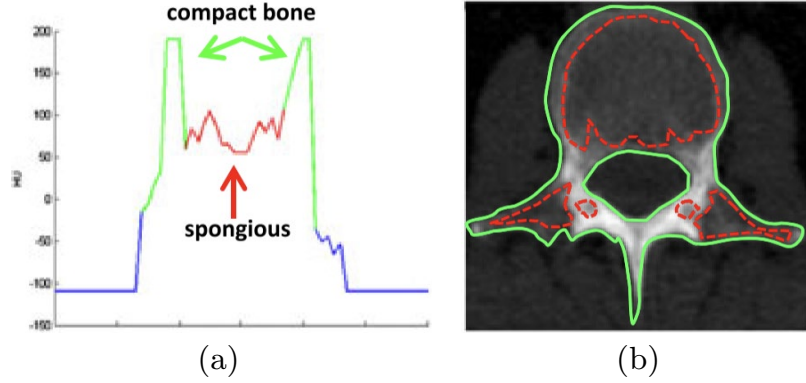


Figure 17. Segmentation of Sambuceti *et al.* [83]. (a) division of the histogram in the different regions; (b) result of the division of the histogram overlaid in the CT slice.

Following the **thresholding** methodology, in 2016 Martínez-Martínez *et al.* [84] developed simple BM segmentation for CT images as a mean for their main objective, the classification of BM involvement in MM. The algorithm used a 250 HU threshold for segmenting the compact bone. Then, for each slice, BM was found by using connected component analysis, i.e. selecting the region enclosed in compact bone. Similarly, in 2017 Nishida *et al.* [85] differentiated three structures from CT images based on the thresholds: bone for values higher than 120 HU; BM for values between -120 and -30 HU; and myeloma for values between -30 and 120 HU. Since both studies had a different objective than BM segmentation, none of them evaluate the quality of the segmentation results. Resulting segmentations are shown in Figure 18.

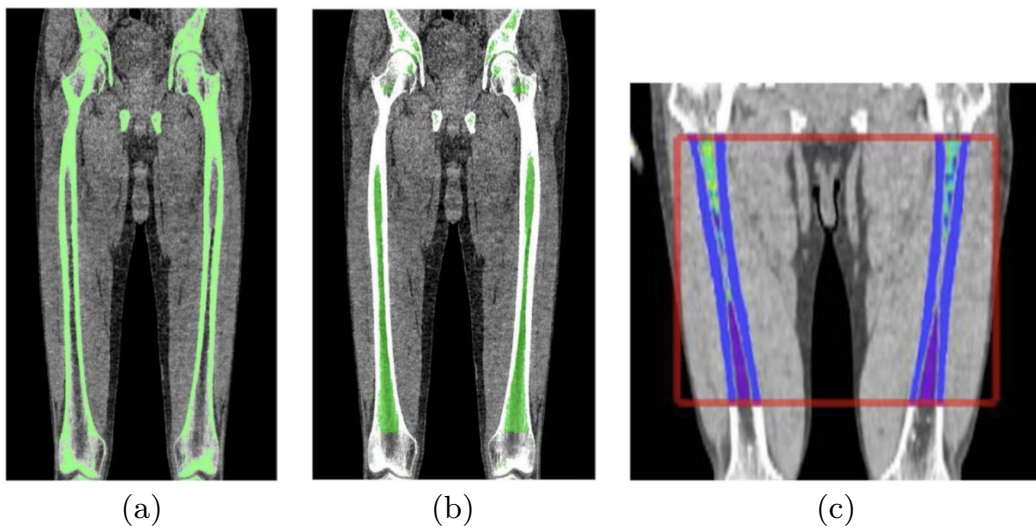


Figure 18. Results for (a) compact bone and (b) BM segmentation in Martínez-Martínez *et al.* [84] and (c) Nishida *et al.* [85] BM segmentation. For (c), red rectangle represents horizontal and vertical ranges selected for the segmentation, and the rest of colours are according to the HU: (yellow-red) high, and (blue-purple) low.

To the best of the knowledge of this thesis, only one study which develops a complete BM segmentation is found. In 2021, Wennmann *et al.* presented to the 18th International Myeloma Workshop two studies [112,113] where they segmented BM from MRI images. In the first study, they trained a nnU-net with 66 images for segmenting the BM from 30 compartments of the skeleton: humerus, femurs, sacrum, hip and vertebral bodies. In this case, they obtained a Dice index individual for each compartment, ranging from 0.76 to 0.97. In the second one, they only segmented the pelvic bones of 158 patients, using also the nnU-Net and obtaining a Dice index of 0.94.

In Table 3, all of the mentioned studies for BM segmentation are included, classified by: method of segmentation, author and year, imaging modality, segmented anatomical regions and results.

Method	Authors, year	Imaging modality	Anatomical regions	Results
Hough Transform	Perasso <i>et al.</i> [82], 2015	CT	Spinal Marrow	-
SUVmean	Li <i>et al.</i> [107], 2017			-
SUVmean + radiomics	Fiandra <i>et al.</i> [108], 2023	PET/CT	Active BM	-
	Sambuceti <i>et al.</i> [83], 2012	CT	Whole-body	-
Thresholds	Martínez-Martínez <i>et al.</i> [84], 2016	CT	Tibia and femur	--
	Nishida <i>et al.</i> [85], 2017	CT	Femur	--
nnU-Net	Wennmann <i>et al.</i> [112,113], 2021	MRI	Several bones	Dice = [0.76, 0.97]

Table 3. Resume of all studies for bone segmentation. (-) not quantitatively evaluated; (--) not quantitatively evaluated and mean for another objective.

3.2. Bone marrow quantification

In current clinical practice, the BM analysis is limited to visual interpretation, and the use of the SUV metric for activity concentration quantification, especially its maximum value (SUVmax) [114,115]. BM segmentation allows a more precise BM analysis, since metrics from the whole-body BM can be taken. This breakthrough, along with the limited visual interpretation, leads some authors to develop more precise and extended methodologies for BM quantification.

3.2.1. Quantification of bone marrow for Multiple Myeloma

Since the IMPeTUs criteria exclude SUV metrics as a MM marker [70] due to its possible affectation by inflammation processes, noise and the reconstruction method [71], the interest for other methodologies of quantification is increasing. There are some reviews [116,117], about articles which use radiomics and/or artificial intelligence in MM field. These articles differ in the final objective, within the MM context, the image modality used, and the methodology applied.

In 2021, Ripani *et al.* [118] studied the potential of radiomics features extracted from the [^{18}F]FDG PET imaging to predict the progression of SMM patients to symptomatic MM. Choosing arbitrarily 4 BM regions (vertebral bodies, iliac crests and part of the femoral diaphyses) from 45 patients, they extracted radiomics features correspondent to first order (histogram-based features), shape, GLCM, Grey Level Run Length Matrix (GLRLM) and Grey Level Size Zone Matrix (GLSZM). Then, each feature was independently studied as a potential differentiator between progression and no-progression to MM groups. Finally, they concluded that shape, second order and higher order texture features were significantly associated with progression to symptomatic MM.

Following with the differentiation of MM from other conditions, there are several studies which distinguish it from bone metastases. In 2021, Xiong *et al.* [119] implemented five ML models based on radiomics features for differentiating MM from metastasis. Based on a cohort of 107 patients with 214 MRI of the lumbar vertebrae, 104 T1-weighted and 107 T2-weighted, they extracted 282 radiomics features, corresponding to histogram variables, gradient variables, GLCM, GLRLM and autoregressive model. After a feature selection, only 13 features from the T1 image and 9 from the T2 image were selected and used for training the ML models. They demonstrated the potential of radiomics for this differentiation, since the best model reached a F1_score_macro of 0.8. In addition, they trained the model to differentiate MM from two subtypes of metastases, acquiring the best model a F1_score_macro of 0.64. One year later, in 2022, Jin *et al.* [120] developed a similar study, analysing 181 bone lesions, changing the image from which to extract the characteristics to [^{18}F]FDG PET/CT. In this case, 279 features were extracted from the same groups and adding features coming from the application of wavelets before extraction. The feature selection reduced the dimensionality to 8 features from the CT and 10 from the [^{18}F]FDG PET. Then, three different models were implemented: (1) with the CT features, (2) with the PET features, and (3) with the features of both modalities, reaching Area Under the Curve (AUC) of the Receiver Operating Curve (ROC) of 0.90, 0.93, and 0.95, respectively. In the same year, Mannam *et al.* [121] developed the same study, but taking as region of interest (ROI) a sphere centred in the position of maximum SUV. They extracted a total of 138 features

from each image modality, CT and PET. The extracted features correspond to the groups of SUV metrics and Total Lesion Glycolysis (TLG), first order, histogram, shape, GLCM, GLRLM, Neighbouring Gray Tone Difference Matrix (NGTDM) and GLSZM. After the feature selection, only 16 CT features and 14 PET features were used for the models training. The results demonstrated the potential of these features with F1_scores of 0.88, 0.61 and 0.88 for the models trained with CT features, PET features and the total of features, respectively.

Related to MM condition, there are also several studies about the prognosis prediction. In 2021, Li *et al.* [122] implemented a nomogram for overall survival prediction in MM patients. From 121 MRI scans and the manual segmentation of its lumbar region (from L1 to L5), 1316 radiomics features were extracted from each vertebral body. Apart from the usual radiomics features, i.e. those coming from first order, shape, GLCM, Grey-Level Dependence Matrix (GLDM), GLSZM, GLRLM and NGTDM, 744 wavelet features and 279 local binary pattern features were extracted. From the total of radiomics features, the vast majority were discarded based on a Cox Regression and Pearson correlation. Finally, the radiomics-based nomogram obtained a C-index (Harrell's concordance-index) of 0.81, allowing the prediction of the overall survival. In the same year, Jamet *et al.* [123] also proposed a new prognosis model for progression-free survival prediction based on random survival forest differing in the image modality. From 139 patients with [^{18}F]FDG PET/CT images, a total of 15 features were extracted: number of focal lesions, creating two groups depending on the threshold of 3 focal lesions; the Metabolic Tumour Volume (MTV); TLG; MTV and TLG from the hottest focal lesion; TLG of all focal and extramedullary disease lesions; SUVmax of a sphere manually located in some vertebral body, for representing diffuse uptake; SUVmax from vertebral body, focal lesions and extramedullary disease lesions; and some features extracted from the radiomic matrixes: GLCM, GLRLM and GLSZM. Finally, a robust model for survival prediction was obtained, with a mean log-rank p-value of 0.03 ± 0.05 .

As Jamet *et al.*, other authors used [^{18}F]FDG PET/CT radiomics features for survival prediction but combining and/or comparing with the potential of clinical features for the same aim. In 2023, Ni *et al.* [124] developed a model for prognosis MM prediction, based on clinical and radiomics features. The cohort was formed by 98 [^{18}F]FDG PET/CT images for MM diagnosis, where the location of a maximum SUV in the BM was manually selected as the volume of interest (VOI). From this VOI located in the CT and PET, 1702 radiomics features were extracted, including features from: wavelets, first order, shape, GLCM, GLSZM, Grey Level Scale Matrix, NGTDM and GLDM. The authors performed Cox Regression models based on three collections of features: (A) only clinical features, (B) only radiomics features, and (C) both of them. These models reached an AUC of 0.58, 0.65 and 0.65, respectively.

In the same year, Zhong *et al.* [125] developed a similar study. They also obtained the radiomics features (from the same groups, apart from the Gray Level Scale Matrix) from CT and PET and used clinical features. However, the study differs from the previous one in: (1) the number of models, 6 different ML models based on Cox Regression and a Random Survival Forest; and (2) in the VOI, since they used as VOI several vertebrae. In addition, they obtained remarkable better results with a maximum C-index of 0.88.

Other authors focus on improving diagnostic imaging evaluation by attempting to quantify BM infiltration, rather than prognostic prediction. The article of Martínez-Martínez *et al.* [84] from 2016 mentioned above in the subsection 3.1.2. *Bone marrow segmentation*, developed a study for evaluate BM infiltration in the medullary cavity of long bones using CT images. Since long bones are characterised by a low calcium content and in adults it contains predominantly low-density fat, they used both femurs, the largest bones. First of all, they modelled the bone density in a homogeneous population of healthy people, assuming that BM infiltration are outliers of this model. Then, for the subjects used to train and test the classifiers, they extracted two features lately normalised: the total volume detections and the total number of connected components of the detection of both femurs. Based on these features, they tried to differentiate between 3 groups, based on two tests: a blood test to differentiate between Monoclonal Plasma Cell Disorder (MPCD) positive or negative; and visual assessment of the BM infiltration in CT images by expert radiologists. The three proposed groups were: (A) MPCD- and BM infiltration negative; (B) MPCD+ and BM infiltration negative; and (C) MPCD+ and BM infiltration positive; composed by 51, 39 and 37 subjects, respectively. With two classifiers, k-Nearest Neighbours (k-NN) and Support Vector Machines (SVM), three experiments were performed: (1) A vs. C, (2) A+B vs. C, and (3) A vs. B+C, reaching an AUC of 0.99 and 0.99 for the experiment 1, 0.91 and 0.94 for the experiment 2 and 0.89 and 0.86 for the experiment 3, for each K-NN and SVM models, respectively. Other authors mentioned above for its segmentation algorithm which developed a study for BM infiltration in MM are Wennmann *et al.* [126], who implemented the BM segmentation to predict the BM plasma cell infiltration from the pelvic BM of 541 MRI scans from 270 patients. The automatic prediction model based on radiomics features reached a mean absolute error of 14.3.

Following the BM infiltration topic several authors opt for the ^{18}F FDG PET/CT imaging due to its functional nature. In 2021, Mesguich *et al.* [127], after proving the efficacy of ^{18}F FDG PET/CT imaging in bone involvement detection [63], developed a methodology for diffuse BM uptake detection from ^{18}F FDG PET/CT images of MM diagnosis. Taking MRI evaluation as GT, they implemented five ML models based on radiomics features extracted from CT and PET imaging. In addition to the

SUV metrics, 87 radiomics features were extracted from each image: 19 first order, 20 from GLCM, 16 from GLRLM, 16 from GLSZM, 12 from GLDM and 4 from NGTDM. The model with the best performance, Random Forest (RF), was characterised by a F1_score of 0.75. Similar to this study but without radiomics or artificial intelligence use, Takahashi *et al.* [128] proposed a new metric for evaluating BM involvement, with a cohort of 59 MM patients and their ^{18}F FDG PET/CT images before treatment. From the mask of the bone along with the BM and from a sphere manually located in the liver, SUVmax, SUVmean and the standard deviation of the SUV are calculated. Based on these metrics, two new metrics were proposed: the Percentage of Bone Involvement (PBI) and the Intensity Bone Involvement (IBI). The PBI is calculated as the division of the volume with uptake above the liver SUVmean (VOI) plus twice the standard deviation by the total bone mask volume. Meanwhile, IBI is calculated as the product of PBI and SUVmean of the VOI. The IBI metric showed significant differences between the groups distinguished by the number of lesions, while an inverse correlation between the level of haemoglobin and IBI and PBI was found. In addition, in 2021, the same authors developed another study [129] based on the previous one. In this case, the IBI variation between two consecutive ^{18}F FDG PET/CT images of 29 patients was evaluated. The experts in nuclear medicine classified the pair of images between: PET-remission, PET-progression and PET-stable. They observed that these groups showed a variation of IBI lower than 0, greater than 0 and approximately 0, respectively. Moreover, they found significant differences between these three groups. Despite the results of these studies, the methodology for mask creation is hugely manual, with a manual extraction of the stretcher, catheter, and areas of high physiological uptake, such as heart, kidneys and bladder. Also, the mask includes the bone, not limiting the evaluation to the BM.

As it could be seen the main topics of the articles for BM quantification in MM mentioned are focus on the diagnostic imaging (Table 4). Only the last article of Takahashi *et al.* [129] evaluated the variation before and after the treatment, but there is no mention to residual burden. However, no articles about MRD detection are found. The only two studies about this topic included in Manco *et al.* [117] were developed for the present thesis. Despite this lack of articles on the topic, it could be seen in the Table 5 how the radiomics are gaining importance in the MM field.

Context	Authors, year	Imaging modality	Anatomical region	Results
SMM to MM	Ripani <i>et al.</i> [118], 2021	[¹⁸ F]FDG PET/CT	Four arbitrary regions	Significant features in the progression
MM vs. Metastases	Xiong <i>et al.</i> [119], 2021	MRI	Lumbar vertebrae	F1_macro = 0.8
	Jin <i>et al.</i> [120], 2022	[¹⁸ F]FDG PET/CT	Bone lesions	AUC = 0.95
	Mannam <i>et al.</i> [121], 2022	[¹⁸ F]FDG PET/CT	VOI around the SUV _{max}	F1_score = 0.88
MM prognosis (survival)	Li <i>et al.</i> [122], 2021	MRI	Lumbar vertebrae	C-index = 0.81
	Jamet <i>et al.</i> [123], 2021	[¹⁸ F]FDG PET/CT	Whole-body	p-value = 0.03
	Ni <i>et al.</i> [124], 2023	[¹⁸ F]FDG PET/CT	VOI around the SUV _{max}	AUC = 0.65
	Zhong <i>et al.</i> [125], 2023	[¹⁸ F]FDG PET/CT	Vertebrae	C-index = 0.88
BM infiltration	Martínez- Martínez <i>et al.</i> [84], 2016	CT	Femurs	AUC = 0.91
	Wennmann <i>et al.</i> [113], 2023	MRI	Pelvis	Mean absolute error of 14.3
	Mesguich <i>et al.</i> [127], 2021	[¹⁸ F]FDG PET/CT	Spine	F1_score = 0.75
	Takahashi <i>et al.</i> [128], 2019	[¹⁸ F]FDG PET/CT	Bone	New metrics: PBI and IBI
Pre and post treatment comparison	Takahashi <i>et al.</i> [129], 2021	[¹⁸ F]FDG PET/CT	Bone	Groups by IBI: remission, progression and stable

Table 4. Resume of all studies for BM quantification in MM field.

Author		Clinical	SUV metrics	TLG, MTV	Gradient	First Order (Histogram)	Shape	GLCM	GLDM	GLRLM	GLSZM	NGTDM	GLSM	ARM	Wavelets	Local binary pattern	Filtering	IBI/PBI
A	Ripani <i>et al.</i> [118]					X	X	X		X	X							
B	Xiong <i>et al.</i> [119]				X	X		X		X				X				
	Jin <i>et al.</i> [120]				X	X		X		X				X	X			
	Mannam <i>et al.</i> [121]		X	X		X	X	X		X	X	X						
C	Li <i>et al.</i> [122]					X	X	X	X	X	X	X			X	X		
	Jamet <i>et al.</i> [123]		X	X				X		X	X							
	Ni <i>et al.</i> [124]	X				X	X	X	X		X	X	X		X			
	Zhong <i>et al.</i> [125]	X	X	X		X	X	X	X	X	X	X					X	
D	Mesguich <i>et al.</i> [127]		X	X		X	X	X	X	X	X	X						
	Takahashi <i>et al.</i> [128]		X	X														
E	Takahashi <i>et al.</i> [129]																	X

Table 5. Resume of the features used in the studies of Table 4. Martínez-Martínez *et al.* [84] is not included since it used bone density as the only feature. Wennmann *et al.* [113], is not included since they mentioned only radiomics, no groups. A: SMM to MM; B: MM vs. Metastases; C: MM prognosis (survival); D: BM infiltration; E: Pre and post treatment comparison. GLSM: Grey Level Scale Matrix.

3.2.2. Quantification of bone marrow for Non-Hodgkin Lymphoma

In contrast to the MM investigation, the BMI in NHL is less common, since it is not the main lesion. As mentioned before, the BM biopsy is the first-line method for detecting this condition. However, the interest in using ^{18}F FDG PET/CT is increasing. As outlined in the reviews developed by Pakos *et al.* [130], Chen *et al.* [131] and Almaini *et al.* [132], respectively, the potential of the ^{18}F FDG PET/CT assessment in the detection of BM biopsy results is demonstrated and depends on the NHL subtype.

Already in 2005, Pakos *et al.* [130] reviewed 13 studies of which 9 included patients from NHL. For the NHL group, they calculated an average sensitivity and specificity of the ^{18}F FDG PET imaging, obtaining 0.43 and 0.88, respectively. In addition, they differentiated the potential of this image modality for different NHL subtypes. For the most aggressive subtypes, 76.2% of positive biopsies were detected as PET+ too. Meanwhile, for the indolent or less aggressive subtypes, only 30.2% were correctly classified. It should be noted that only 1 (Elstrom *et al.* [133]) out of the 9 studies uses SUV metrics and not only qualitative assessment for differing between PET+/PET-. Six years later the research line remained the same. Chen *et al.* [131] reviewed a total of 8 studies which only includes NHL patients, matching 3 of them with Pakos *et al.* review. In this case, they calculated the median value of sensitivity and specificity for aggressive subtypes, reaching 0.79 and 0.94 respectively, and for indolent subtypes 0.25 and 1.

Subsequently and reaching the same conclusions than the reviews of Pakos *et al.* and Chen *et al.*, in 2022, Almaini *et al.* [132] reviewed a total of 41 articles collected since 2011, i.e. after the Chen *et al.* review. From these studies, a total of 9 used as gold standard not only de BM biopsy but also the ^{18}F FDG PET/CT follow-up. Joining all the studies by different groups, they estimated the median sensitivity and specificity for the PET BMI assessment. The evaluated groups obtained the median sensitivity and specificity of 0.77 and 0.92 for DLBCL, 0.6 and 0.81 for FL, 0.77 and 0.938 for aggressive, 0.585 and 0.851 for indolent, and 0.73 and 0.90 for all NHL patients. Furthermore, they also estimated the median sensitivity and specificity for the BM biopsy BMI assessment from those 9 articles with ^{18}F FDG PET/CT follow-up as gold standard. The same groups obtained sensitivity values of 0.47, 0.81, 0.94, 0.58 and 0.56, respectively, and a value of 1 for specificity in all cases except for FL, which was not reported in any of the articles. Although the similarity of the results with the previously mentioned reviews, 18 out of the 41 articles included a quantitative assessment of the ^{18}F FDG PET imaging, being the most remarkable Adams *et al.* [134,135] with SUV metrics and Aide *et al.* [136] for DLBCL based on radiomics features. In these three reviews, the sensitivity results remark the better

potential for aggressive subtype, than indolent ones. In addition, the outstanding specificity results defined the potential for discarding negative cases.

As SUV metrics are the most standardised methodology for measuring [^{18}F]FDG PET imaging, several studies propose methodologies based on these metrics for BMI detection. In 2003, Elstrom *et al.* [133] differentiated BMI groups using SUVmax. They defined the [^{18}F]FDG PET scans as positive when the maximum SUV of a located lesion was superior to 2.5. Then, they calculated the sensitivity and specificity differentiating in NHL subtypes, obtaining 0.60 and 0.95 for DLBCL and 0.13 and 1 for FL, respectively.

Then, in 2014, Adams *et al.* [134] applied two different quantification methods for FL patients. The first method was based on considering PET+ all the studies with a BM FDG uptake higher than the liver FDG uptake, demonstrating a sensitivity and specificity of 0 and 1, since all the cases were reported as PET-. The second method consisted in extracting a VOI in the shape of a sphere located in the right posterior iliac crest and calculating the mean, maximum and peak SUV values. These three metrics were demonstrated to be a better quantification method, with sensitivities and specificities of 0.75 and 0.86, 0.88 and 0.86, and 0.88 and 0.86, respectively. The same authors implemented the same methodology in other study [135] for DLBCL patients. In this case, only a sensitivity of 0.69 was reported for the first method. No specificity was reported since the authors determined that a BM biopsy with a negative result cannot discard a BMI. In addition, they performed a progression-free survival analysis based on the second method of quantification, demonstrating that a positive BM biopsy implied a worse survival than a negative one, but that there was no significant difference in survival regardless of the outcome of the PET scan.

Three studies with the aim of using SUV metrics for differentiating BMI+/BMI-groups, but not included in the previous reviews are El-Najjar *et al.* [137], Perry *et al.* [138] and Asenbaum *et al.* [139]. In 2014, El-Najjar *et al.* [137] quantitatively evaluated the potential of [^{18}F]FDG PET imaging for identifying diffuse BM involvement in FL. In this case, the SUVmax at mediastinal blood pool, spleen, liver and four skeletal regions (sternum, both iliac crests and T12 vertebra) were calculated. From these values, a total of 4 metrics were acquired: (1) SUVmax from all the mentioned locations; (2) the average SUVmax of the four skeletal regions; (3) the ratio between metric (2) and the SUVmax at the mediastinal blood pool; and (4) the ratio between metric (2) and the SUVmax at the liver. From all these metrics, a cut-off value based on the ROC curve for distinguishing between PET+ and PET-groups was calculated. Then, sensitivities and specificities for each metrics were calculated: 0.31 and 0.92 for (1), 0.58 and 0.96 for (2), 0.75 and 0.83 for (3), and 0.83 and 0.83 for (4). Then, in 2016, Perry *et al.* [138] developed a study to determine the

best threshold of SUVmean for differing between BMI+/BMI- groups in FL patients, similar to the study of Elstrom *et al.* From a cohort of 68 patients, they extracted the SUVmean from the posterior iliac crest in the 35 patients with no increased uptake and the 25 patients with diffusely increased uptake, and from the positions with a focally increased uptake in the rest of cases. They demonstrated that a SUVmean higher than 2.7 obtains a sensitivity and positive predictive value of 0.69 and 0.82, respectively, while a threshold of 1.7 obtains a negative predictive value of 1. This way, they limited the doubtful cases to those with a SUVmean ranged between both thresholds. According to this study, with SUVmax instead of SUVmean, Asenbaum *et al.* [139] proposed a methodology with the same aim but different methodology, regions to evaluate and cohort. From a cohort of 60 patients of HL and NHL, they extracted VOIs corresponding to some vertebrae, from the sacrum and from the right side of the iliac crest. In addition, for focal uptake, they extracted all the pixels above 50% of the SUVmax of the lesion. From all of these VOIs, they extracted the maximum SUVmax. Then, they extracted the SUVmax of the liver as reference, in order to calculate the ratios lesion-to-liver. For determining the current BMI state, they considered as BMI+ when one of the tests (BM biopsy or [^{18}F]FDG PET visual assessment) were positives. Having all measured, they observed different ratios to normal, diffuse and focal uptake: 74.8%, 100.6% and 324.1% respectively. From these measures, they calculated the cut-off value of the ratio lesion-to-liver for detecting a no-normal uptake, obtaining 95.25% which, tested in the cohort, reached an AUC, sensitivity and specificity of 0.87, 0.88 and 0.86, respectively.

In spite of the utility of SUV metrics for this quantification, other authors opt for radiomics analysis, trying to study not only the conventional metrics, but also the textural features. In 2018, Aide *et al.* [136] proposed a quantification method for DLBCL subtype based on radiomics features. They extracted a VOI through a 150 HU threshold over the CT scan and a manual selection of the spine and pelvis. From this VOI, they extracted radiomics features correspondent to first order, GLCM and GLSZM. For the GT of the 82 patients, they established two rules: (1) one or several bone focal uptakes are considered PET+, while diffuse or heterogeneous uptake is not; (2) a patient with BM biopsy and PET negatives is considered as negative, while one of them as positive is enough to consider the case positive. A ROC analysis was performed on every feature in order to determine the one with the highest ROC, i.e. skewness with an AUC of 0.82. From this variable, a cut-off value was estimated and used to differentiate positive and negative groups, reaching a sensitivity of 0.82 and a specificity of 0.82. They also used this feature to establish survival differences between groups, demonstrating a significant difference between both groups, with a progression-free survival in 2 years of 88.8% and 70.8% for negative positive BMI.

Following the study, in 2024, Faudemer *et al.* [140], the same authors, presented a new metric based on radiomics features extracted from ^{18}F FDG PET imaging to predict the BMI in FL patients. From 66 patients, they proposed a similar study than Aide *et al.*: (1) same criteria for differing between BMI+/BMI-; and (2) extracting of radiomics features, including variables from SUV metrics, GLCM, GLSZM, and NGTDM. However, they developed a new metric (named pred.score) using a LASSO regression. This metric was obtained as a weighted sum of three GLCM variables (variance, correlation and joint entropy) and one NGTDM (busyness). With this metric, they reached an AUC of 0.82, sensitivity of 0.7 and specificity of 0.83, demonstrating the potential of these features and the resultant metric.

The studies mentioned evaluate the differentiation between BMI+ and BMI-. However, the heterogeneous and diffuse uptake pattern of the BMI observed in ^{18}F FDG PET scans is also seen in patients affected by inflammation or infection as a response to the need for the overproduction of immune cells, since the BM is the main hematopoietic organ [9,141–143]. No studies have been identified that differentiate this inflammatory status from those previously mentioned.

On the other hand, although many studies which differentiate BMI are mentioned, most of them perform no or very poor quantification based exclusively on SUV metrics. Only a few use more complex methodologies [134–136,140]. For this reason, it is worth mentioning two studies that, within the field of NHL quantification, implement these more complex methodologies but for different purposes. In 2024, Chen *et al.* [144] developed seven ML models based on radiomics features and conventional PET parameters to predict the treatment efficacy in DLBCL patients. Since they were not focused on the BM, they interpreted as VOI everything with an uptake higher than 41% of the SUVmax in the PET image. From the VOIs of 177 patients, they obtained 110 features: SUVmax, MTV, TLG and radiomics features. The trained ML models reached a maximum AUC of 0.74 for predicting mid-term efficacy of the treatment. In the same year, Häggström *et al.* [145] implemented different variants of a ResNet34 DL model for locating hypermetabolic tumour sites for HL and NHL. They considered the case as positive when one or more lesions had an FDG uptake higher than the liver. Confirming the classification, expert physicians revised the clinical reports and the ^{18}F FDG PET/CT scans to discard those cases marked as positive by infection or inflammation. They trained the model for different databases obtaining MIP or 3D volume from PET, CT or both images. Comparing all the models, the one with the best performance, measured in AUC, was the one based on the average probability of sagittal and coronal PET MIP. This model reached an AUC, sensitivity and specificity of 0.95, 0.87 and 0.95, respectively.

All the studies about the quantification in the NHL mentioned are included in the Table 6. The thick border in the middle of the table separates studies with different objectives. The continuous thinner one separates studies by the methodology applied.

		Author, year	Methodology	Subtype	Results
BMI	SUV metrics	Elstrom <i>et al.</i> [133], 2003	SUV _{max}	DLBCL FL	S = 0.6 and Sp = 0.95 S = 0.13 and Sp = 1
		Adams <i>et al.</i> [134], 2014	SUV _{mean} , SUV _{max} , SUV _{peak}	FL	S = 0.75 and Sp = 0.86 S = 0.88 and Sp = 0.86 S = 0.88 and Sp = 0.86
		Adams <i>et al.</i> [135], 2014	SUV _{max} (liver)	DLBCL	S = 0.69
		El-Najjar <i>et al.</i> [137], 2014	SUV _{max}	FL	S = 0.83 and Sp = 0.83
		Perry <i>et al.</i> [138], 2016	SUV _{mean}	FL	S = 0.69, PPV = 0.82 and NPV = 1
		Asenbaum <i>et al.</i> [139], 2018	SUV _{max} (lesion-to-liver)	HL and NHL	AUC = 0.87, S = 0.88 and Sp = 0.86
	Radiomics	Aide <i>et al.</i> [136], 2018	Skewness (histogram)	DLBCL	S = 0.82 and Sp = 0.82
		Faudemer <i>et al.</i> [140], 2024	New metric: pred.score	FL	AUC = 0.82, S = 0.7 and Sp = 0.83
		Chen <i>et al.</i> [144], 2024	ML models	DLBCL	AUC = 0.74
		Häggström <i>et al.</i> [145], 2024	CNN	HL and NHL	S = 0.87 and Sp = 0.95
Other					

Table 6. Resume of all studies for quantification in NHL field. The first column indicates the aim of the study, the second one indicated the main methodology applied. No image modality is mentioned since all the studies use [¹⁸F]FDG PET/CT imaging. S: sensitivity, Sp: specificity, PPV: positive predictive value, NPV: negative predictive value.

4. Materials and methods

Following the present thesis objective: “*Proposal and development of [^{18}F]FDG PET/CT image processing methodologies to quantify bone marrow in diseases such as Multiple Myeloma or non-Hodgkin's lymphoma as an aid to clinical decision making*”, a complete methodology for the quantification is proposed. With this purpose, a segmentation method for obtaining an adequate BM segmentation for its posterior quantification is presented. For the quantification methodology, two different approaches are also proposed. This methodology is applied for multiple clinical contexts encompassing two pathologies:

For MM clinical context:

- I. MRD detection in [^{18}F]FDG PET/CT images acquired for the assessment of the treatment response.
- II. MRD prediction from [^{18}F]FDG PET/CT images acquired for the MM diagnosis, in order to evaluate the potential of the quantification methodologies in predicting patients with more possibilities of relapse.
- III. SMM risk of progression to MM (low, intermediate and high risk), using [^{18}F]FDG PET/CT images of the SMM diagnosis.

For NHL clinical context:

- IV. BM status in NHL patients using [^{18}F]FDG PET/CT images acquired for the NHL diagnosis. In addition to differentiating the classes for NHL, the methodology is also applied dividing the NHL database between its subtypes, due to the differences observed in the BM status between indolent and aggressive subtypes.

For obtaining the quantification of the BM in different clinical contexts, an adequate BM segmentation is essential. First, an automatic segmentation based on image processing algorithms, such as thresholding and morphological operations, is developed. Once the BM masks are obtained, they are manually revised with a platform implemented in the present thesis. This way, a semi-automatic segmentation is achieved by combining the automatic segmentation and the manual revision. Then, a cohort of CT scans from MRD and NHL contexts are semi-automatically segmented to obtain a GT of the segmentation masks, in order to train a DL model for replacing the semi-automatic segmentation by an automatic one. Having a model for automatic BM segmentation, the different databases corresponding to different clinical contexts are used to evaluate the potential of quantifying the BM and its clinical status by means of univariate analysis, ML

models and/or DL models, depending on the characteristics of the clinical context and its database. The workflow followed for obtaining the BM quantification is shown in Figure 19.

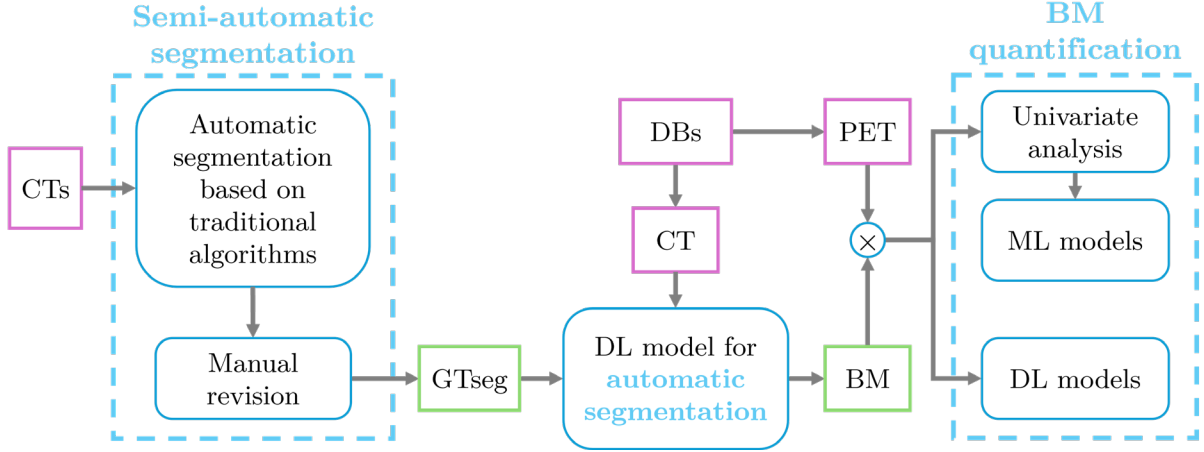


Figure 19. Workflow for the BM quantification. GTseg: GT for the segmentation. DDBB: database of a concrete clinical context.

4.1. Materials

The study cohort for the present thesis is formed by ^{18}F FDG PET/CT images from two different protocols differing in the clinical context: the MM, MRD and SMM diagnostic protocol (whole-body) and the NHL diagnostic protocol (femur-to-head). Siemens Biograph TruePoint 6 PET/CT (Siemens Healthineers, Erlangen, Germany) is used to obtain ^{18}F FDG PET/CT studies at the Nuclear Medicine Department of the Hospital Universitario 12 de Octubre in Madrid. These images are acquired based on the European Association Nuclear Medicine (EANM) procedure guidelines [146]. An intravenous weight-adjusted dose of ^{18}F FDG was injected into the subjects. Sixty minutes later, whole body PET images were acquired with an emission time of 3 minutes per bed position. Random, scatter and attenuation corrections were performed. Reconstructed PET images had a matrix size of 168×168 with a voxel size of $4.0728 \times 4.0728 \times 5 \text{ mm}^3$. Additionally, CT images were obtained using helical CTs (120-140 kVp, 25-170 mAs) with a resolution of 512×512 with a voxel size of $0.9766 \times 0.9766 \times 2.5 \text{ mm}^3$.

Along with the images, their respective visual evaluations conducted by nuclear medicine experts are included in the cohort. These evaluations depend on the clinical context where the BM is assessed. For the MRD (detection and prediction) there are only two classes: MRD+ and MRD-. For the NHL, there are three classes: BMI-, BMI+ and reactive BM (rBM), i.e. when the BM is overproducing immune cells due to infections, inflammations and similar conditions. In addition, the haematological BM assessment based on MFC of the BM biopsy is available in most of the patients,

and the classes in which this test classifies coincide with the ones of the visual assessment.

In the case of the SMM clinical context, there are also three groups depending on the level of risk of developing symptomatic MM: low, intermediate and high. Instead of using the visual assessment or the result of the BM biopsy, there are two types of labelling based on two different criteria: the IMWG labelling [13,24] and the PETHEMA labelling based on the GEM model [26,147].

The collection of all images and data is performed in accordance with the ethical principles of the ‘Declaration of Helsinki’ and ‘Ley Básica de Protección de Datos 15/1999’. In addition, the approval of the Clinical Research Ethics Committee of the Hospital Universitario 12 de Octubre has been obtained for the different studies, with codes 25/002 and 24/363.

4.2. Bone marrow segmentation

In this section, the automatic segmentation of the BM is explained, starting from an approach based on traditional methods, following with a platform implemented for the manual revision, and finishing with a DL model for the automatic segmentation as shown in Figure 19. This way, a semi-automatic segmentation is used for obtaining a GT for an automatic segmentation. For both automatic BM segmentations, first a model for bone segmentation is performed for several reasons:

1. As observed in the section 3.1. *Bone and bone marrow segmentation*, the vast majority of the BM segmentation studies start with the segmentation of the bone for then, removing the compact bone, obtaining the BM.
2. As is also observed in the section 3.1. *Bone and bone marrow segmentation*, several studies obtain outstanding results for bone segmentation, meanwhile those focused on BM segmentation are very limited. Only Wenmann *et al.* [112] obtains acceptable results, however their model is not tested in the whole-body and it is not reproducible since there is not enough information in the published article.
3. The manual revision of the masks of bone (aprox. 2-4h) is faster than the BM (aprox. 6-12h), starting both revisions from the masks obtained with the process described in the subsection 4.2.1. *Semi-automatic segmentation based on traditional algorithms*. The main reason for a slower segmentation is that, in some regions with a thin compact bone layer or with spinal canal, the process of removing this thin layer is time-consuming. This way the obtention of an acceptable number of GT samples for bone segmentation is faster than for BM segmentation.

It must be mentioned that all of the segmentation methodologies developed in the present thesis are limited to the regions from knee-to-head, excluding head, knees, elbows, and the rest of bone tissue below these joints, due to several reasons: (1) the NHL diagnostic protocol is limited from the femur to the head, since it focuses on the main lesions; (2) distal bone movement can cause segmentation errors; and (3) the ^{18}F FDG uptake of the brain can cause mistakes in the BM quantification.

4.2.1. Semi-automatic segmentation based on traditional algorithms

As it has been mentioned, the first step of the methodology is the BM segmentation based on traditional image processing algorithms, which is completely developed in MATLAB R2023b (The MathWorks Inc., Natick, MA, USA) [148]. In Figure 20, the steps for obtaining the BM mask are shown. First of all, CT DICOM files are converted into a CT volume through a tool developed in MATLAB called Statistical Parametric Mapping (SPM12) [149]. Posteriorly, the methodology is based on 4 main steps which are described in detail below: (1) stretcher removal, (2) bone segmentation, (3) spinal canal subtraction and (4) compact bone removal. It should be noted that, following the studies included in the subsection 3.1.2. *Bone marrow segmentation*, the BM segmentation is planned to be obtained after the removal of the compact bone from the bone segmentation.

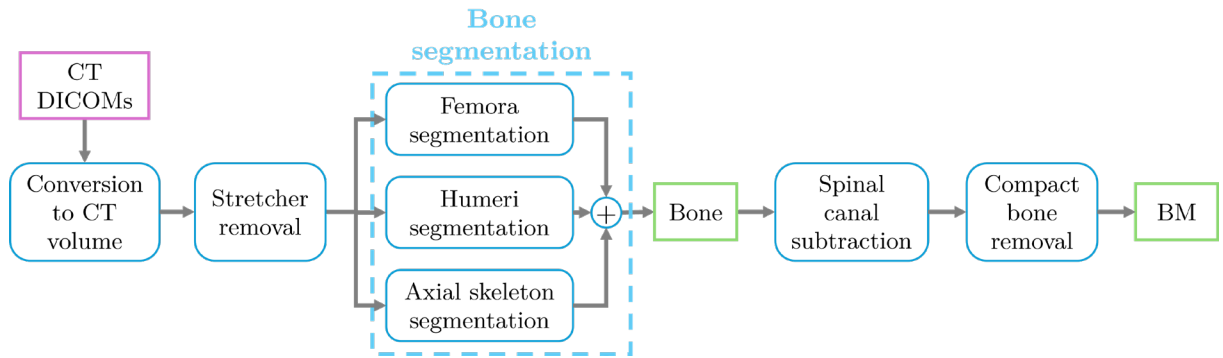


Figure 20. Pipeline of the automatic segmentation based on traditional image processing algorithms. Axial skeleton represents the bone segmentation from the torso, i.e. from neck to pelvis, excluding limbs.

4.2.1.1. Stretcher removal

For the stretcher removal, a threshold of 98 HU is applied to the original CT volume, remaining every voxel over this threshold (Figure 21a). Then, the resulting mask is filled, in order to segment not only the edges of the stretcher, but also the interior part (Figure 21b). Since the bone has HU over 300, it is included in the thresholded mask. For discarding the voxels corresponding to the skeleton, the mask is labelled into any volumetric component, and only the component with the greatest volume,

i.e., the highest number of voxels, is considered stretcher. Since in some cases, the skeleton could be the component with more voxels (Figure 21c), it is checked that the position of the component is the closest to the left edge of the image. If it is not, the second component with more voxels is selected as stretcher (Figure 21d).

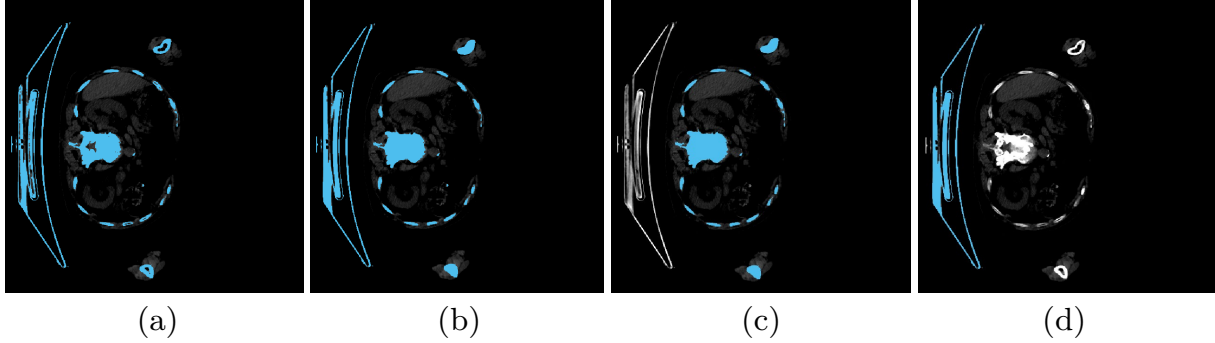


Figure 21. Stretcher removal process represented over a CT thresholded between 0 and 300 HU for a better visualization. The mask of the stretcher correspondent to each step is shown in blue. (a) stretcher mask as a result of a threshold of 98 HU in the CT; (b) mask of (a) after the filling operation; (c) mask after selection of the greatest volume; (d) correction of the volume selected based on the position.

4.2.1.2. Bone segmentation

The first step of the segmentation consists in removing the stretcher from the CT, by considering as 0 all the voxels identified as stretcher. Then, the skeleton segmentation is divided in three different sub-steps according to the anatomical regions: (1) femurs segmentation, (2) humeri segmentation, and (3) axial skeleton segmentation. As mentioned before, the distal bones of the extremities are excluded from the segmentation. In addition, the calotte is also discarded by converting to 0 the superior 16% of the CT scan.

Femurs and humeri segmentations are performed with a collection of traditional image processing algorithms: thresholding, morphological operations and flood-fill operations. For the **femurs**, first, a threshold between 196 and 235 HU, depending on the number of slices of the CT volume, is applied in the CT scan with the stretcher removed. Then a morphological close operation with a structuring element with disk shape of radius of 1 pixel is applied in every slice. Finally, the holes are filled in order to include not only the compact bone, but also the BM. This filling operation is applied in the slices of the lower part of the image (inferior 30% of the image). However, in the cases of whole-body protocol, i.e. scans with the knee to the feet included, detected by having more than 537 slices, the filling operation is applied in the slices of the second third of the scan. Finally, for smoothing the obtained segmentation, a morphological open operation with a structuring element with disk shape of radius of 9 pixels is applied slice-by-slice. Since the lower part is the only one filled, the volumetric component check performed is applied in the resultant

part, being both femur the first and second component with the greatest volume. Then, for the next skeleton segmentation sub-steps, the edges of the femurs are obtained, i.e. the limits for each direction: width, length and height (number of slices).

For the **humeri**, the inferior 60% of the CT scan with the stretcher removed is discarded, i.e. converted to 0. Then, the CT scan is filled from the superior limit of the femurs to complete height of the image. Then, the same smoothing process than for the femurs is applied with a structuring element of disk shape with a radius of 7 pixels in the mentioned slices. Then, the two volumetric components of highest number of voxels are selected. Again, the limits of humeri are calculated. For some cases of the femur-to-head protocol, the humeri are located over the head, making impossible humeri location. In other cases, because the arms are located over the patient body, the regions from the elbow to the hand are wrongly included.

For the rest of the skeleton, i.e. the **axial skeleton**, the CT with the stretcher removed is thresholded between 0 and 200 HU. Then, two different thresholds are applied, 102 and 124 HU: one for the scapula and ribcage segmentation, and the other for the rest of structures. A median filter is applied in order to remove regions with HU higher than the defined thresholds. To this filtered mask, a morphological close operation and a flood-fill operation are consecutively applied to avoid small and incorrect structures. The limits of femurs and humeri are removed from the mask, to apply the volumetric component step for discarding every structure not contained in the greatest volume. Finally, a morphological close operation with a structuring element with disk shape of radius 5 pixels is applied in every slice to smooth. The radius is smaller than the ones used for femurs and humeri since there are small structures such as ribs. Examples of slices and the total volume where the structures are differentiated by colours can be found in Figure 22.

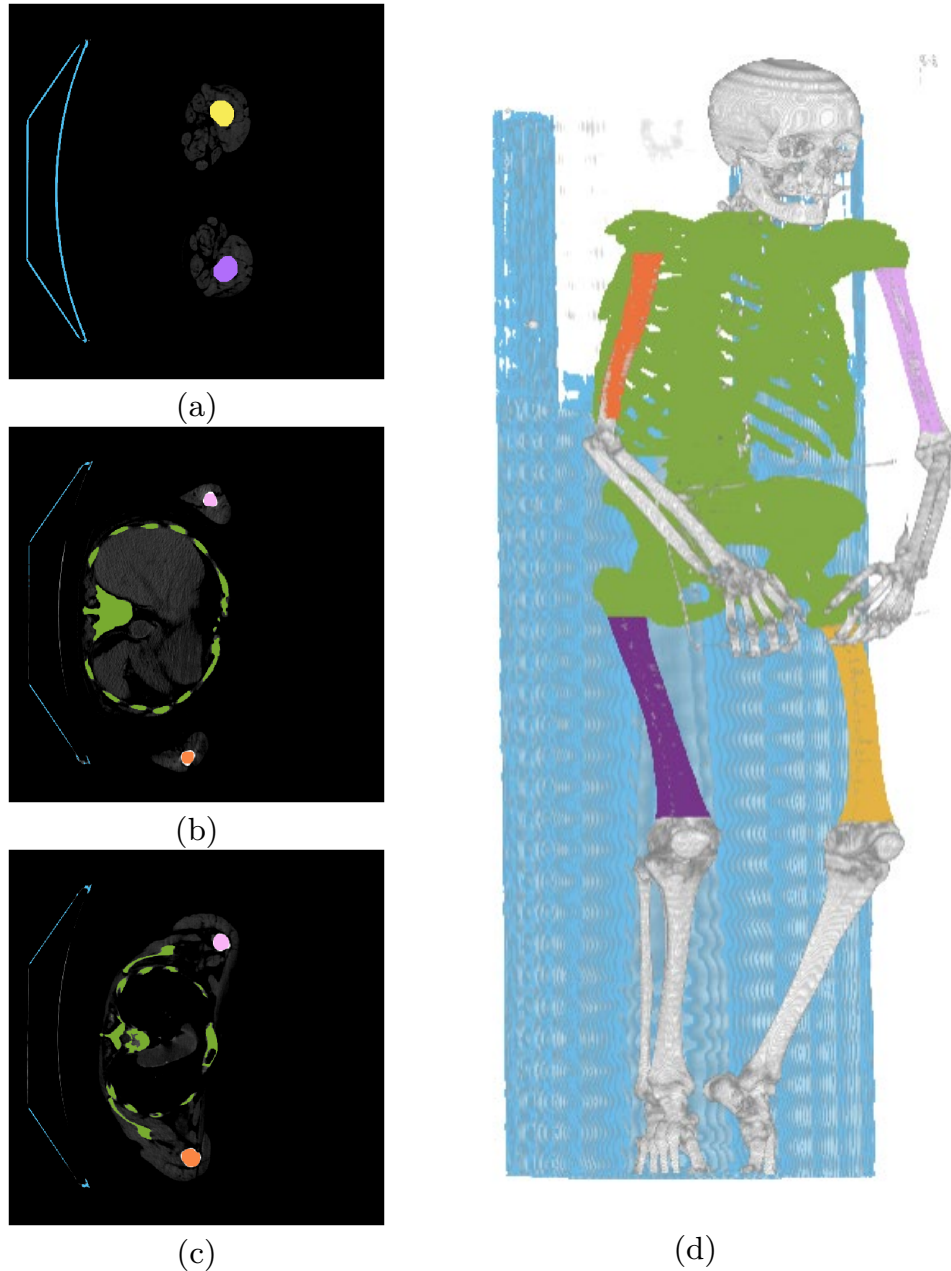


Figure 22. Representation of the whole bone segmentation, with different colours for each segmented structure: (blue) stretcher, (purple) right femur, (yellow) left femur, (orange) right humerus, (pink) left humerus, (green) axial skeleton. Examples of slices at different heights (a) femurs region; (b and c) ribs region. (d) complete volume over thresholded CT differing by the same colours.

Once the three individual volumes are obtained, femurs and humeri are reassembled to the axial skeleton by addition. To the resulting bone mask, three consecutive flood-fill operations on the sagittal, coronal, and axial plane are applied, in order to fill any possible holes. Then, the surface of the bone mask is smoothed by a median filter in 3D. A morphological close operation with a structuring element of disk shape of radius depending on the number of slices (between 11 and 17 pixels) on the area corresponding to the sacrum is applied taking into account its low HU compared to

the rest of the axial skeleton. Figure 23 shows the bone mask reassembled and after the flood-fill operations.

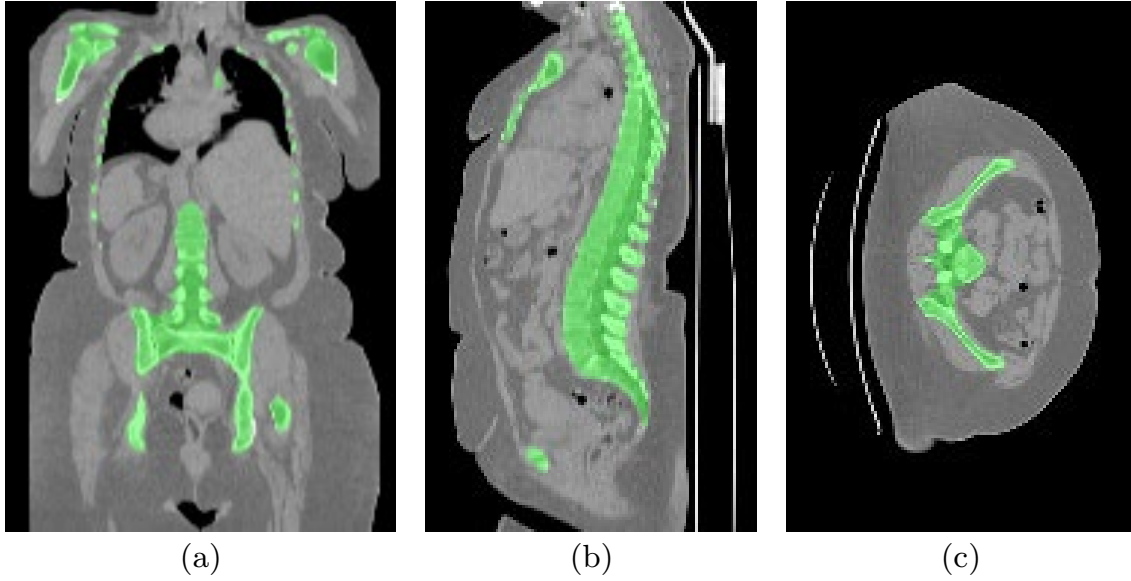


Figure 23. (a) Coronal, (b) sagittal and (c) axial planes of the bone mask (green) over the CT.

4.2.1.3. Spinal canal subtraction

As seen in Figure 23, due to the multiple flood-fill operations, the spinal canal is segmented as part of the bone. The spinal canal subtraction process starts calculating the mean HU value of the CT of the bone mask voxels. The CT image is thresholded with a value of 60 or 77 HU depending on the mean calculated, whether it is lower or equal, or higher than 230 HU, respectively. This threshold is applied considering that canal HUs are lower than those of bone tissue. In addition, any pixel not represented in the previous bone mask, is deleted from the thresholded spinal canal mask. Then, small objects around the canal or inside some vertebrae are removed. Finally, a loop for distinguishing canal from inner vertebrae, based on a morphological close operation is applied. The spinal canal mask (Figure 24abc) is then subtracted from the skeleton mask obtained in the previous step (Figure 24def).

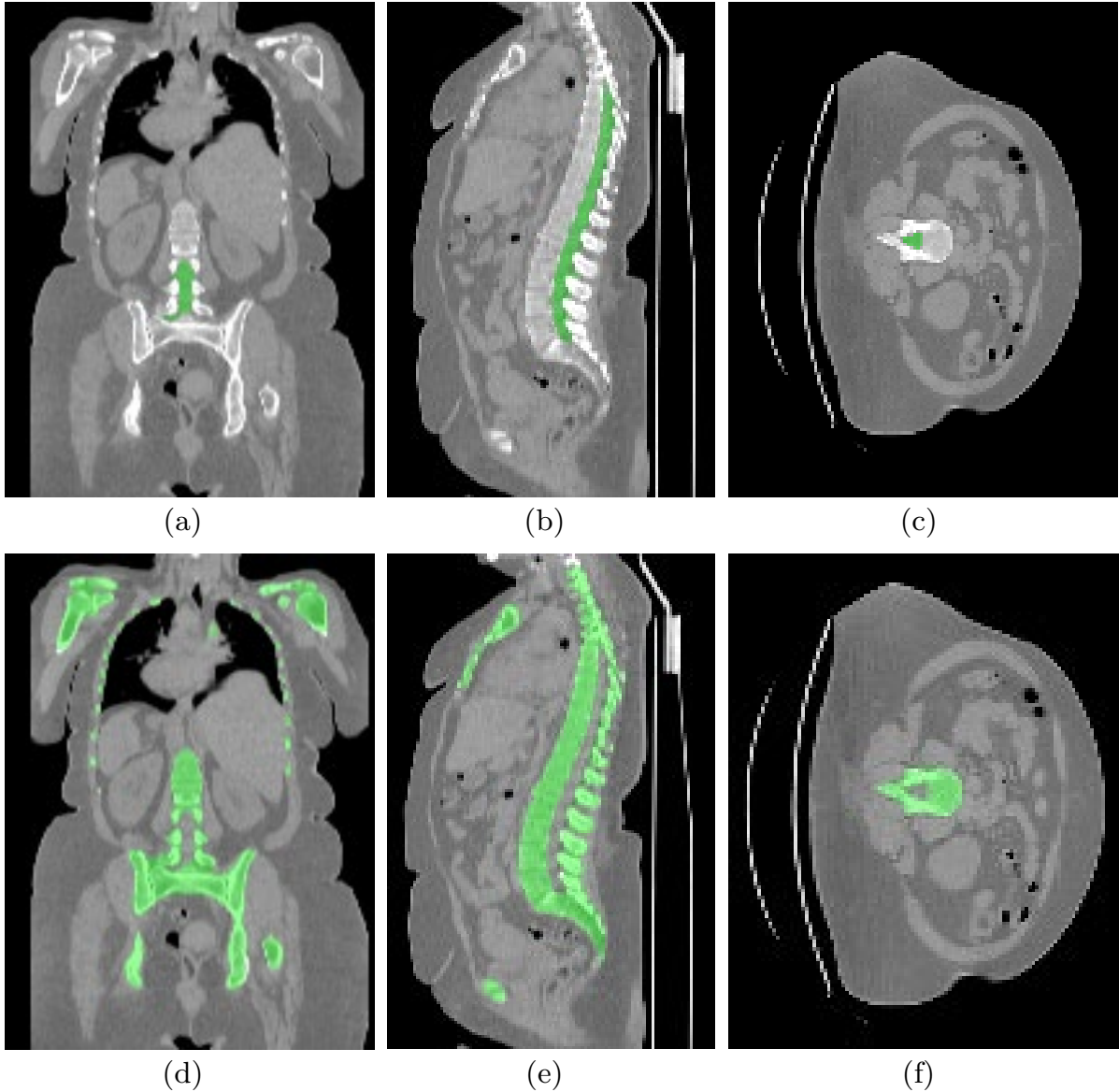


Figure 24. (a) Coronal, (b) sagittal and (c) axial planes of the spinal canal mask (green) over the CT. (d) Coronal, (e) sagittal and (f) axial planes of the bone mask (green) after the spinal canal subtraction over the CT.

4.2.1.4. Compact bone removal

Finally, the compact bone is removed from the bone mask without the spinal canal in order to obtain the BM contained inside. To extract the BM in humeri and femurs a single threshold is applied followed by inverting the resulting mask. Then the selection of the two greatest volumes to avoid including small wrong structures is applied. For smoothing the results, a morphological close operation with a structuring element with disk shape of radius 4 and 7, for femurs and humeri respectively, is performed. For the axial skeleton voxels, an iterative subtraction with 5 repetitions drawing from the mask perimeter is applied. The resulting masks are again assembled into the final BM mask. Figure 25 shows the final result.

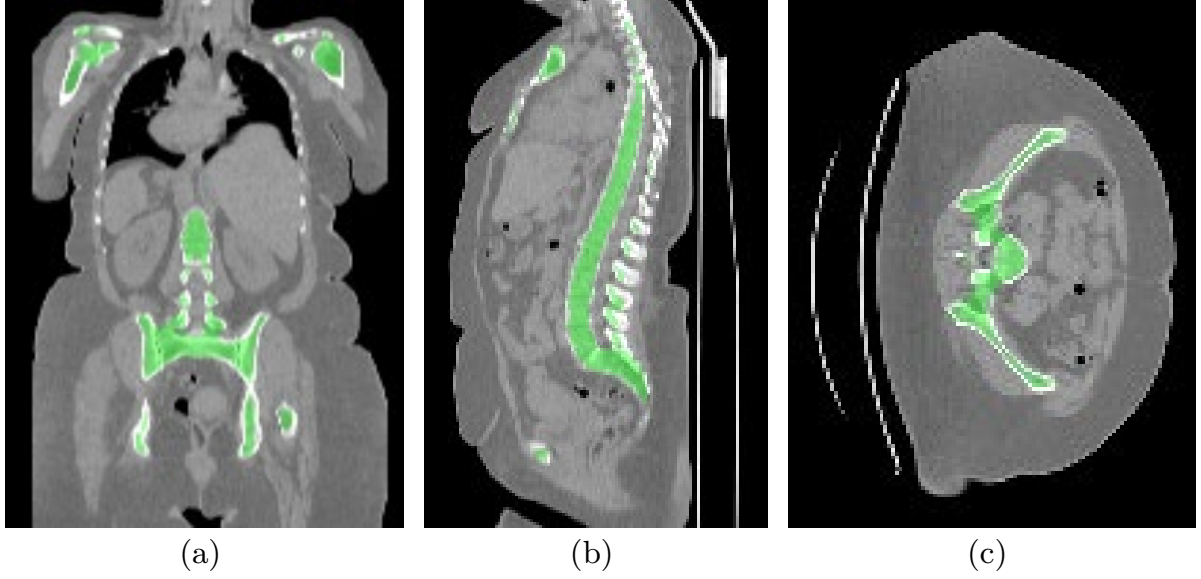


Figure 25. (a) Coronal, (b) sagittal and (c) axial planes of the BM mask (green) over the CT.

4.2.1.5. Manual revision

Once only the BM is present in the mask, the masks of the patients are manually revised using the platform implemented for the present thesis, which is explained in the subsection 5.1.1. *Platform for bone and bone marrow segmentation* The manual revision of the results is a time-consuming process, especially for the BM. After the manual revision of the masks in the platform developed in MATLAB R2023b (The MathWorks Inc., Natick, MA, USA) [148], the rest of segmentation and quantification methodologies are implemented in Python 3.6 [150] and Python 3.9 [151], depending on the necessary operation.

4.2.2. Automatic segmentation based on Deep Learning

As mentioned before, both automatic segmentation methodologies are planned to be developed first for the bone and then, for the BM mask. In the case of the methodology described previously in subsection 4.2.1. *Semi-automatic segmentation based on traditional algorithms*, the BM is obtained by means of the compact bone removal from the bone mask. However, in the case of the DL model approach, the usefulness of training the bone segmentation model first lies in: (1) the speed of obtaining GT from bone versus BM samples; (2) the results which in the first run are expected to be better for bone than for BM; (3) the confidence in the potential of the U-Net model for bone segmentation based on the studies of Klein *et al.* [78] and Noguchi *et al.* [79]; and (4) the possible independence of the compact bone removal process.

Considering these reasons, Figure 26 explains why the bone segmentation model is necessary for the different approaches of BM segmentation (posteriorly detailed in

the subsection 4.2.2.2. *Bone marrow segmentation*). In the first two approaches (marked as ①), the pre-processing defined for the bone segmentation is adapted for the pre-processing necessary for the BM segmentation. In addition, for this approach, depending on the database selected, some of the skeleton utilised for obtaining the BM mask used as GT are obtained through the bone segmentation DL model. In the last approach (marked as ②), the bone segmentation model is retrained for BM segmentation model, benefiting from the transfer learning nature of DL models.

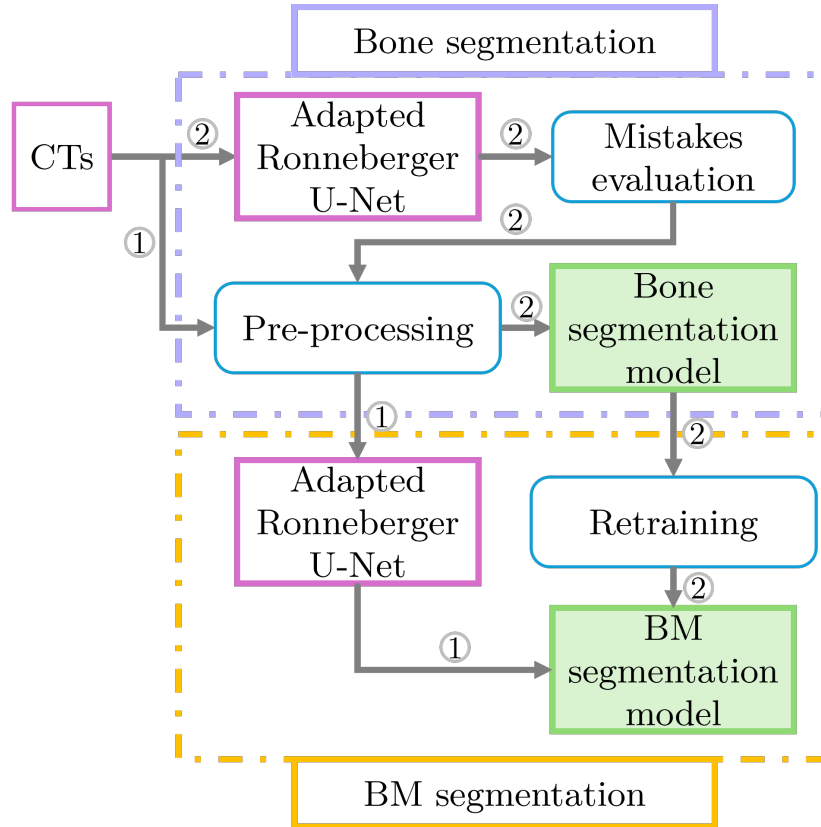


Figure 26. Simplified pipeline for obtaining the BM segmentation model. ① BM segmentation based on an untrained U-Net model; ② BM segmentation based on retraining the bone segmentation model.

4.2.2.1. Bone segmentation

In the Figure 27, the pipeline to obtain an optimal pre-processing is shown. First the bone segmentation model is trained and tested with unprocessed data. Once the mistakes of this segmentation are evaluated, the pre-processing steps can be defined. However, for understanding the results of the unprocessed data model, first the architecture of the model and the dataset division for train and test must be explained. Both steps are the same for both, the model with unprocessed data and for the model with the pre-processed data.

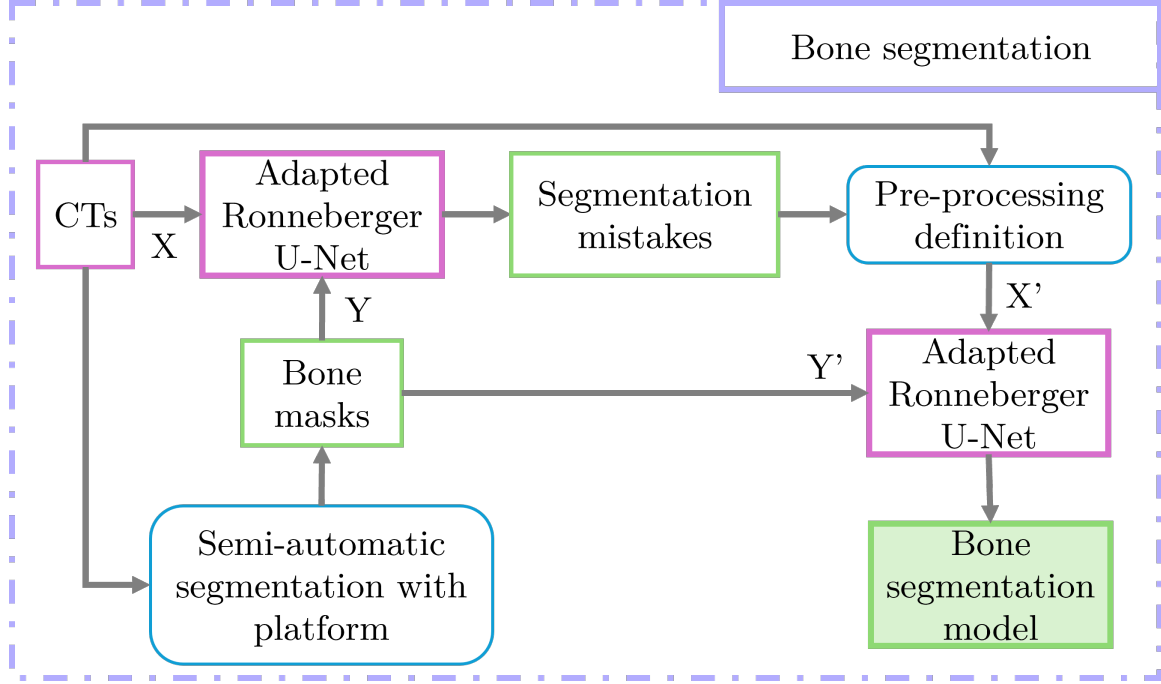


Figure 27. Pipeline for the pre-processing steps of the bone segmentation model. X: unprocessed data for train set; Y: unprocessed data for test set; X': pre-processed data for train set; and Y': pre-processed data for test set.

4.2.2.1.1 Model architecture

As mentioned before in subsection 3.1.1. *Bone segmentation*, the model of Ronneberger *et al.* [102] appears to be the best option for segmenting the bone from the CT slices. This model is a CNN designed for biomedical image segmentation tasks called U-Net due to its architecture, since it is composed of a contracting path to capture the context of the structures to segment and a symmetric expanding path for precise localization, giving to the architecture a U shape, as seen in Figure 28. It is characterised by its ability to handle limited training data.

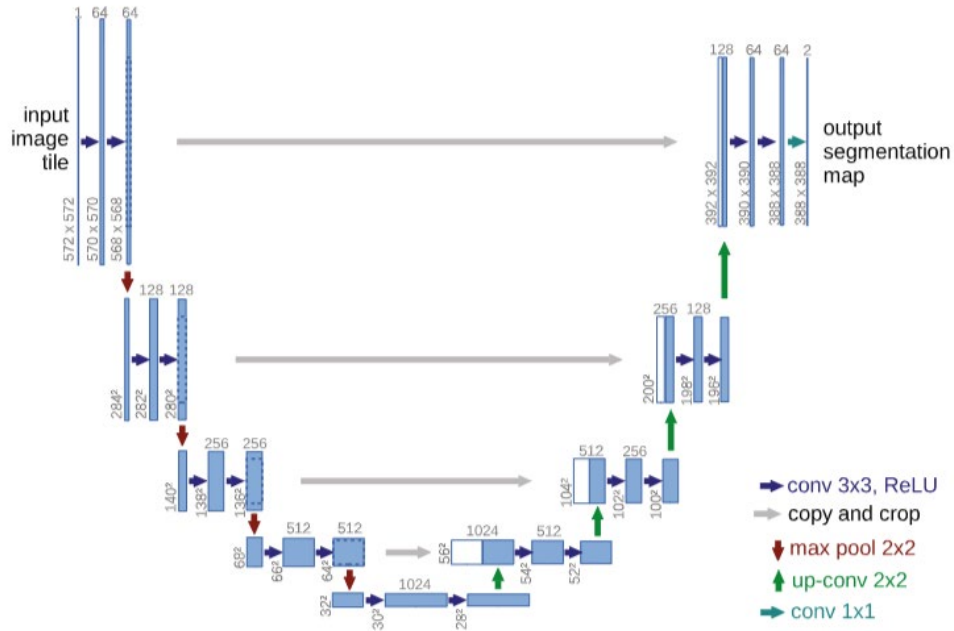


Figure 28. U-Net architecture from Ronneberger *et al.* [102]. Blue boxes: multi-channel feature map; white boxes: copied feature maps; arrows: operations; number on the top of the box: number of channels of each feature map; number at the lower left edge: x-y-size of the image.

In this case, i.e. for bone segmentation, different number of channels or filters in the first feature map, represented as F in Figure 29, are tested (2, 4 and 8). No higher values are tried due to the GPU limitations (only 6 GB of VRAM). As it could be seen, the number of channels for the remaining feature maps are adapted to the first one. In addition to the number of channels, the model is tested with different number of epochs (5, 10 and 20) for the training. With the objective of not overfitting the model, no higher number of epochs are tested.

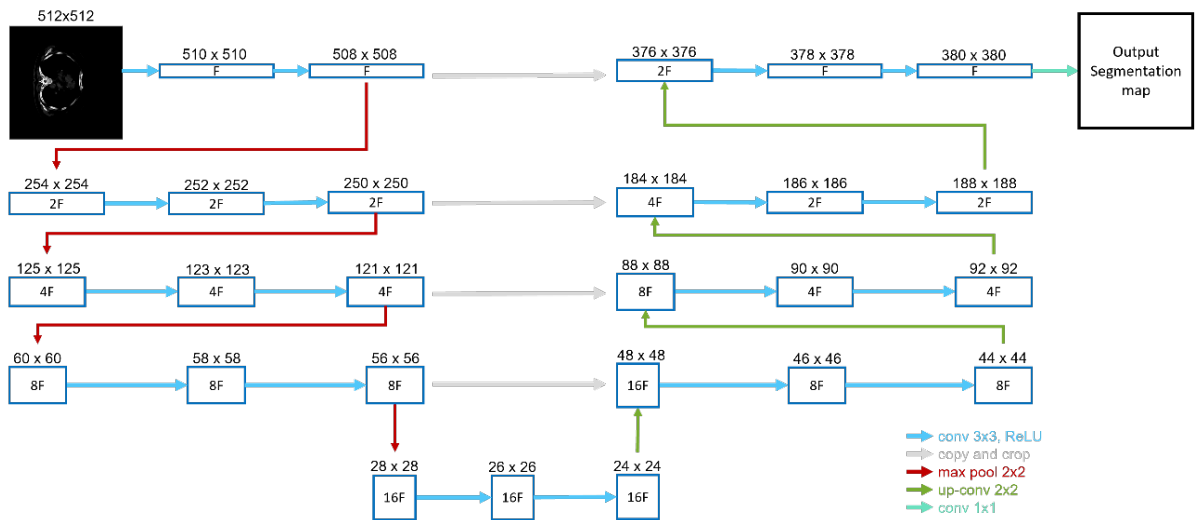


Figure 29. U-Net architecture adapted for input image of 512x512 resolution. Blue boxes: multi-channel feature map; arrows: operations; number inside the box: number of channels of each feature map (F , $2F$, $4F$, $8F$ and $16F$); number at the top of the box: x-y-size of the image. The left half of the diagram represents the contractive path and the right half the expansive path.

Using this model, an axial slice (2D image), i.e. the input image, is converted into an output segmentation map with the probability of being part of the binary mask, from 0 to 1, for each pixel. If the probability value is higher than or equal to 0.5, the pixel is included in the mask (value = 1). Otherwise, the pixel is interpreted as background (value = 0). In this way, the input image is converted into a binary image conformed of pixels with value 1 for the bone mask and pixels with value 0 for the background (non-skeleton regions, including soft tissues).

4.2.2.1.2 Dataset creation

First, all patients manually revised after the automatic segmentation based on traditional image processing algorithms are divided into two sets: 85% for the train and 15% for the test. In order to avoid overfitting the model to images of the same anatomical region, such as slices corresponding to the femur region, all training CT image slices are randomly ordered, both by patient number and by slice number.

Figure 30 shows an example of how a dataset is created from the CT scans of 5 patients. Firstly, a patient is randomly selected. Secondly, a slice of the CT scan of this patient is randomly selected. The ID of the patient and the number of the slice are saved in a shorted list. Then, this process is repeated until there are no slices left in any patient of the train set. Once the list is completed, it is divided into 5 different list sets. Every list set is converted into two volumes with the same corresponding slices: one for the CT scans (inputs), and other for the bone mask (outputs). This division avoids creating large collections of slices and allows to randomise the input order of the sets too. This way, 10 different training orders of the model are randomly generated, varying the order of the sets. All sets appear only once in each order. This way, the impact of the order of appearance of the sets on the training of the model is evaluated.

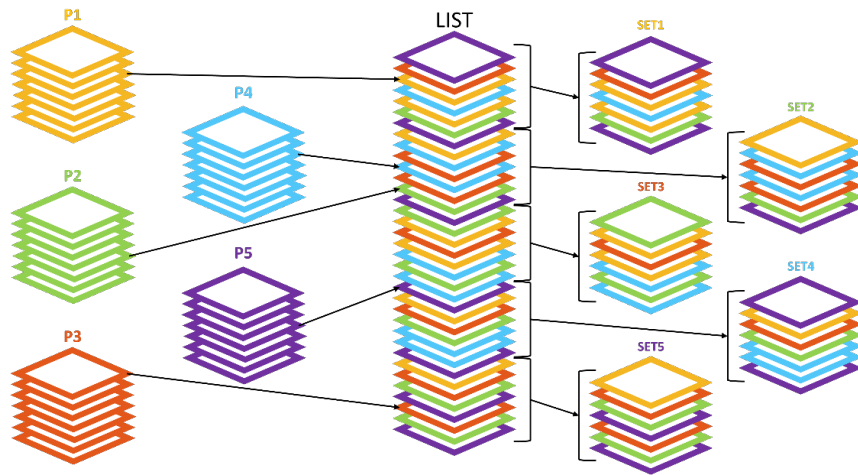


Figure 30. Dataset creation. Example of dataset creation with patient and slice randomization.

For the test set, the set creation is obtained by the addition of CT scans to a new list set, test list, slice-by-slice and patient-by-patient, without randomization, since

it does not provide learning to the model. Equally as for the train set, an input volume with the slices of the CT scans and an output volume with the slices of the bone masks are obtained for the test set following the defined test list.

4.2.2.1.3 Pre-processing definition

Before designing the pre-processing steps, the models are trained without any pre-processing, only clipping from the first femur slice to the last slice without head and the datasets randomization (as seen in Figure 30). A conversion to integer images is applied in order to consume less computational resources. These models result in an unacceptable segmentation, shown in Figure 31 and zoomed in Figure 32 for a better visualization. The main segmentation mistakes are: (1) only pixels with high grey intensity levels are included in the segmentation, causing not only an exclusion of regions with low grey intensity levels such as BM (Figure 31a and Figure 32a) or bone with reduced thickness (Figure 31b and Figure 32b), but also an over-segmentation of organs with high HU (Figure 31c and Figure 32c) or regions such as stretcher (Figure 31d and Figure 32d). Based on these observable mistakes, a pre-processing based on limiting the segmentation to the interesting grey intensity values and discarding the stretcher from the input image is proposed.

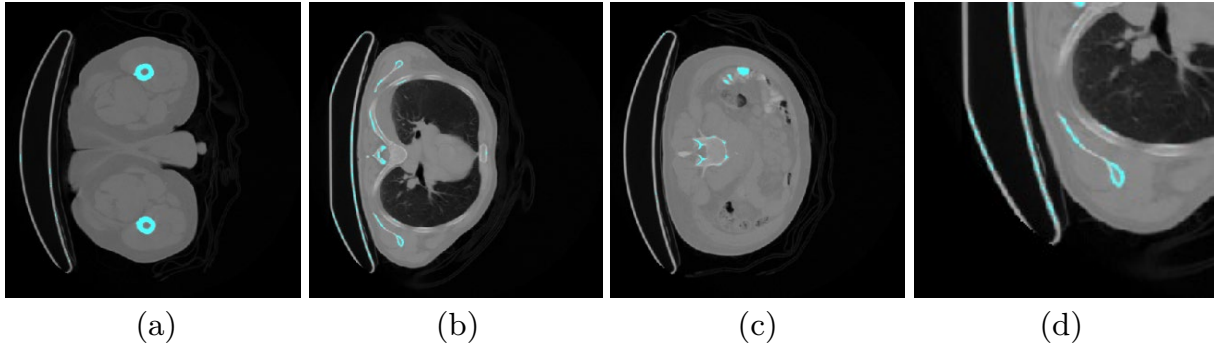


Figure 31. Bone segmentation mistakes for unprocessed database: (a) lack of BM in skeleton segmentation; (b) Exclusion of cortical bone regions of low grey intensity levels i.e. reduced thickness; (c) organs with higher grey intensity levels than bone inclusion; (d) stretcher inclusion due to its high grey intensity level.

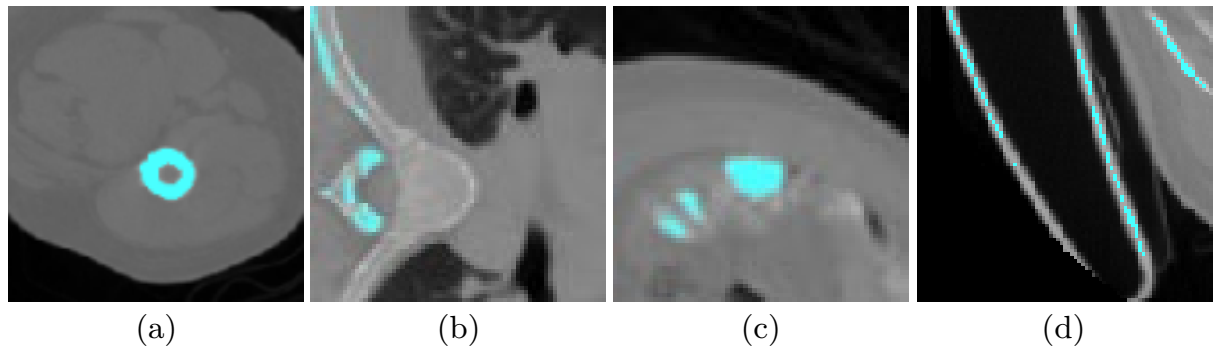


Figure 32. Zoom of Figure 31 for a better visualization of the mistakes for the bone segmentation model trained with unprocessed database. Each subfigure is the correspondent zoom of the Figure 31, i.e. the (a) is the zoom of Figure 31a, (b) is the zoom of Figure 31b, etc.

4.2.2.1.4 Pre-processing steps

Evaluating the mistakes from the segmentation of unprocessed data, the CT image is pre-processed in order to facilitate the prediction of the bone by the model. This pre-processing consists of 4 different steps: (1) stretcher removal, to prevent the model from confusing the stretcher with the bone due to the similarity in high-intensity levels in the CT image (over 300 HU); (2) thresholding, to focus the segmentation on the grey intensity levels corresponding to the compact bone and BM, as well as to unify the region removed from the stretcher with the image background; (3) image clipping, to unify both image acquisition protocols, whole-body and femur-to-head; and (4) normalization, to use integer images which consume less computational resources. This way, the final pipeline for the bone segmentation model is shown in Figure 33.

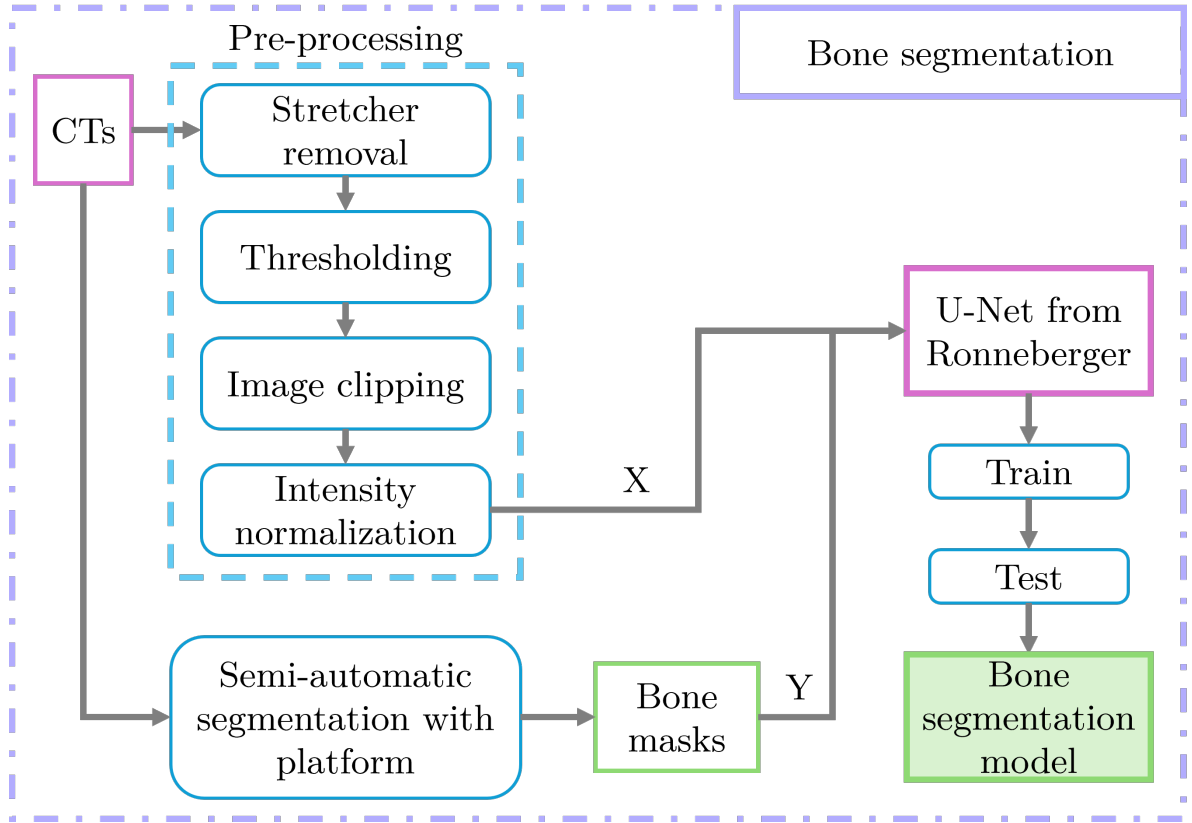


Figure 33. Pipeline for obtaining the bone segmentation model. X: train set data; Y: test set data.

4.2.2.1.4.1 Stretcher removal

Although a stretcher removal is developed for the semi-automatic segmentation, the results of the stretcher mask often fail, since for some patients the first thresholding detects as the greater volume the skeleton and not the stretcher (Figure 21c). In addition, the inside of the stretcher is not always completely removed. For this reasons, a new stretcher removal methodology is proposed. This methodology is

based on: margin location (Figure 34, Figure 35 and Figure 36), and stretcher filling (Figure 37), adaptation and removal (Figure 39).

First, the middle slice of the CT volume (Figure 34a) is binarised with a threshold of 98 HU, converting all values greater than this threshold into 1 and the rest into 0. However, for a better visualization, the rest of subfigures (Figure 34c, all from Figure 35, Figure 36, Figure 37, Figure 38 and ab from Figure 39) are shown with the CT slice thresholded between 0 and 300 HU as shown in Figure 36b, overlaying what is necessary for the explanation. After the binarization, the pixel of the binary mask closest to the image margin in the stretcher direction, i.e., at the left edge, is located along with the vertical line which crosses this pixel (i0_1, shown in red in Figure 36c). This way, this vertical line is assumed as the beginning of the stretcher. The next margin of the stretcher in the horizontal central zone that is included in the binary image, thus eliminating the soft zone of the stretcher, is located along with its vertical line (i0_2, shown in yellow in Figure 34c). Finally, to include also the extremes of the stretcher, a third index is obtained, which is the result of adding 20 pixels in the horizontal plane to the i0_2 index, and again the vertical line which crosses this point is obtained (i0, shown in green in Figure 34c).

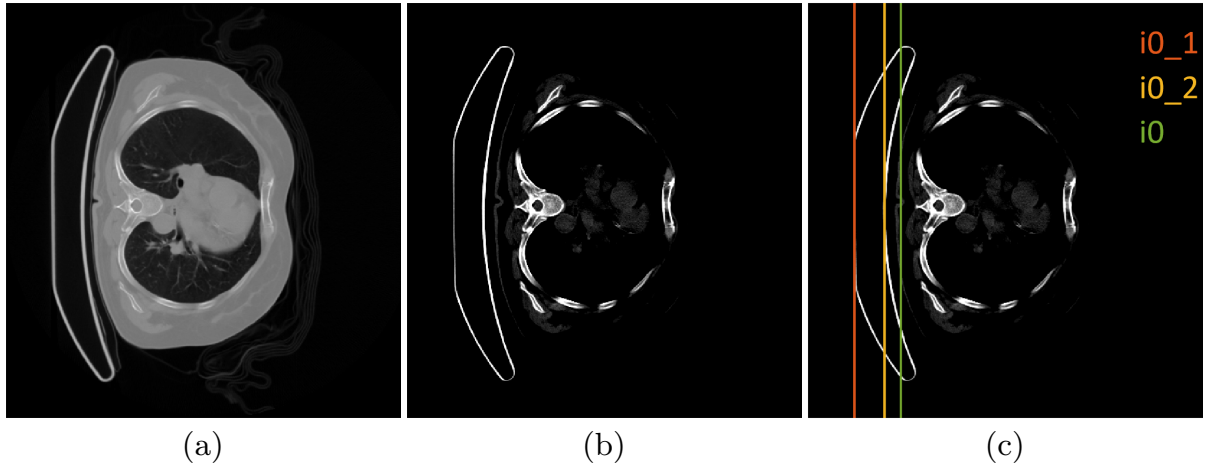


Figure 34. Stretcher removal I: margin location I. (a) Axial slice of the torso area extracted from the original CT scan; (b) slice shown in (a) threshold between 0 and 300 HU for a better visualization; (c) location of the stretcher points over the CT thresholded slice: closest to the left image margin in the horizontal central zone (i0_1, red), furthest from the left image margin in the central zone (i0_2, yellow), the point representing where most of the elevation of the stretcher ends (i0, green).

Based on the calculated indexes, it is considered that the whole stretcher is shown when the difference between i0_1 and i0_2 is higher than 50 pixels. Otherwise, only its upper half is present. This knowledge about the appearance of the stretcher is necessary to calculate the indexes that define the ends of the stretcher in the central zone, in the curved zone, the one with more elevation, and in the limits of the stretcher (i1, i2 and i3, shown as its vertical lines in green, yellow and red,

respectively, in Figure 35a). If the complete stretcher is present in the image, the indexes are calculated following (Eq. 1). Otherwise, if only the upper side, the half closest to the patient, is included, the indexes are calculated following (Eq. 2).

$$\begin{aligned} \text{Complete stretcher} \begin{cases} i1 = 105 + i0 \\ i2 = i1 - 20 \\ i3 = i2 - 10 \end{cases} & \text{(Eq. 1)} \quad \text{Half stretcher} \begin{cases} i1 = 45 + i0 \\ i2 = i1 - 10 \\ i3 = i2 - 10 \end{cases} & \text{(Eq. 2)} \end{aligned}$$

Drawing from these indexes ($i1$, $i2$ and $i3$), the slice is vertically divided into 8 equidistant regions defined as R1-R8, from top to bottom, divided by horizontal lines (shown in blue in Figure 35b). Then, the previously calculated indexes are related to the lines which separate the regions, for finding the intersection points, exterior (Pe) defined by intersection with $i1$ and $i3$, and interior (Pi) defined by the intersection with $i2$. The 2 superior regions (R1 and R2) and the two inferior ones (R7 and R8) are associated with index $i3$, i.e., the ends of the stretcher. In this way, the intersection points between the horizontal lines that limit R2 inferiorly and R7 superiorly with the vertical line of the index $i3$ are found and defined as Pe1 and Pe2, respectively. The regions around the middle of the image (R4 and R5) are associated with the central zone, i.e. index $i1$. In the same way, the intersection points of the horizontal lines that limit R4 superiorly and R5 inferiorly with the vertical line corresponding to index $i1$ are also found and defined as Pe3 and Pe4, respectively. This way, all the exterior points are found and shown in white in Figure 35c. Finally, the regions with the highest inclination, i.e. the transit regions between the central and the end regions, corresponding to R3 and R6, are associated with index $i2$. In order to ensure that the margin which represents the separation between stretcher and body of the patient is formed by vertical and horizontal lines, and that no inclination confuses stretcher pixels on the other side, the intersection points of all the horizontal lines corresponding to all the Pe (below R2, above R7, above R4 and below R5) with the vertical line corresponding to the index $i2$ are found and named Pi1, Pi2, Pi3 and Pi4, respectively (shown in pink in Figure 35c).

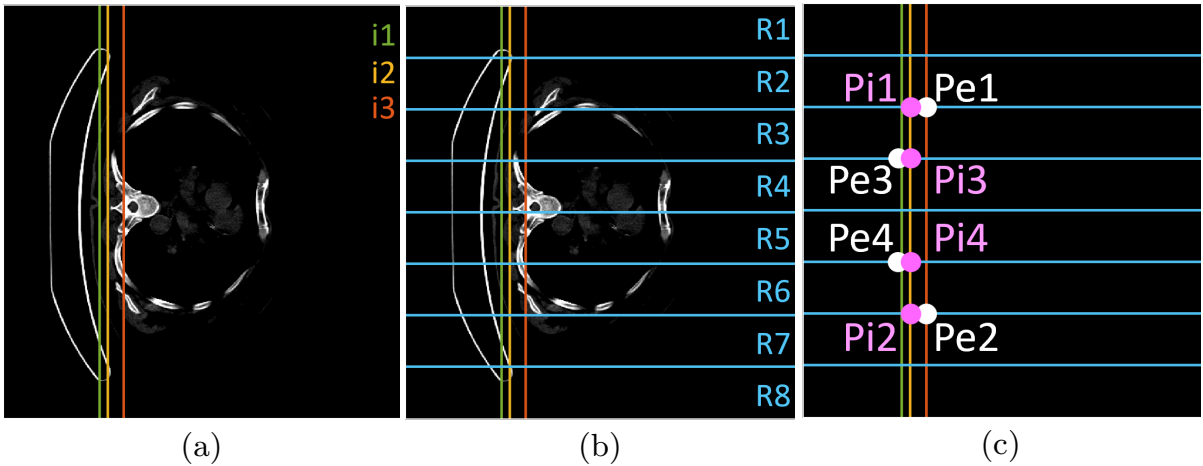


Figure 35. Stretcher removal I: margin location II. All images are shown over Figure 34b. (a) Location of the vertical lines that separate the stretcher from the body of the patient: in the central zone (i1, green), in the curve that goes from the central zone to the ends of the stretcher (i2, yellow), and at these ends (i3, red); (b) horizontal division of the slice in 8 equidistant regions (R1-R8, blue); (c) intersection points described as Pe (white) or Pi (pink), exterior or interior, shown over the correspondent vertical and horizontal lines (CT scan is not included for getting an easier visualization).

As it could be deduced in Figure 36a, all the intersection points are located between the stretcher and the skeleton. Subsequently, to create the margin which separates both structures, the intersection points defined are joint as follows: the upper point of the vertical line defined by the index i3 (at the upper margin of the slice) with Pe1; from this point the intersection points are joint as follows: Pe1 - Pi1 - Pi3 - Pe3 - Pe4 - Pi4 - Pi2 - Pe2; and finally Pe2 with the lower point defined by the vertical line of the index i3 (at the lower margin of the slice). The result of this union is the margin separating the stretcher from the skeletal area shown in Figure 36b with the auxiliar lines and in Figure 36c with only the thresholded CT.

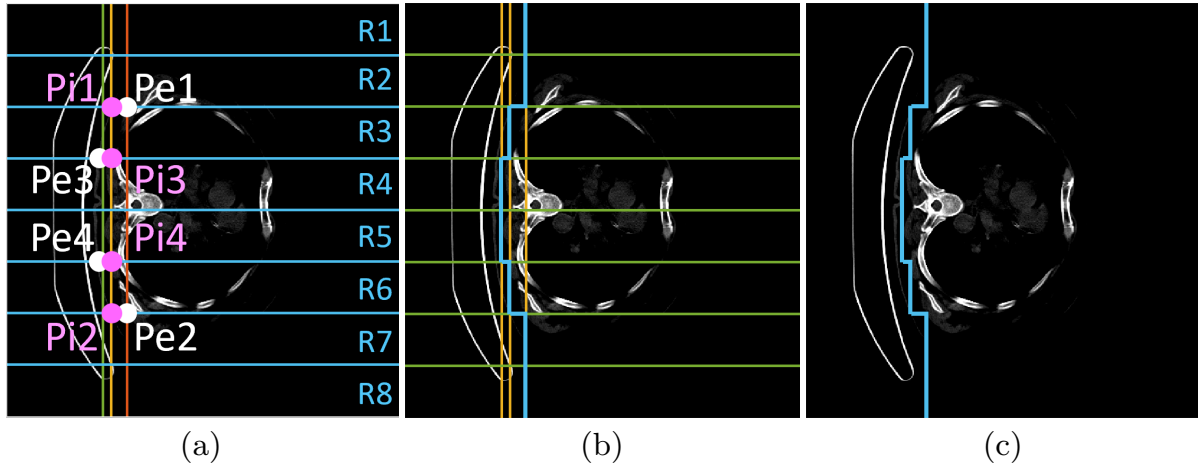


Figure 36. Stretcher removal I: margin location III. All images are shown over Figure 34b. (a) Combination of (Figure 35b) and (Figure 35c) to visualise the intersection points over the thresholded CT scan; (b) generation of the margin (blue) of separation between body and stretcher based on the previously calculated equidistant regions (green) and vertical lines of the indices (yellow), going through every intersection point; (c) margin (blue) separating the stretcher from the skeletal area shown in the thresholded CT slice.

Once the margin separating stretcher from the bone is obtained, the pixels of the binarised image in the region between the left edge of the image and the calculated margin are found and defined as the mask corresponding to the stretcher. In addition, especially for the cases where not the complete stretcher appears, the first 20 vertical rows of pixels, i.e., the closest pixels to the left edge of the image, are also defined as stretcher (Figure 37a). Subsequently, not only the hole generated by the soft material of the stretcher, but also the space between the stretcher and the left margin, are filled to obtain a single structure (Figure 37b). Then, the space between

superior and inferior image margins and the stretcher is filled too, since no anatomical regions should be included below the stretcher (Figure 37c). Finally, to avoid bigger holes and to avoid missing the borders of the structure, two morphological close operations are performed (Figure 37d).

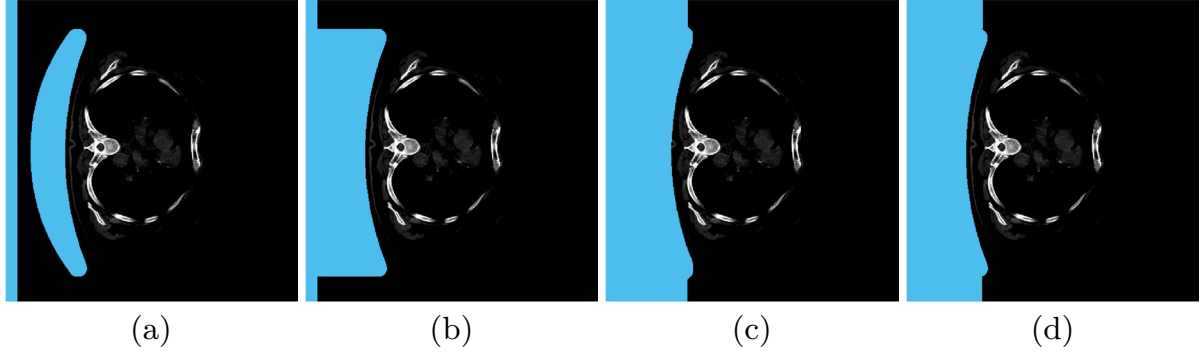


Figure 37. Stretcher removal II: stretcher filling. Blue mask represents the stretcher mask. All images are shown over Figure 34b. (a) Threshold slice of the CT scan highlighting the pixels with intensity level higher or equal to 98 HU in the region defined by the left border and the calculated margin; (b) stretcher mask after filling holes; (c) stretcher mask after filling the area between stretcher and superior and inferior edge of the image; (d) stretcher mask after filling holes by morphological close operation.

Since this stretcher mask is obtained for the middle slice, the mask is then adapted to every slice of the same patient. For this purpose, the pixels classified as stretcher mask are assigned as 0 in the binarised image. Then, the dilated mask, with a structuring element of circular shape and radius 7, is checked slice-by-slice to see if there are coincidences in any pixel with the binarised mask. If coincidences are found, the shape of the stretcher for the original calculated margin remains the same (Figure 38a). Otherwise, the dilated shape of the stretcher is used (Figure 38b and its zoom Figure 38c, with the green region corresponding to dilatation).

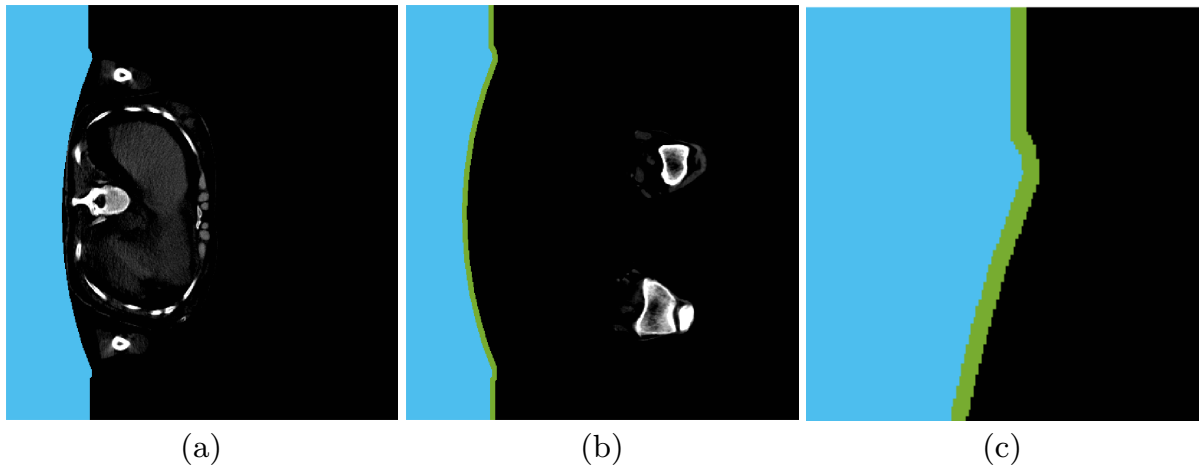


Figure 38. Stretcher removal III: stretcher adaptation I. Blue mask represents the stretcher mask. All images are shown over Figure 34b. (a) Stretcher mask for a CT slice of the torso area without adaptation; (b) stretcher mask for a CT slice of the end of the femur with adaptation by means of dilatation; (c) zoomed-in view of the adapted stretcher mask showing the green region corresponding to dilatation.

of dilatation, the green mask represents the adaptation with respect to the original stretcher; (c) zoom of (b) for a better visualization of the dilatation.

Finally, for the superior slices, a stretcher accessory that is placed on some patients in the head region to prevent movement and improve their comfort, is removed. In this case, a slice in which the stretcher presents this accessory is found by the dilation and the check of coincidences, and the lower and upper slices are searched to adjust the shape of the stretcher to the shape of the accessory (Figure 39a and its zoom in Figure 39b). Once the stretcher mask is obtained, it is removed from the CT scan defining every pixel present in the stretcher mask as a 0 in the CT scan (Figure 39c).

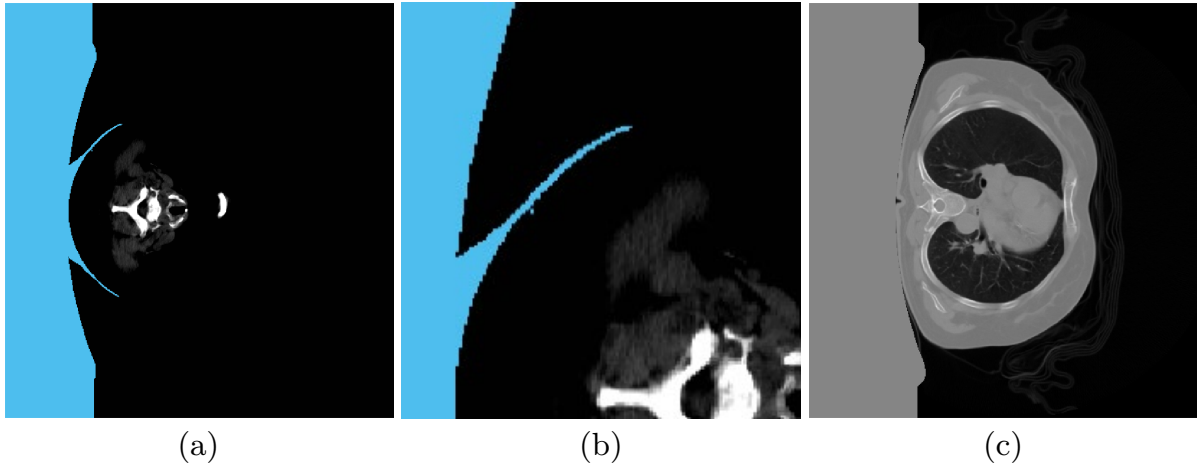


Figure 39. Stretcher removal III: stretcher adaptation II. Blue mask represents the stretcher mask over the thresholded CT. (a) Stretcher mask for a CT slice of the neck region with adaptation by means of stretcher accessory searching; (b) zoom of (a) for a better visualization of the dilatation; (c) original CT scan slice substituting the values of the stretcher region by 0.

4.2.2.1.4.2 Thresholding

In order to focus the segmentation not only on compact bone but also on the BM, the CT scan is processed using two different thresholds. First, all voxels with a value lower than or equal to 0 HU are converted into 0 (Figure 40a), since the regions with these HU values in CT correspond to uninteresting elements for analysis such as air (background), some soft tissues (such as adipose tissue) and the removed stretcher. Then, to include both structures of the skeleton, i.e. the compact bone and BM, a threshold of 300 HU is applied, assigning a value of 300 HU to all voxels with a value greater than or equal to 300 HU (Figure 40b), as they are the only tissues in the human body with a HU value close to and above 300 on a CT scan, which prevents further tissue mistaking at the upper threshold value.

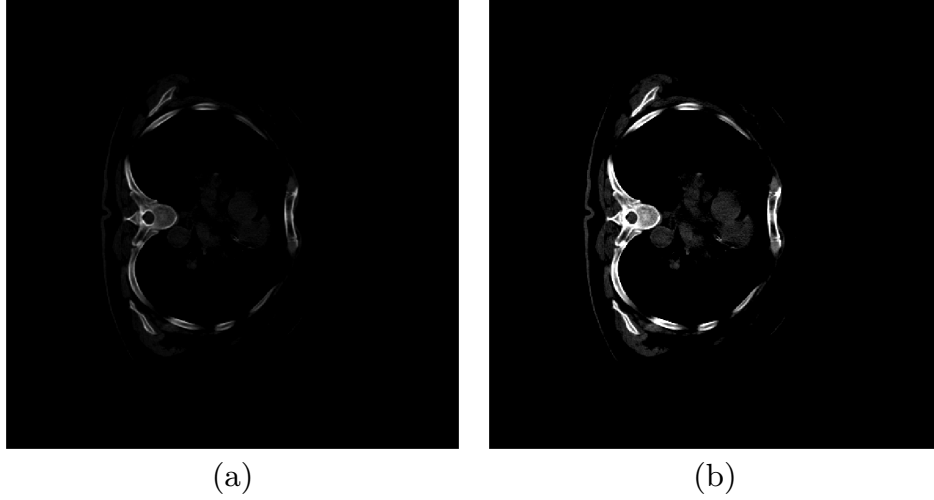


Figure 40. Thresholding process. (a) CT scan slice in (Figure 39c) thresholding pixels with values lower than or equal to 0 HU; (b) CT scan slice in (a) thresholding pixels with values greater than or equal to 300 HU.

4.2.2.1.4.3 Volume clipping and intensity normalization

As mentioned in the introduction of the section 4.2. *Bone marrow segmentation*, the segmentation is limited to the volume which goes from the femur to the beginning of the head. Culminating with the pre-processing, the image is normalised, since an input manageable in terms of size is necessary. For this reason, the main idea is to convert the image into a format of integers between 0 and 255, i.e. in 8 bits. For obtaining these images, first, the volume (V) is normalised between 0 and 255 by means of (Eq. 3), being $minGL$ and $maxGL$ the minimum and maximum value of grey level of the input volume, respectively, i.e., 0 and 300 after the thresholding step. Finally, the values are converted into integer values.

$$V_{normalized} = \frac{255 \times (V - minGL)}{(maxGL - minGL)} = \frac{255 \times (V - 0)}{(300 - 0)} = \frac{255 \times V}{300} \quad (\text{Eq. 3})$$

Once all the images are pre-processed, the model is again trained and tested with the different values of F , E and order of the dataset. From all of these combinations, the model with the best performance is selected to be used as the model for bone segmentation.

4.2.2.2. Bone marrow segmentation

For the BM segmentation, the problem is more complex since the process of manual BM segmentation requires a large amount of time, between 6 and 12 hours per patient (approximately 300 slices). This way, the obtention of several images for an acceptable database is a time-consuming process. For this reason, the pipeline shown in Figure 41 is proposed. First, it is worth mentioning that, as explained before, the manual segmentation is time-consuming. Consequently, a pseudo-GT (pGT) of the

BM mask is obtained from the bone masks. Secondly, some steps of the pre-processing are partially changed from the bone segmentation pre-processing, since the ROI has changed. Finally, the U-Net from Ronneberger *et al.* [102] included in Figure 28, adapted in Figure 29, and previously explained in subsection 4.2.2.1.1. *Model architecture*, is trained differently for each of the different approaches:

- I. Approach based on the untrained U-Net model using as input the pre-processed CT scan.
- II. Approach based on the untrained U-Net model using as input the pre-processed CT scan in the regions of the bone mask. This approach has no independence of the bone mask.
- III. Approach based on the U-Net model trained for the bone segmentation using as input the pre-processed CT scan.

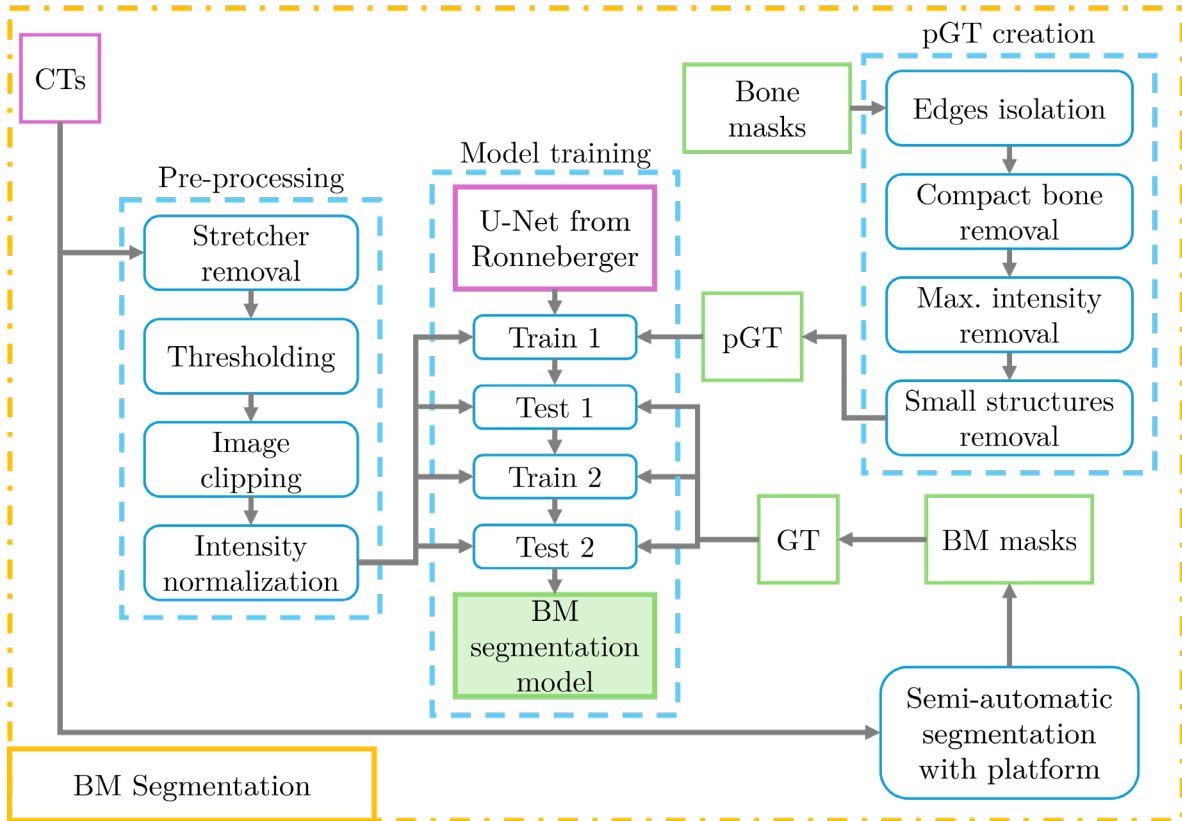


Figure 41. General pipeline for obtaining the BM segmentation model. For approaches I and II the U-Net from Ronneberger is the original architecture. For approach III, the model is the one obtained as the bone segmentation model. Although it is not marked for an easy visualization, in the approach II, the bone mask is also multiplied with the CT scan for creating the input of the model.

In addition to the different approaches, three different collections combining both image cohorts and both bone masks origins are proposed, resulting in a total of nine combinations of possible models. The proposed collections differ in the number of patients and the precedence of the bone masks used for creating the pGT:

- A. The same cohort used for the bone segmentation model, using as bone masks those manually segmented for being the GT of the bone model.
- B. The same cohort used for the bone segmentation model, using as bone masks those obtained from the bone segmentation model.
- C. A larger cohort than the one used for the bone segmentation model, using as bone masks those obtained from the bone segmentation model.

Although different bone masks are used, all of them are processed to obtain the BM pGT. Since, as mentioned before, the manual segmentation is time-consuming, obtaining a complete BM GT collection is complex. In the following subsection, the pGT creation process is described.

4.2.2.2.1 Pseudo-ground truth creation

This process of creating the BM pGT includes four steps: (1) dilation of bone mask, predicted or manually revised, with a structuring element of disk shape with 3 pixels of radius followed by subtracting the original mask from the dilated mask to leave only the edges; (2) the dilation of the edges mask with a structuring element of disk shape with a radius of 2 pixels followed by its removal from the bone mask; (3) removal of all the pixels with a grey value equal to the maximum after the normalization (which corresponds to all the pixels over 300 HU in the original CT scan) from the remaining mask; (4) smoothing of the resulting mask by removing structures with less than 10 pixels. Figure 42 shows some examples where the difference between the masks resulting from each step can be seen.

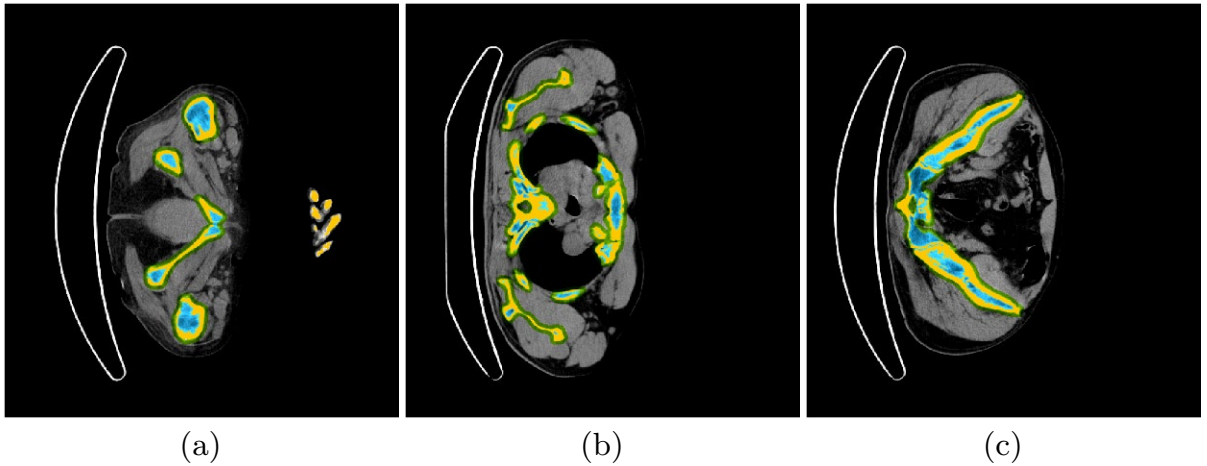


Figure 42. Examples of the pGT created for BM segmentation in: (a) and (c) pelvis regions, and (b) torso area. (Green) edges removed as a result of the step (2); (yellow) pixels over 300 HU in the original CT removed in the step (3); (blue) resulting BM mask after applied the four steps.

Although this BM pGT can be used for training the models, it can be seen in Figure 43 that there are several mistakes which reduce the usefulness of these masks to be used in BM quantification. First, some regions between compact bone of different

bone structures, which should be considered background, are marked as BM. In addition, some thin compact bone is also considered BM.

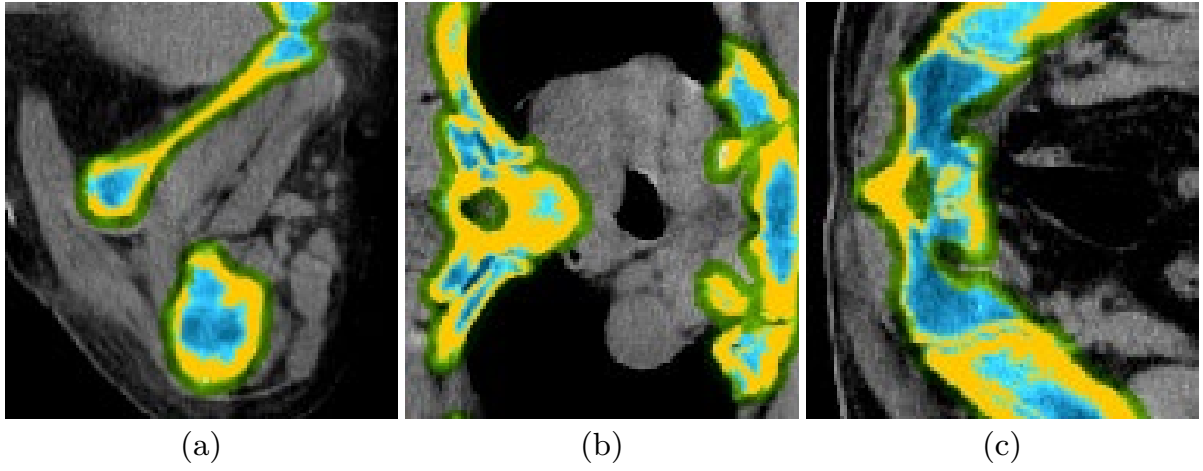


Figure 43. Zoom of the Figure 42 for a better visualization. Each subfigure is the correspondent zoom of the Figure 42, i.e. the (a) is the zoom of Figure 42a, (b) is the zoom of Figure 42b, and (c) is the zoom of Figure 42c.

4.2.2.2.2 Changes in the pre-processing

Same as in subsection 4.2.2.1. *Bone segmentation*, the pre-processing consists of four steps: (1) stretcher removal, (2) thresholding, (3) image clipping and (4) intensity normalization. However, some changes are applied. First, for step (1), and based on the original CT images and the stretcher removal process described in subsection 4.2.2.1.4.1. *Stretcher removal*, the same U-Net architecture for bone segmentation is trained for obtaining the stretcher masks. As GT for training the stretcher removal model, the final stretcher mask (shown in Figure 38) multiplied with its correspondent CT scan slice, and then thresholded with 98 HU, is used (see Figure 44). Similarly to the bone segmentation, different number of channels or filters in the first feature map ($F = 2, 4$ and 8) and different number of epochs ($E = 5, 10$ and 20) are tested.

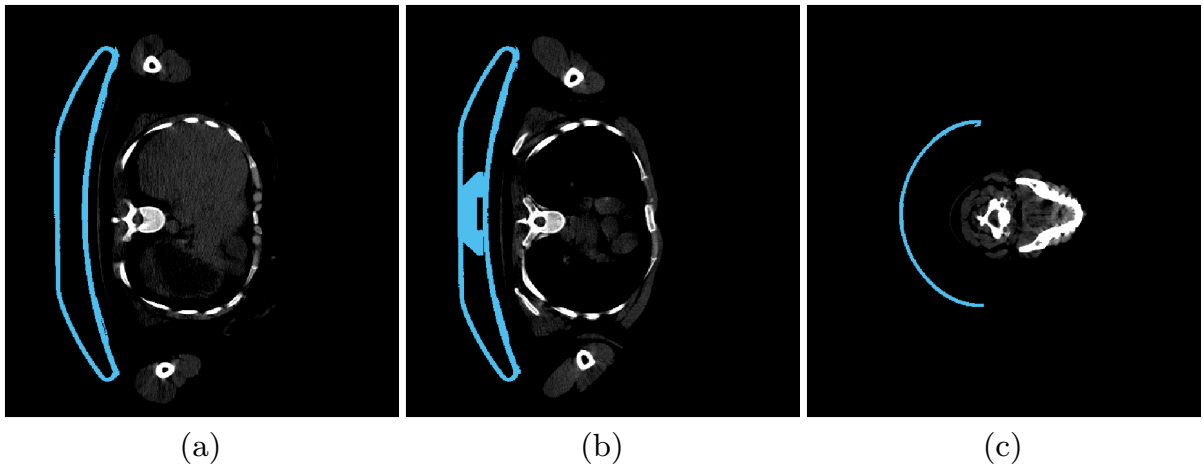


Figure 44. Examples of stretcher masks used as GT for the stretcher removal model: (a) usual stretcher shape; (b) usual stretcher shape with some metallic material inside for a better support of the patient weight; (c) accessory metallic object for supporting the patient head.

Once the stretcher is identified and removed, step (2) is lightly changed. To focus the segmentation on the BM, the thresholds selected and applied in the CT images are -100 HU and 300 HU. Then, step (3) is equally applied. Finally, the images are normalised following step (4) between 0 and 255. Nonetheless, due to the new thresholding, the normalization is made following (Eq. 4).

$$V_{normalized} = \frac{255 \times (V - minGL)}{(maxGL - minGL)} = \frac{255 \times (V - (-100))}{(300 - (-100))} = \frac{255}{400} \times (V + 100) \quad (\text{Eq. 4})$$

4.2.2.2.3 Model training

For a clearer explanation of model training, a schematic diagram detailing the process is shown in Figure 45. For each pair of approach (I and II) and collection (A, B or C), the model is trained varying the same parameters changed for bone segmentation, i.e. the number of channels or feature maps in the first layer (F), the number of epochs and the training order. In this case, the number of training orders is superior to those created for bone segmentation since the masks used as GT are less precise (pGT) than in the skeleton, giving more opportunities to converge. In the approach III, the retrained model is previously defined with a specific number of F and epochs, since it is optimal for bone segmentation. In this case, only the number of training orders is varied.

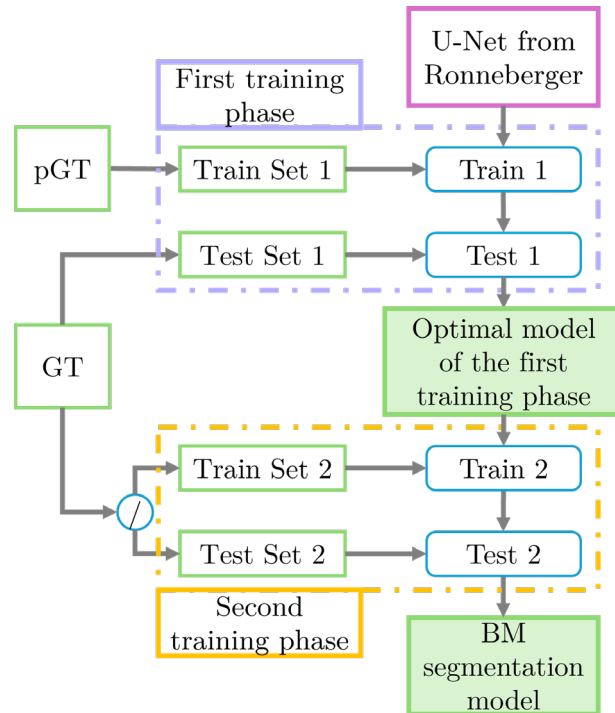


Figure 45. Model training pipeline for BM segmentation.

Thus, a **first training phase** is performed using the pGT masks as the train set and all the GT masks, i.e. the ones manually segmented ones, as the test set. Once all combinations of parameters are tested in this first training phase, the optimal combination for each pair of approach (I, II and III) and images collection (A, B and C) is retrained in a **second training phase**, with part of the patients manually segmented, and tested in the rest of them.

4.2.3. Evaluation of segmentation models

For the performance assessment of all the segmentation models, multiple metrics are calculated. The most commonly used metrics for measuring the superposition of predicted and GT area are the Jaccard index (IoU) [152] and the Dice Index [153]. These metrics were proposed in 1912 and 1945, respectively, for purposes very different to analysing medical images. However, due to their simplicity, they have been used in the last three decades for assessing medical segmentation model performances, as it can be seen in the studies [154–156].

In terms of definition, the IoU index measures the proportion of the intersection area between two masks, in relation to area resulting of the union of both masks, while the Dice index measures the proportion of the same intersection in relation to the mean of both individual area sets. Both metrics are associated with Sensitivity and Specificity coefficients, since they measure the amount of area correctly segmented or correctly excluded from the segmentation [157]. Both metrics, IoU and Dice indexes, are expressed following (Eq. 5) and (Eq. 6), being A the GT mask and B the predicted one:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (\text{Eq. 5}) \qquad \qquad \qquad Dice = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (\text{Eq. 6})$$

To convert these indexes in terms of the confusion matrix variables, as easily observed in Figure 46, there are some considerations which must be considered:

- The intersection of A and B, i.e. the correctly segmented pixels, conforms the True Positive (TP) region. Summarised in (Eq. 7).

$$|A \cap B| = TP \quad (\text{Eq. 7})$$

- The union of A and B, i.e. the pixels in one or both masks, includes: the region corresponding to TP; the region where A is included but B is not, i.e. the False Negative (FN) region; and the region where A is not included, but B is, i.e. the False Positive (FP) region. Summarised in (Eq. 8).

$$|A \cup B| = TP + FN + FP \quad (\text{Eq. 8})$$

- The sum of A and B, correspond to all both masks, including: the FN region, the FP region, and twice the intersection region between A and B, i.e. the TP region. Summarised in (Eq. 9).

$$|A| + |B| = 2 \times TP + FN + FP \quad (\text{Eq. 9})$$

- Pixels not included in A or B, i.e. pixels not segmented (neither incorrectly nor correctly), are the True Negatives (TN).

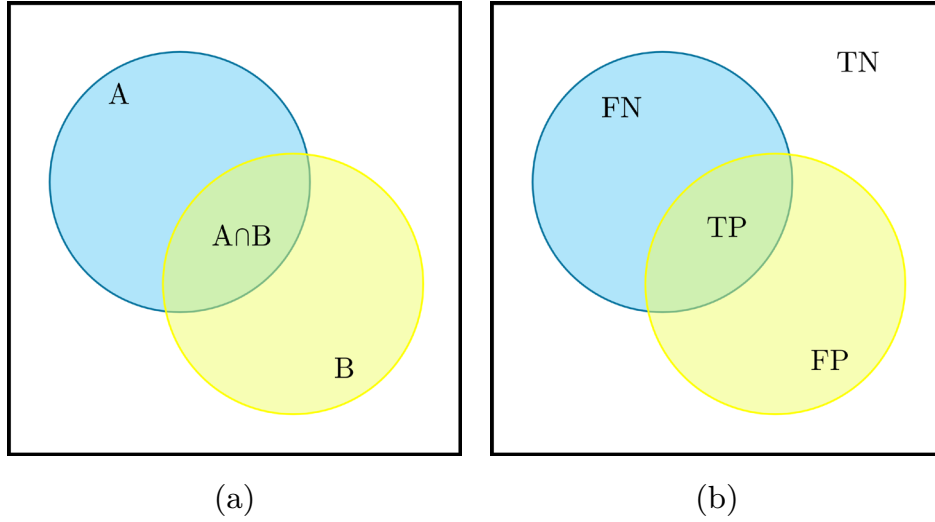


Figure 46. Example of two masks to visualise the IoU and Dice indexes. (Blue or A) GT mask; (yellow or B) predicted mask. (a) the example is explained with set theory; (b) the example is expressed as confusion matrix metrics.

With these considerations, (Eq. 5) and (Eq. 6) are converted into (Eq. 10) and (Eq. 11), respectively. In this way, the calculation from the obtained mask is easier, since: TP represents the pixels correctly assigned as mask, FP represents the pixels wrongly assigned as mask, and FN represents the pixels wrongly assigned as background.

$$IoU = \frac{TP}{TP+FN+FP} \quad (\text{Eq. 10}) \quad Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (\text{Eq. 11})$$

This way, Jaccard or IoU and Dice indexes for the assessment of the bone, BM and stretcher segmentation models, are calculated as the mean of both labels, i.e. background and mask. First, for the background mask, (Eq. 10) and (Eq. 11) are calculated taking as TP which in Figure 46 is the TN, i.e. pixels correctly assigned as background. Then, the value of both indexes for the mask is calculated, following (Eq. 10) and (Eq. 11), literally. Then the average of both, the corresponding index for background and mask, is calculated.

Although the segmentation based on traditional algorithms is an unsupervised segmentation, in order to make easier the comparison between segmentation methodologies, the same patients from the test set of the DL approach are used for

evaluating the performance of this first approach. It must be noted that manual revisions of patients which are used posteriorly for the DL segmentation models are used, as mentioned before, as the GT for evaluating the performance of different segmentation approaches.

4.3. Bone marrow quantification in [^{18}F]FDG PET/CT

As mentioned before, an adequate segmentation of the BM is essential for its quantification. The use of artificial intelligence has hugely increased not only for segmentation tasks, but also for quantification. In fact, there are multiple possibilities which are being used for quantifying biomedical images [158].

For this thesis, as shown in Figure 47, the activity of the BM, extracted from the [^{18}F]FDG PET image is quantified by means of two different approaches: (1) ML models based on radiomics features, implementing a univariate analysis of the features and an oversampling method; (2) DL models, concretely CNNs, using as input the 3D image or the MIP image, and with and without data augmentation. For both quantification approaches, both labelling methods, i.e. visual assessment of the PET and MFC result of the biopsy, are used for classifying the images in the different classes. In addition, in some of the clinical contexts, another way of labelling, mixing both previous mentioned labels, is used and defined as pGT labelling.

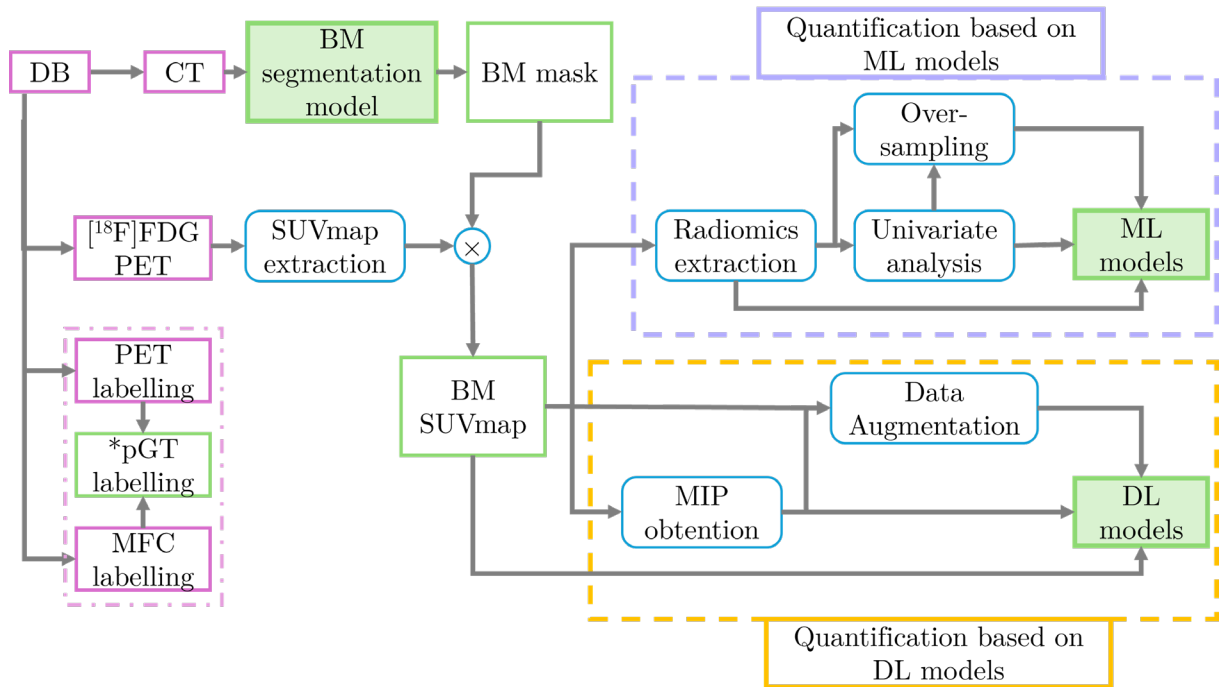


Figure 47. General pipeline of the BM quantification methodology. (Purple dashed box) approach corresponding to ML models; (yellow dashed box) approach corresponding to DL models. DB: database.

4.3.1. Pseudo-ground truth labelling

This pGT labelling mentioned above, is applied in two different clinical contexts: MRD and NHL. In the case of MRD, the pGT labelling is exactly the same as in MFC, nevertheless, only patients with PET- are included. This way, the possibility of classifying the MFC results in the cases where it is not possible to observe any pathological uptake through the visual assessment by experts of nuclear medicine is tested. The main objective of using only these patients is to study the potential of both methodologies, ML and DL models, to find differences between MFC- and MFC+ classes which the human eye is incapable to see.

In the case of NHL, the pGT labelling is evaluated with another purpose. There are multiple discrepancies between BM assessment based on MFC and the visual assessment of the $[^{18}\text{F}]$ FDG PET imaging depending on the uptake pattern. A patient with PET+ and MFC- often indicates a missed BMI, due to focal uptake [130], since the sample taken for the biopsy could not be representative for the whole BM. In addition, as mentioned before, the heterogeneous pattern observed in $[^{18}\text{F}]$ FDG PET images in patients with BMI+ is very similar to the pattern observed in those patients with an inflammation or infection, i.e., a rBM [10,48]. For these reasons, a pGT, where information about both tests is taken account, is proposed as an additional labelling to predict. The rules applied to this pGT, from the most to the least important, are defined as follows and shown in the Table 7:

1. If only PET assessment is available, the pGT is defined as the PET labelling using the same class than the described by the experts in nuclear medicine.
2. If both assessments, MFC and PET, have the same class, the pGT is defined as this coincident class.
3. If the MFC assessment is BMI+ or rBM, the pGT is the same class as the MFC independently of the PET, due the MFC technique is highly sensitive.
4. If the MFC assessment is BMI-, but the PET assessment is BMI+, the pGT is defined as BMI+, since the sample taken for the biopsy could be not representative of the whole BM due to its small size.
5. If the MFC assessment is BMI- and the PET indicates rBM, the pGT is defined as the BMI-, since the rBM identification by the visual assessment of the PET image is considered tough.

		PET		
		BMI-	BMI+	rBM
MFC	BMI-	BMI-	BMI+	BMI-
	BMI+	BMI+	BMI+	BMI+
	rBM	rBM	rBM	rBM
	X	BMI-	BMI+	rBM

Table 7. Confusion matrix for the pGT labelling definition. X: no MFC test available.

4.3.2. PET processing

For the use of [^{18}F]FDG PET imaging, having a co-registered PET image taking as reference the respective CT, since it represents the anatomical location, is necessary. The term registration for two biomedical images implies the spatial alignment between them. For this alignment, the obtention of the transformation which can associate the position in one coordinate space with the coordinate space of the other image must be found. For the obtention of this transformation, based on sampling and interpolation, the co-registration process identifies features that correspond to the same anatomical entity. If the transformation is properly found, the overlap of both images is possible in the same coordinate space [159]. Although the process of co-registration is applied, the acquisition device for both image modalities is the same, Siemens Biograph TruePoint 6 PET/CT (Siemens Healthineers, Erlangen, Germany), implying a co-registration after the acquisition process, due to the coordinate space is the same for both images. For this reason, the only step in adapting [^{18}F]FDG PET imaging to CT is to resample the resolution to CT, since the PET resolution is worse than the CT one. Due to both images share the coordinate space, the process of adapting the resolution is based on an interpolation of the PET image to transform its dimensions to the CT ones. In addition, the same acquisition device includes an attenuation correction of the PET using the low-doses CT, consequently, a value intensity registration is not necessary.

Once the resolution is adapted, the aim is to differentiate the activity of each tissue by measuring the uptake of the radiotracer, in this case [^{18}F]FDG, since the variation of the radiotracer uptake is directly related to a different metabolism of the regions which is cause of the intratumoural heterogeneity [160]. For this reason, as mentioned in several sections (*1.3. Bone marrow assessment* and *3.2. Bone marrow quantification*), the SUV metrics are the most extended and commonly used metrics for quantifying PET images. For obtaining these metrics (SUVmax, SUVmean, SUVpeak, etc.), the obtention of the SUVmap is a compulsory step. This SUVmap is calculated following (Eq. 12) described in [161]. In this way, from the activity concentration (measured by the PET image), the FDG dose injected for the acquisition, corrected for physical decay, and the weight of the patient, the SUVmap is calculated.

$$SUV = \frac{\text{Activity Concentration (kBq/ml)}}{\text{FDGdose (MBq)/Body Weight (Kg)}} \quad (\text{Eq. 12})$$

Using this image described as SUVmap, with the concrete positions of the BM, i.e. its segmentation, the metabolism of the BM is possible to study. However, the quantification of PET images based on SUV metrics is very limited (*3.2. Bone marrow quantification*), since no heterogeneity appreciations of the metabolism are made. For this reason, a more complex quantification is proposed.

4.3.3. Quantification by Machine Learning models based on radiomics

Using the [^{18}F]FDG PET imaging, i.e. its SUV map, and the BM mask obtained with the BM segmentation DL model, 125 radiomics features are extracted from each image with the library pyradiomics [162], varying the discretization in 4 different combinations. Once these features are extracted, four different characteristic groups are created: the original radiomics features (ORIG); the reduced one (RED), applying a univariate analysis in order to select only the most important features; the same radiomics features from the ORIG group, but with the database previously oversampled (OS), i.e. creating some false samples of radiomics features; and the OS group but with the same reduction that is applied in the RED group (OS_R). Then, eleven ML models are trained and tested for these four groups. This quantification approach is detailed in Figure 48.

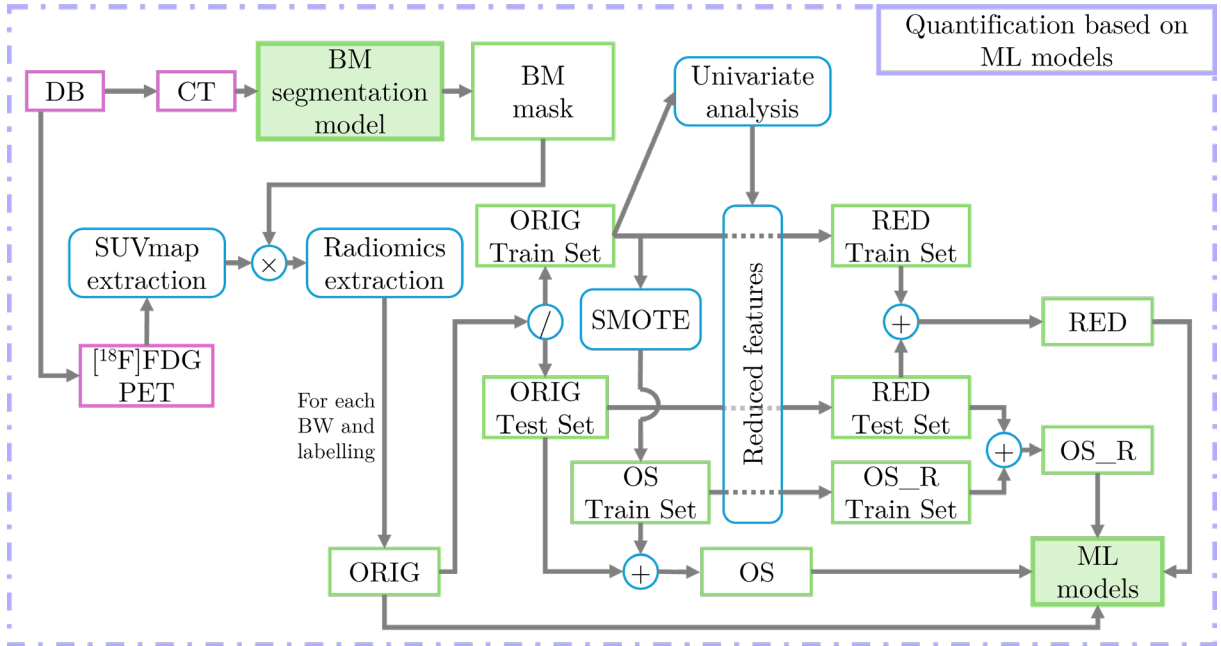


Figure 48. Pipeline of the quantification based on ML models. ORIG: original group of radiomics features; RED: reduced group of radiomics features; OS: oversampled group maintaining the original radiomics features; OS_R: oversampled group reducing the number of radiomics features. SMOTE: Synthetic Minority Over-sampling TEchnique. BW: Bin width.

4.3.3.1. Radiomics features extraction

As mentioned above, [^{18}F]FDG PET/CT imaging can potentially be used for evaluating the BM pattern and, consequently, its condition in multiple clinical contexts [46,58,68,69,76]. The visual assessment of this image modality, based on the detection of regional variations in metabolism which cause the intratumoural heterogeneity [160], provides valuable information for choosing treatment strategy

and performing a more personalised medicine [163]. However, medical images are demonstrated to contain more information than the human eye can observe. This way, the extraction of radiomics features emerge as a necessary tool for additional information [164].

For radiomics extraction, many research groups have developed their own radiomics software packages. Nonetheless, each of these packages has their own algorithms for features calculation and for imaging analysis, i.e. for a concrete tissue or image modality. These variations produce substantial differences between the radiomics features estimations [165,166]. For this reason, the Image Biomarker Standardization Initiative (IBSI) arises in order to avoid the problems of radiomics features related to the variation and reproducibility [167]. The IBSI proposes the pipeline shown in Figure 49. As it could be seen, there are several main steps, from obtaining the ROI to calculating every textural feature from the radiomics matrixes. From the IBSI recommendations, the pyradiomics [162] library for python language [150] is created. This library includes multiple functions which implement the steps between image interpolation and radiomics features, standardizing the whole process of extraction. This way, a dictionary of all the calculated features is returned with this library [168].

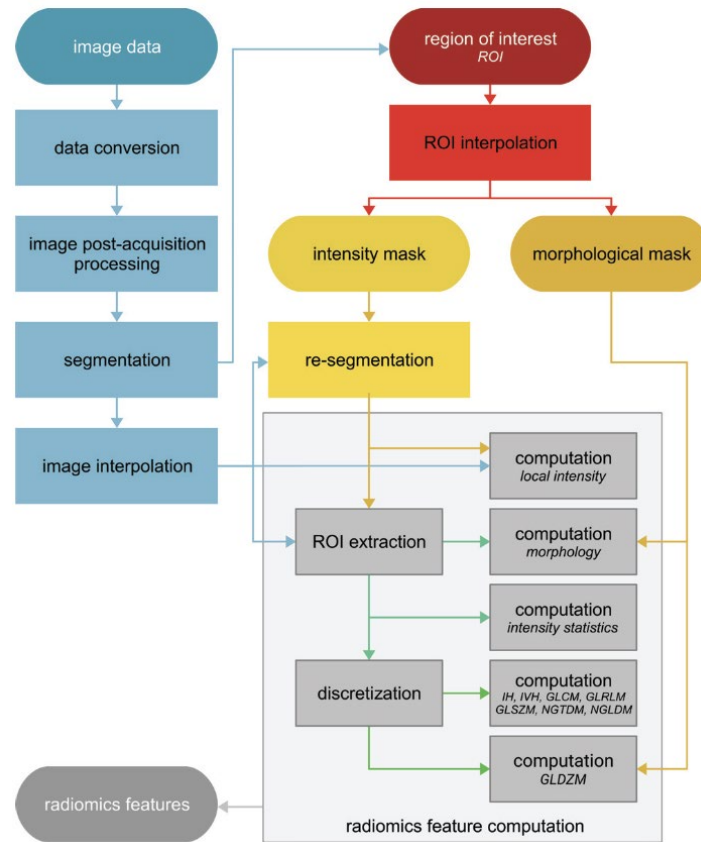


Figure 49. Pipeline of the radiomics image processing scheme for computing radiomics features. IH: intensity histogram; IVH: intensity-volume histogram; GLDZM: Grey Level Distance Zone Matrix [167].

Using the library pyradiomics [162], a feature extractor¹ is configured for obtaining radiomics features from the SUVmap using the BM mask as ROI. As mentioned, the extractor develops the complete process for having a standardised radiomics features extraction. Following the pipeline represented in Figure 49, first, for the image interpolation, the volume is resampled to an equidistant space, i.e. the resampled voxels have the same length in all directions, since many features are sensitive to the voxel size [169]. Taking the original CT resolution, 0.9766x0.9766x2.5 mm³, which is the SUVmap resolution since the PET is previously adapted to the CT, the images are resampled to 1x1x1 mm³. Then, having the ROI of the BM and the SUVmap interpolated, the ROI is extracted from the image, since many radiomics features require an isolated ROI. Finally, the image in the ROI is discretised. However, there are two different discretization methods which are being commonly used, [170] and their differences have to be considered to decide the optimal discretization for the present study.

4.3.3.1.1 Discretization of the images

The discretization method for the SUVmap can considerably change the radiomics features values and their interpretation, making crucial an adequate discretization for obtaining a good quantification [171]. Two main discretization methods are applied for radiomics features extraction:

- **Fixed bin number.** This discretization method adapts the image to a fixed number of bins, following (Eq. 13) described in [170]. In this way, the intensity of the voxel ($X_{gl,k}$) is corrected with the lowest intensity of the ROI ($X_{gl,min}$), taking the width of the bin into account ($X_{gl,max} - X_{gl,min}$), and rounding down to the nearest integer.

$$X_{d,k} = \begin{cases} \left\lfloor N_g \frac{X_{gl,k} - X_{gl,min}}{X_{gl,max} - X_{gl,min}} \right\rfloor + 1 & X_{gl,k} < X_{gl,max} \\ N_g & X_{gl,k} = X_{gl,max} \end{cases} \quad (\text{Eq. 13})$$

- **Fixed bin size.** This discretization method, defined in [170] by (Eq. 14), assigns a new bin for an interval of intensity values based on the defined bin width, starting at the lowest intensity of the ROI or the minimum defined by the user, being recommended a value of 0 for the SUV.

$$X_{d,k} = \left\lfloor \frac{X_{gl,k} - X_{gl,min}}{w_b} \right\rfloor + 1 \quad (\text{Eq. 14})$$

¹ <https://pyradiomics.readthedocs.io/en/latest/radiomics.html#module-radiomics.featureextractor>

In Figure 50, an example of both discretization methods applied to two images with different range of SUV values is shown. The fixed bin number discretization method spoils the relationship between the image intensity and the physiological meaning of this intensity, in this case, the functional nature of the PET image. Meanwhile, the fixed bin size discretization respects the contrast between images. For this reason, and due to the functional nature of the PET, the discretization applied is the fixed bin size with a minimum of 0 (by recommendation of [170]) and a maximum based on the maximum SUV_{max} of the population of the Train Set of the respective clinical context.

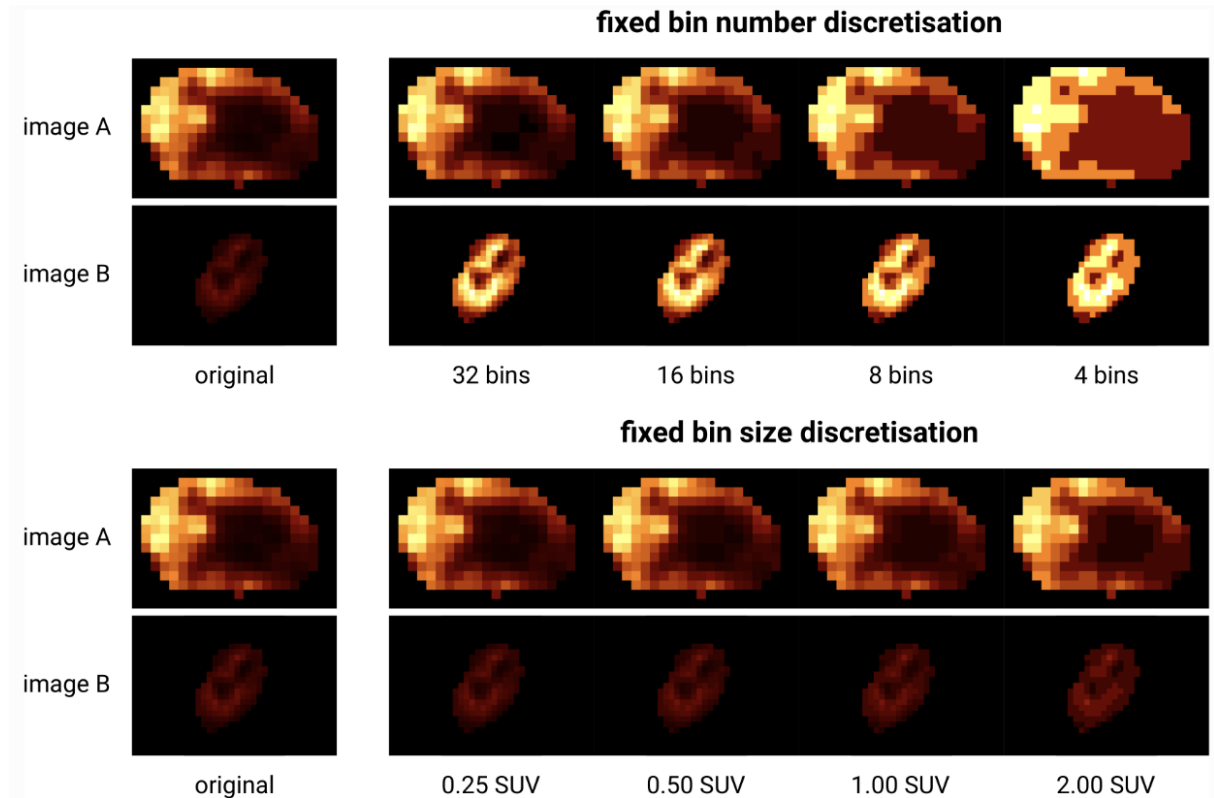


Figure 50. Example of both discretization methods in two $[^{18}\text{F}]\text{FDG}$ PET images with a SUV_{max} of 27.8 and 6.6 for image A and B, respectively [170].

As mentioned, the fixed bin size approach is selected with the aim of maintaining image contrasts and not losing information about the heterogeneous BM uptake pattern. Using this discretization method, the number of intensity values present in the discretised image must be taken into account depending on the aim of the analysis. Following (Eq. 14), Table 8 represents an example of the creation of the new values of the discretised image for a determined bin width (BW). As it can be seen, the maximum value of the number of intensity values directly depends on the selected BW.

N original	N discretised
$X_{gl,min} - [X_{gl,min} + BW]$	1
$[X_{gl,min} + BW] - [X_{gl,min} + 2*BW]$	2
$[X_{gl,min} + 2*BW] - [X_{gl,min} + 3*BW]$	3
...	...
$[maxSUVmax - BW] - maxSUVmax$	$maxSUVmax/BW$

Table 8. Example of discretised intensity values for fixed bin size discretization. N: intensity values.

Since the main objective of the BM quantification is evaluating its heterogeneity in all the clinical contexts, the selection of an adequate BW is essential, since a greater BW can relate some of the heterogeneities within the same bin, impeding differ metabolism variations. However, a compromise between the size of the bin and the level of heterogeneity pursued must be made, as the difference between the radiomics features extracted with the different BWs can be negligible. For this reason, four different BW are selected for extracting the radiomics features: (1) 0.5SUV, (2) 0.25SUV, (3) 0.125SUV and (4) 0.0625SUV.

4.3.3.1.2 Extracted Features

Using the different values of BW, radiomics features are extracted and then used to train the ML models. The radiomics features are divided into different groups [168,172,173]:

- **Image and ROI descriptors:** the first 18 extracted features correspond to the **mean**, **minimum** and **maximum** of the original image, the interpolated image, the interpolated ROI, and the ROI re-segmented in the image; and **number of voxels** and **volume** of the original ROI, the interpolated ROI and the re-segmented ROI in the image.
- **First Order Statistics:** features which describe the intensities distribution in the ROI. They are mainly based on the original histogram, so there is no need of discretization for their extraction. This group includes a total of 18 features: **energy**, **total energy** and **Root Mean Squared**, representing both of them the magnitude of voxel values, the second one scaled by the volume of the voxel; **entropy**, which interprets the randomness of the image values; **minimum**, **10th percentile**, **90th percentile**, **maximum**, **mean**, **median**, **interquartile range** (calculated from the 25th and 75th percentile) and **range** (calculated as the subtraction of the minimum to the maximum), which are simple measures obtained from the histogram of the original image in the ROI; **Mean Absolute Deviation** and **Robust Mean Absolute**

Deviation which measure the deviation of the intensity values; and **skewness** and **kurtosis**, which measure the asymmetry and the concentration of the distribution of intensity values; **variance** and **uniformity**, which measures the mean distance of each intensity value to the mean of the total histogram and the sum of squares of each intensity value, respectively.

- **Shape-based:** descriptors of the volume size and shape of the ROI. For the extraction of these features, the intensity value distribution is not important. The descriptors derive from a mesh, calculated following a standard. First, the vertices of the mesh are defined as the points in the middle of a voxel from the ROI and the nearest voxel outside the ROI. Applying this process iteratively, a triangular mesh is created. This group includes a total of 14 features: **mesh volume** and **voxel volume**, which measure the volume of the ROI with a more precise and more general approximations; **surface area**, **surface area to volume ratio**, and **sphericity**, which measure the area, the compactness, and the roundness of the mesh; **maximum 3D diameter** and **maximum 2D diameter** (slice, column and row), which measure the maximum diameter in the volume and in the mesh in the three directions, with a different feature for each direction; **major**, **minor** and **least axis length** from the ellipsoid which can confine the ROI; **elongation** and **flatness**, which estimates a ratio between the minor and major and the least and major axis lengths, respectively.
- **GLCM:** is the first of the textural matrixes described, concretely in 1973 by Haralick *et al.* [174], which represents how combinations of discretised intensities of neighbouring voxels are distributed in all the image directions. For these calculations, a neighbouring matrix of 26-connection is used. This group of textural features includes 24 features: **autocorrelation** and **joint average**, which measure the fineness and coarseness and the mean grey level for all of the intensities; **cluster prominence**, **shade** and **tendency**, which measures the skewness and asymmetry, the skewness and uniformity, and the groupings of voxels with similar intensity values in the GLCM, respectively; **contrast**, which estimates the local intensity variation; contrary, the **correlation** measures the linear dependency of intensity values to their voxels; **difference** and **sum average** measures the relation between the pair of occurrences with similar intensity values and the occurrences with different intensity values; **difference** and **sum entropy**, and **difference variance**, which measure the variability in the intensity values of the same neighbourhood, differing in that the variance gives greater weight to the variability of the intensity values furthest from the mean; **joint energy** and

joint entropy, which measure the homogeneous pattern and the randomness in neighbourhood intensity values, respectively; two features of **informational measure of correlation**, which measure the correlation between the distribution probabilities of two intensity values; **inverse difference** and its **normalised** version, **inverse difference moment** and its **normalised** version, which measure the local homogeneity; **maximal correlation coefficient**, which measures the complexity of the texture; **inverse variance**; **maximum probability**, for estimating the most predominant pair of intensity values; and **sum of squares**, which measures the distribution of a pair of intensity values in terms of the mean intensity level.

- **GLDM**: a matrix defined in 1983 [175] which quantifies the dependences of the intensity values of an image, being a dependence, the number of voxels connected to a central pixel with a similar intensity value and with a determined distance. This matrix allows to extract 14 different features for measuring: small dependences (**Small Dependence Emphasis**, **Small Dependence Low Gray Level Emphasis** and **Small Dependence High Gray Level Emphasis**); large dependences (**Large Dependence Emphasis**, **Large Dependence Low Gray Level Emphasis** and **Large Dependence High Gray Level Emphasis**); non-uniformity (**Gray Level Non-Uniformity**, **Dependence Non-Uniformity** and **Dependence Non-Uniformity Normalised**); the image variance and entropy of dependences (**Gray Level Variance**, **Dependence Variance** and **Dependence Entropy**); and the intensity values emphasis of the dependences independently of the size of the dependence (**Low Gray Level Emphasis** and **High Gray Level Emphasis**)
- **GLRLM**: is the matrix which quantifies the length, in number of pixels, of the segment with pixels with the same intensity value [176]. As for the GLCM, each feature is calculated separately for an angle. Then, GLRLMs are weighted by the distance between neighbouring voxels, summed and normalised. This group includes 16 features focused on measuring: short run lengths (**Short Run Emphasis**, **Short Run Low Gray Level Emphasis** and **Short Run High Gray Level Emphasis**); long run lengths (**Long Run Emphasis**, **Long Run Low Gray Level Emphasis** and **Long Run High Gray Level Emphasis**); non-uniformity of the image (**Gray Level Non-Uniformity**, **Gray Level Non-Uniformity Normalised**, **Run Length Non-Uniformity** and **Run Length Non-Uniformity Normalised**); the coarseness, variance and entropy in the runs (**Run Percentage**, **Gray Level Variance** and **Run Variance Run Entropy**);

and the runs in terms of low or high intensity values, independently of its length (**Low Gray Level Run Emphasis** and **High Gray Level Run Emphasis**)

- **GLSZM:** is a matrix similar to GLDM which quantifies grey level zones, instead of runs, in the image. In this case, the grey level zone is interpreted as the voxels with the same intensity value which are connected, using the 26-conection matrix [177]. Contrary to the rest of matrixes, only one matrix is calculated for all directions in the ROI, since it is rotation independent. From this matrix, a total of 16 features are extracted, grouped by purpose: assessing the small areas (**Small Area Emphasis**, **Small Area Low Gray Level Emphasis** and **Small Area High Gray Level Emphasis**); assessing the large ones (**Large Area Emphasis**, **Large Area Low Gray Level Emphasis** and **Large Area High Gray Level Emphasis**); evaluating the non-uniformity of the areas (**Gray Level Non-Uniformity**, **Gray Level Non-Uniformity Normalised**, **Size-Zone Non-Uniformity** and **Size-Zone Non-Uniformity Normalised**); to measure the coarseness, variance and entropy of the grey level zones (**Zone Percentage**, **Gray Level Variance**, **Zone Variance** and **Zone Entropy**); and the zones in terms of low or high intensity values, independently of its size (**Low Gray Level Zone Emphasis** and **High Gray Level Zone Emphasis**).
- **NGTDM:** a textural matrix corresponding to textural properties [178] which measures the difference between an intensity value and the average of its neighbourhood within a determined distance. This group includes 5 different features: **coarseness**, for intensity values uniformity; **contrast**, for variation between a voxel and their neighbours; **busyness**, for sharp or rapid changes between neighbours; **complexity**, for non-uniformity and with rapid changes; **strength**, for evaluating slow changes in the intensity but with larger coarse differences.

In order to explain the textural features extraction, i.e. those corresponding to the matrixes GLCM, GLDM, GLRLM, GLSZM and NGTDM, an example of the GLCM calculation is shown in Figure 51 based on [179]. As it can be seen, for the original image shown in Figure 51a, the combination of the intensity values of 2 and 3 in the direction 0° with 1 pixel of distance appears 3 times, rounded in red. The GLCM shown in Figure 51b, correspondent to this direction and distance, represents the number of times which is repeated each combination of intensity values, marked in red. Equally, the GLCM shown in Figure 51c, represents the same matrix but corresponding to direction 135° and distance of 1 pixel, marked in green the same combination of intensity values.

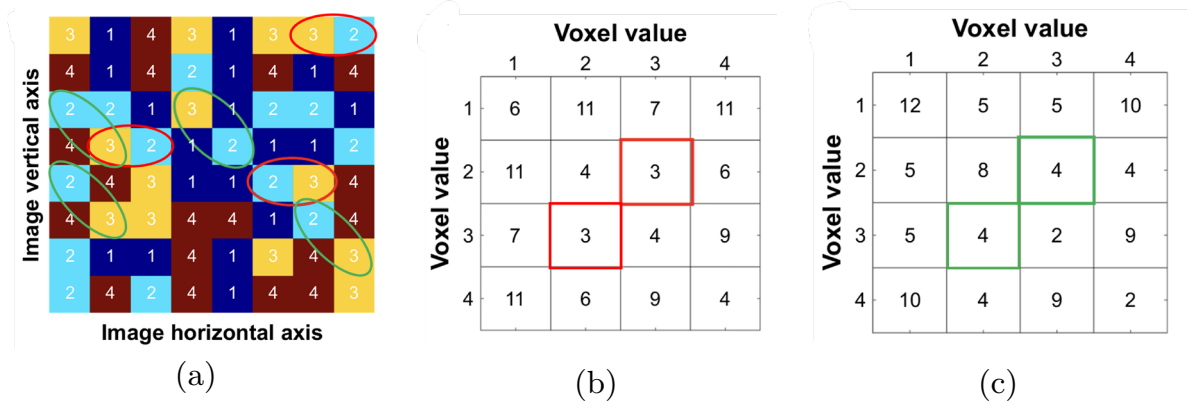


Figure 51. Example of the GLCM matrix calculation. (a) Representation of the different intensity values of an image, rounded in red the combination of intensity values 2 and 3, with a direction of 0° and distance of 1 pixel and in green the combination of the same intensity values with a direction of 135° and distance of 1 pixel; (b) GLCM for the pair of direction 0° and distance of 1 pixel, marked in red the number of repetitions for the intensity values 2 and 3, rounded in red in (a); (c) GLCM for the pair of direction 135° and distance of 1 pixel, marked in green the number of repetitions for the intensity values 2 and 3, rounded in green in (a). Adapted from [179].

Following the example, all the features corresponding to the matrixes are extracted. These features, along with the ones corresponding to the descriptors of the image, mask, shape and histogram, generate a total of 125 textural features. As mentioned above, these 125 features are extracted for four different values of BWs, forming 4 groups. In addition, from the original 4 groups of features (ORIG), another new three are created. For the creation of these new three groups of features, first, the ORIG is divided into its Train and Test Set (85%/15%), maintaining the proportions of each class in both sets.

4.3.3.2. Feature groups creation

As shown in Figure 48, from each BW and labelling, a different ORIG group of features is created. Considering the four different BWs, four different radiomics extractions are performed. In addition, depending on the clinical context, each of this extraction is used for developing a ML model able to distinguish each of the labelling method. Furthermore, since the number of extracted characteristics is large for the number of samples, a features selection or reduction is applied. In addition, due to the disbalance between classes, an oversampling method is also performed.

4.3.3.2.1 Univariate analysis

For the feature selection or reduction, a univariate analysis is performed using the Train Set as the ORIG samples. The main objective is to evaluate the potential of the extracted features, individually, for being part of the ML model training. However, only the Train Set samples are used in order to avoid bias caused by the

Test Set. This way, the Test Set does not affect the training of the model in any manner. Furthermore, in the clinical context where the number of samples is not enough to obtain an adequate model, this analysis allows to observe the possible potential of radiomics features for future works, such as in SMM. Depending on the number of classes, two different but equivalent methodologies are proposed.

First, a normality study is performed for each of the variables, independently of the number of classes. For this, the **Shapiro-Wilk** test [180] is used, which returns a $p\text{-value} < 0.05$ for features that follow a non-normal distribution, and a $p\text{-value} > 0.05$ otherwise. None other normality tests are used, since the number of samples is less than 2000 in any clinical context [181]. In the case of binary classification, the proposed methodology is shown in Figure 52. Statistically significant differences between groups are assessed with **Mann-Whitney** [182] and **T-test** [183] analyses depending on whether the distribution of the variable is non-normal or normal, respectively. For both tests, a variable is considered to have significant differences between classes if, and only if, the $p\text{-value}$ of the test is lower than 0.05.

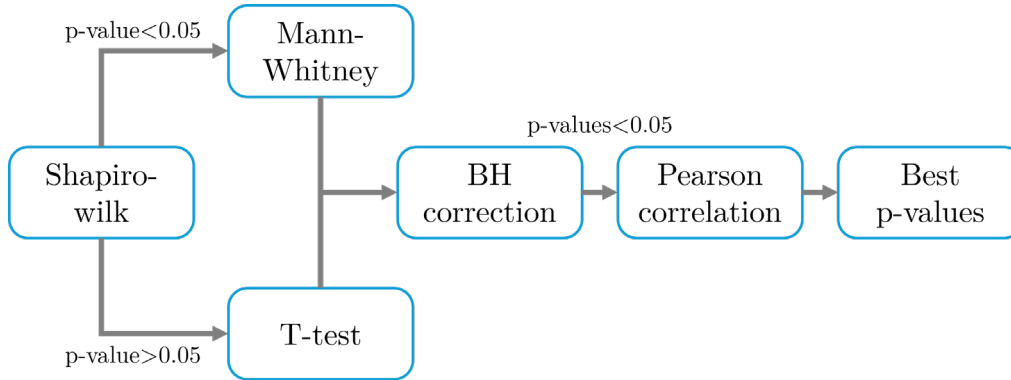


Figure 52. Pipeline of univariate analysis for clinical contexts with only two classes. BH: Benjamini-Hochberg.

In case of more than two labels, the methodology shown in Figure 53 is followed. Although the last-mentioned methodology is adequate for distinguishing three or more classes, in the present thesis only clinical contexts of a maximum of three classes are assessed. In this case, **Kruskal-Wallis** [184] and **ANOVA** [185] tests are used for features with non-normal and normal distribution, respectively. However, these tests only allow to assess whether there are differences between groups, but not between which groups. For this reason, and only if the $p\text{-value}$ from the previous test is lower than 0.05, i.e. there are significant differences between some classes, it is necessary to subsequently perform the post hoc tests, which assess the significant differences between each pair of classes, which are those already mentioned for two classes, i.e. **Mann-Whitney** and **T-test**, depending on the distribution of the variable.

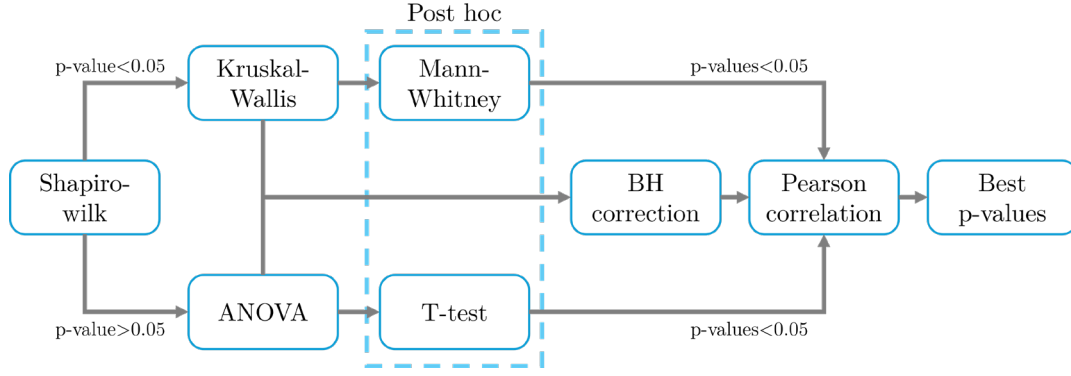


Figure 53. Pipeline of univariate analysis for clinical contexts with three or more classes. BH: Benjamini-Hochberg.

After these analyses, the methodology followed is shared between both proposals. Based on the premise that when multiple hypotheses are tested simultaneously (one hypothesis for each individually evaluated feature) there is a higher probability of committing false statistical inferences, a multi-testing correction is applied [186]. For each feature and its p-value obtained from the initially applied test, i.e. for those with 2 classes Mann-Whitney or T-test and for those with three classes Kruskal-Wallis or ANOVA, a multi-testing correction by means of the **Benjamini-Hochberg procedure** [187,188] is applied. This correction method is selected because of its control over the False Discovery Rate with its consequent better compromise between sensitivity and specificity than other correction methods, and its suitability for a large number of comparisons. In this way, all variables with a $p\text{-value} < 0.05$ in all the tests applied (Mann-Whitney, T-test, Kruskal-Wallis, ANOVA and Benjamini-Hochberg) are considered as variables with significant differences between classes. The p-value obtained after the Benjamini-Hochberg correction is always greater than that of the original test, so there is no way for a variable to have significant differences between classes if it was not significant in its initial test. Moreover, for a more correct analysis, only variables with significant differences between all classes (for cases where there are more than two classes) are considered significant.

For the features with significant differences between classes, in order to avoid redundant information, a **Pearson correlation** [189] analysis is performed. With this analysis, the correlation is calculated for each pair of remaining features. Of all the correlations, the one with the highest coefficient is selected and the variable with the highest p-value of the two is eliminated. Recursively, one variable is eliminated from each pair until the maximum correlation is 0.9.

Finally, in order to perform a reduction adapted to the number of samples of the database, one feature is kept for every 10 samples [190]. In this way, once all the above tests have been performed and if, and only if, the number of variables is still

higher than 10% of samples, a discarding of features is applied, leaving only those with the lowest p-values.

After applying the whole process of features reduction to the four ORIG groups of features, i.e. in their Train Sets, once per each BW, a new group for each of them is obtained, named RED group. Although the samples to develop the reduction are the ones corresponding to the Train Set, the Test Set is posteriorly reduced, removing the same features which have been removed in the Train Set. As explained, the RED group of features has the same number of samples than the ORIG, the only variation is the number of features presented. In the case that none of the features obtain a $p\text{-value} < 0.05$ after the multi-testing correction by means of Benjamini-Hochberg procedure, i.e. there are no variables with significant differences between classes, the RED group is not created. Although the number of features change for this RED group, there is still a disbalance in the number of samples of each class, which can cause an overfitting of the models to the majority class.

4.3.3.2.2 Oversampling

The classes imbalance is a very common situation in biomedical contexts. The main problem of this imbalance is the way it affects the models training. When the database is highly imbalanced, the model ignores the minority class, since the performance could be better if everything is classified as the majority class [191]. One of the main solutions for addressing the problem is re-sampling the original database. In the re-sampling process, there are two different methodologies: undersampling the majority class or oversampling the minority class. However, the undersampling technique is limited to large databases.

For the oversampling, Chawla *et al.* [192] proposed in 2002 the Synthetic Minority Over-sampling TEchnique (SMOTE). This technique creates synthetic samples of the minority class. Drawing from real data, they create new data by selecting each minority class sample and inserting synthetic examples throughout the segments which join all the k minority class neighbours. The number of neighbours is randomly selected depending on the amount of synthetic data required. Those synthetic examples are created multiplying the difference between the feature vector and the nearest neighbour by a random number between 0 and 1, adding it posteriorly to the vector under consideration. This way, the region which represent the minority class is more general, making the classifier to create less concrete decision regions.

Trying to avoid the effect of imbalanced databases to the ML models, a SMOTE process for increasing each minority class by 33% is applied to the ORIG and to the RED groups of features, again only in the Train Set. The main purpose of not being applied into the Test Set is not to fake the test results by trying to classify synthetic data. This way, from each ORIG group of features, two new ones are created: an

oversampled one (OS), consequence of increasing the number of samples from the ORIG group; and the reduced and oversampled group of features (R_OS) which is the reduction of the OS group, applying the same reduction as the one used to obtain the RED group. It must be considered that, if RED group is not created for lack of features with significant differences between groups, the R_OS is also no created.

4.3.3.2.3 Summary of groups creation

After both processes, features reduction and oversampling technique, finally a total of four different feature groups are created. In order to easier understand the groups creation, a scheme with the creation of each group is shown in Figure 54, which represents a part of the Figure 48. Following this scheme, the four groups are created by its corresponding Train and Test Sets, being some of them shared between groups:

- **ORIG**: formed by the ORIG Train set and the ORIG Test set, i.e. the sets from the original radiomics extraction for the correspondent BW, divided by 85% and 15%, maintaining the proportion of classes.
- **RED**: formed by the RED Train Set and the RED Test Set, i.e. the Train Set obtained after the univariate analysis for features selection applied in the ORIG Train Set, and the RED Test Set result of the removal of the same features from the ORIG Test Set.
- **OS**: formed by the OS Train Set and the ORIG Test Set, since there is no need in oversampling the test because it can cause bias in the test results.
- **OS_R**: formed by the OS R Train Set, i.e. the OS Train Set after the removal of the same features than those removed for the RED Train Set creation, and the RED Test Set, again to not cause bias.

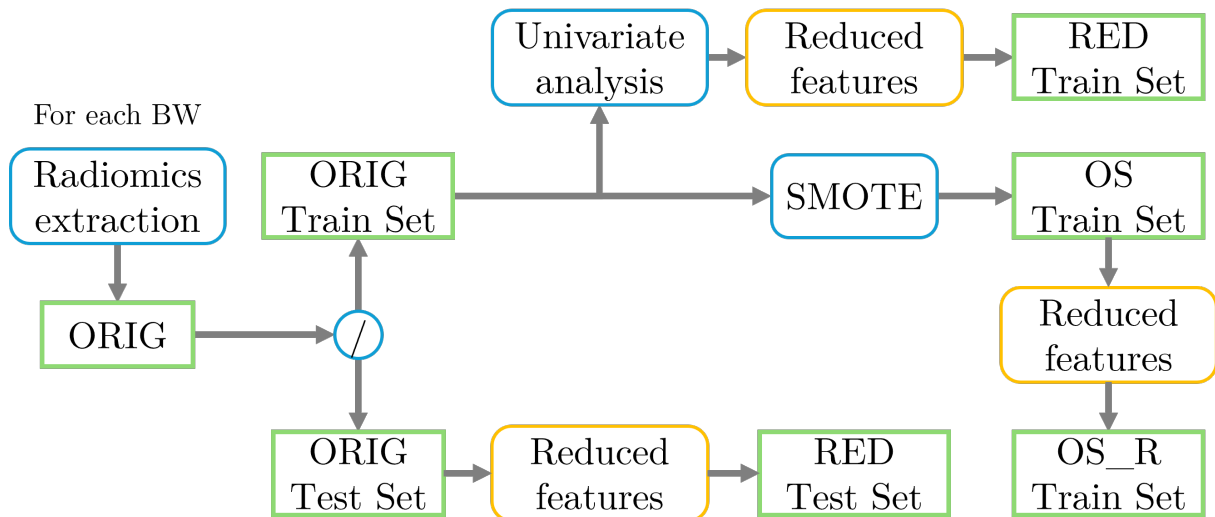


Figure 54. Summary of groups creation from each of the radiomics extraction, i.e. for each BW. The square for reduced features is obtained only once (from the ORIG Train Set), and then, it appears in other parts of the scheme representing that, although no other univariate analysis is applied, the same features removal is applied.

This way, there are 4 different groups for each BW. Having 4 BWs, there are 16 different groups, which are used for training and testing the ML models in order to classify the corresponding classes, varying in the labelling used.

4.3.3.3. Machine Learning implementation

ML is a science focused on implementing computational systems with the ability to improve automatically from experience [193]. As early as 1997, the potential of ML had been demonstrated in multiple disciplines [194]. There are many ML systems divided in the following categories [195]:

- Supervised or unsupervised, depending on the human supervision of the training. Commonly, the supervised models have a target defined by the human, while the unsupervised models study the data trying to group the data in different clusters.
- Incrementally or batch learning, if the models are constantly learning or they learn from a concrete collection of data, respectively.
- Instance-based or model-based, depending on the way they learn, by comparing known and new data or detecting patterns, respectively.

These categories are not exclusive between them, they can be mixed. For the present thesis, supervised models based on batch learning are used, since the target is previously defined by the labelling methods and the database is limited to a concrete batch. The classification ML models used for the BM quantification are [195–198]:

- Decision Tree (DT) [199]: is a model which classifies a target by learning simple decision rules based on data features. However, the depth of tree can be selected, being a more complex and fitter model when it is deeper. The main advantages of this ML model are that it is easy to understand, requires minimal data preparation, and it is an efficient, flexible and robust model. However, it is a model easy to be overfitted, unstable against light variation in the data, difficulty for learning complex contexts, and with a high bias for imbalance data.
- Support Vector Machines (SVM) [200]: is a ML model based on the main idea that the feature data is non-linearly mapped to a huge-dimension space of features. The transformation to high-dimensional space is based on a kernel function, which allows saving computational resources by making the calculations in the high-dimensional space without calculating the actual coordinates of the transformed data. The kernels used in this classification problem are: radial basis function (RBF) and sigmoidal (SIG), useful for data with more complex relationships; and linear (LIN), for data easy to separate linearly. Contrary to the DT, this model is useful in high dimensional spaces,

being still good in cases of a dimension higher than the number of samples. In addition, it is a versatile model since can adapt to different dimensional shapes thanks to the use of different kernel functions.

- Random Forest (RF) [201]: is a model which combine multiple DTs, training each of them with a random sample of the data, reducing the variance and the possibility of overfitting. The main advantages of this model are the robustness, since it is less sensitive to data noise than a simple DT; and the ability to evaluate features importance in the prediction.
- Gradient Boosting (GB) [202,203]: similarly to the RF model, is a model based on the combination of multiple DTs. In this case, the trees are sequentially combined in order to correct the mistakes of the previous trees instead of optimizing the general performance. The main advantages of this model are the high performance, since the mistakes are recursively corrected, flexibility to different types of data and loss functions, and efficiency with imbalanced data. However, it tends to overfit with an excessive number of iterations, is time-consuming, needs high computational resources and is sensitive to outliers.
- Logistic Regression (LR) [204]: estimates a class probability as a non-linear function of the independent variables. Although it was first defined in 1944, it is still considered as a useful model thanks to its advantages: easy interpretation of the results, since are estimated as a probability of belonging to the class, and the simplicity and efficiency, since it requires less computational resources than complex models. However, it is not adequate for linear models, is inefficient for imbalanced data and is sensitive to outliers.
- Bernoulli Naïve Bayes (NB) [205–207]: is an interpretation of the Naïve Bayes model according to multivariate Bernoulli distributions, i.e. it assumes that the data is composed only by binary features. The Naïve Bayes model is based on the Bayes' theorem, which classifies the data depending on conditional probabilities between the, assumed as independent, features. Despite its simplicity and efficiency, the model always assumes the independence between variables, which is usually improbable.
- k-Nearest Neighbours (k-NN) [208]: is an instance-based model, i.e. there is not an internal model. The idea is to collect the training data and evaluate the distance of the new samples to the neighbours. The main advantages are based on its simplicity, due to the lack of necessity of more calculations apart from the distances; the model makes no assumptions about decision function; and the efficiency in classification problems with a few classes and separated data. Although the calculation is based only on the distances, the model

requires many computational resources in case of having several samples. In addition, the model is sensitive to the selected k , and to a high dimensionality space.

- Multi-Layer Perceptron (MLP) [209]: is a supervised model based on the human brain. The model is composed of multiple layers, connected between them, with multiple nodes or neurons. Through the training, the weight of the connection between the layers is adjusted to minimise the loss. The main advantages are the capability to learn non-linear models and the flexibility, since the selection of layers is unlimited. The disadvantages are that the initialization weights, which are randomly selected, can lead to different models, even non-convergent ones, and the model requires a huge volume of data to learn correctly. For the present thesis, simple versions of this model are trained, with 100 and with 200 neurons in only one hidden layer (MLP1 and MLP2), since posteriorly CNNs are trained.

As explained, the models described above can be useful for the BM quantification, since all of them are efficient. However, since the features are based on the radiomics extraction, no relationship or independence between them is assumed. Although the models are trained and tested with the same data, i.e. the same Train Sets and Test Sets, some processing is necessary in order to not prioritise some features against others.

4.3.3.3.1 Data processing

Although the majority of models are efficient in problems with many features, in the vast majority, using a very different range of values for the different variables can cause a disproportional influence on the model [195,203]. There are two commonly used ways of scaling the features: (1) normalization, where values are rescaled into the range 0-1, following the equation (Eq. 15); and standardization, following the equation in (Eq. 16) where μ represents mean and σ the standard deviation. The standardization against the normalization converts the data into a different range of values, not limited to 0 and 1, and is less affected by outliers. For this reason, it is the selected way to scale the databases.

$$X_{normalized} = \frac{(X - \min X)}{(\max X - \min X)} \text{ (Eq. 15)} \quad X_{standardized} = \frac{(X - \mu)}{(\sigma)} \text{ (Eq. 16)}$$

In terms of the NB model, as mentioned, the Bernoulli interpretation is implemented. For this interpretation, the variables are assumed to be binary, i.e. acquiring only values of 0 or 1. The optimal cut-off values for features defining high- and low-value sub-groups are obtained through the application of the ROC and the closest-to-(0,1) criterion. Then, each of the variables are converted into its binary version, each of them with each particular cut-off value.

4.3.3.3.2 Hyperparameters optimization

Once the data is prepared, a hyperparameter optimization is applied for each model, in order to obtain the best combination of hyperparameters. The scorer used for measuring the optimal combination of parameters is `F1_score_macro`, i.e. mean of `F1_score` for each class, but it must be considered that this metric is measured for the validation of the model, i.e. inside the training process. The mentioned metrics are explained posteriorly in the subsection 4.3.5. *Performance assessment*.

In order to create a reproducible model, the same randomization seed is defined for all the models. In addition, for each model, considering their characteristics, a grid with different values for each parameter is defined, but these parameters are always the same. The values of these parameters are defined following the description of each of them for the scikit-learn library [198], and they can be consulted in *Annex A. Hyperparameters grid*.

Then, with a process of cross-validation with 5 folds, each combination of parameters is tested. Cross-validation is a technique used to measure model performance by assessing possible model overfitting. For this purpose, the train set is divided into k subsets with the same number of samples. In each of the k iterations, $k-1$ subsets are used for training and the remaining subset for validation. In this way, each iteration has a different validation set, as shown in Figure 55. Finally, the performance of the model is measured as the average of the performances for each iteration, which allows to evaluate the generalisation of the model and possible overfitting [210,211].

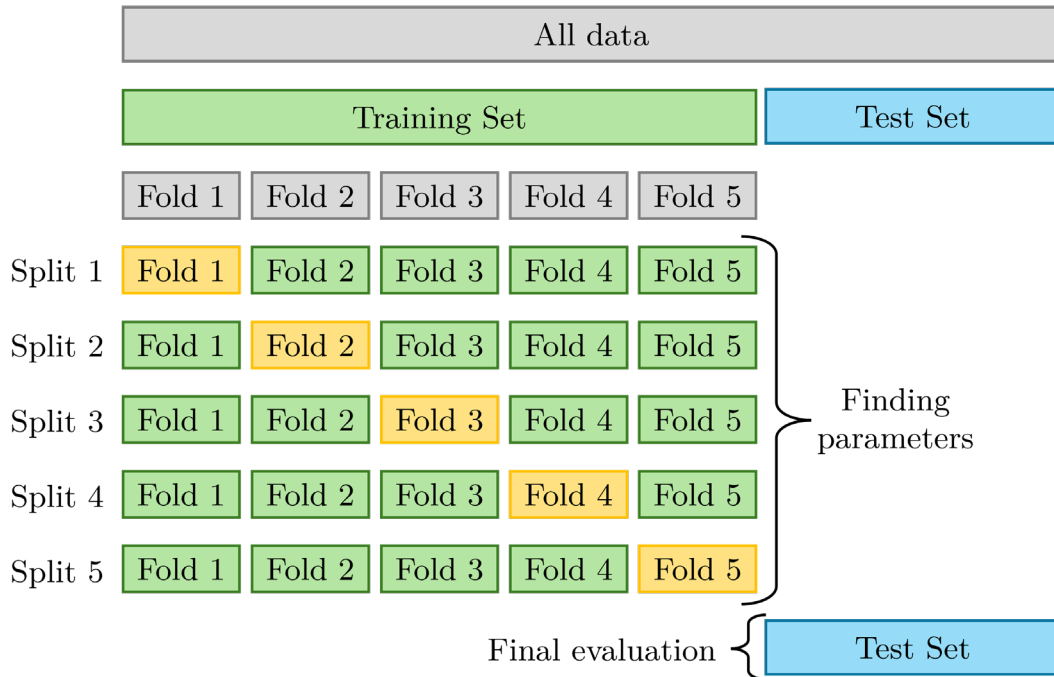


Figure 55. Cross-validation illustration adapted from [211]. (Grey) unassigned to any process (training, validation or test) data; (green) training data; (yellow) validation data; and (blue) test data.

Although the majority of models improve with the iterations, due to the number of samples, more than 5 folds makes not representative train and validation sets. Moreover, the time of computation is lower with a lower number of folds, making the training process faster with only 5 folds. As no randomization processes are used at any stage of hyperparameters optimization or training, the model is fully reproducible using the same database.

4.3.4. Quantification using Convolutional Neural Networks

Similarly to ML models quantification, using the $[^{18}\text{F}]\text{FDG}$ PET imaging, i.e. its SUV map, and the BM mask obtained with the BM segmentation DL model, different DL models, concretely CNNs, are used as classifiers. For training and testing the CNNs, the patients present in both sets, Train Set and Test Set, are the same as for the previous quantification. However, in this case, the input to the models is the image, there is no feature extraction. There are three variants of the same model architecture depending on the input used (Figure 56):

- I. Original model architecture for the 3D input ($N \times N \times N$), i.e. the SUVmap in the BM region.
- II. Adapted model architecture for a pseudo-RGB (pRGB) input ($N \times N \times 3$). The input of the model is the same 3D volume, but a block of layers is included at the beginning of the model architecture to convert the last dimension to just 3, like an RGB image.
- III. Adapted model architecture for a 2D input ($N \times M$). The input of the model is a MIP image calculated for the coronal plane.

Although the pRGB model is adapted to a 2D CNN, the input size of the model is different, forcing to create two different structures for the pRGB and MIP models. In this way, a total of 3 different adaptations are considered: 3D, pRGB and MIP models.

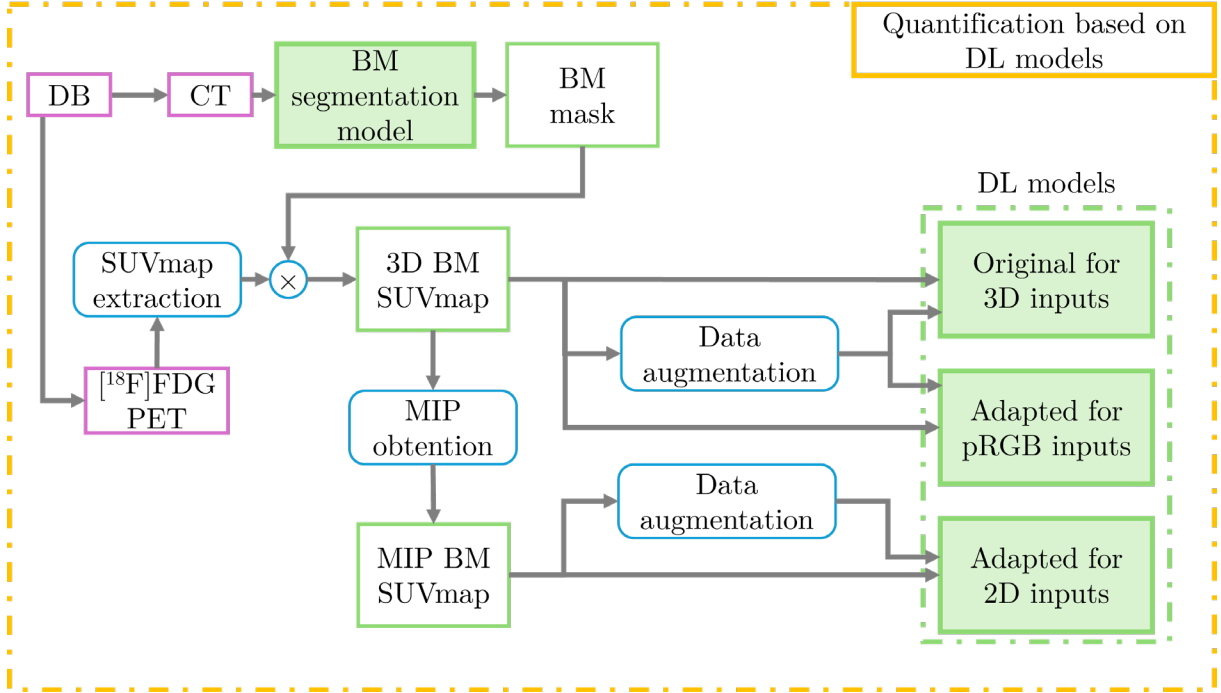


Figure 56. Pipeline of the quantification based on DL models. DL models square content three variants of the same model, based only on its input.

4.3.4.1. Inputs creation

Following the scheme in Figure 56, first the 3D inputs are obtained. Drawing from the original multiplication of both, the SUVmap and the BM mask, the volume is processed:

1. Resize. Due to the limitations of computational resources, and considering that a higher volume, i.e. a greater number of voxels, implies a higher number of features extracted in the CNNs, the volume is resized to $\frac{1}{4}$ of its original resolution (for two out of the three dimensions) to avoid resources issues. The original volumes have dimensions of $512 \times 512 \times N$ voxels, being N variable depending on the acquisition protocol and the height of the patient. After the resize process, the volumes are all converted into a size of $128 \times 128 \times 128$ voxels. The third dimension has to be the same for all patients, since the CNNs must be trained for a particular image size.
2. Normalization. Following the main idea of the radiomics feature extraction, and to respect the functional nature of the PET, the normalization is applied between 0 and the maximum SUVmax of the population of the respective clinical context, obtained only from the Train Set.
3. Conversion to integers. Again, due to the computational limitations, the dynamic range of the image is adapted into 0 to 255 to, posteriorly, convert

the image into integers of 8 bits. This way, the created mask is less heavy than if the type were a float.

Then, considering that RGB CNNs are 2D although they have 3 channels as the third dimension, the models are adapted to receive a 3D image, go through a convolutional block (Table 9) to convert the 3D image into a pRGB one, i.e. with dimension of 128x128x3, and then applied the original architecture of the CNN but adapted to a 2D classification. The convolutional block is composed of 4 layers which reduce the dimensionality by its half, i.e. from 128 to 64, from 64 to 32, from 32 to 16, until the last layer, which convert the image into a pRGB one. In this way, the input in the 2D model is an image of dimensions 128x128x3.

Layer	Input shape	Output shape
Conv2D, 64, (3x3), ReLU	128x128x128	128x128x64
Conv2D, 32, (3x3), ReLU	128x128x64	128x128x32
Conv2D, 16, (3x3), ReLU	128x128x32	128x128x16
Conv2D, 3, (3x3), ReLU	128x128x16	128x128x3

Table 9. Representation of the layers of the convolutional block for the pRGB adaptation.

Then, for the MIP creation, the original volume resulting from the SUVmap and BM mask multiplication is projected into coronal plane, maintaining its maximum value in this projection. Having this 2D image, a similar processing than the 3D one is applied. The resize step is slightly different, since it is applied due to the necessity of the networks of having a certain input size, the volume is resized to Mx512, being M the average number of slices of the Train Set of the correspondent clinical context. In this case, the other dimension is not reduced since, being a 2D image, the number of parameters of the network and, consequently, the required computational resources, are hugely less than for 3D images. After this step, the normalization and conversion to integers is exactly the same as explained for 3D inputs.

The different adaptations of the models are trained individually on the original batch of images and on a version of the same batch with data augmentation. The data augmentation process is the same for both 3D and pRGB and adapted to a 2D for the MIP images.

4.3.4.1.1 Data augmentation

With the increased use of DL for biomedical tasks, such as segmentation and classification, the problem related to the limited databases is being extendedly studied. Data augmentation is a popular technique to create new samples from the real ones, when the available data is not large enough to generalise the model and avoid the overfitting [212]. For the present thesis, the data augmentation process is

based on simple transformations: (1) rotation of a random angle between -45° and 45° ; (2) translation of a random number of pixels between -10 and 10; (3) randomised horizontal flip, i.e. a random factor decided whether it is applied; and (4) zoom function with a zoom factor between 0.8 and 1.2, which can zoom out or in of the image. No intensity value transformations are applied to respect the idea that a higher uptake is related to an increase in the metabolism due to a functional change. All of the transformations are consecutively applied in the slices of the 3D image, applying the same transformations to all of the slices to create a realistic volume, or in the MIP image. For the pRGB variants, the data augmentation process is applied to the 3D image, i.e. before the convolutional block, so there is no data augmentation of the pRGB images specifically. In this way, six different types of input are used to train each of the models.

In addition to randomise the factors of each transformation, the data augmentation is applied in the training, i.e. on-the-fly, being random for each iteration, in order to generalise the model as much as possible. This way, any model with their correspondent combination of factors is trained with a random train set, making unreproducible models.

4.3.4.2. Deep Learning models

DL is a subfield of ML, which is drawn from neural networks in 2006 and is based on deep architecture of learning. This deep architecture consists of several layers between the input and output layers, which implies many units of non-linear information processing for feature learning and pattern classification. Some of the advantages of DL are that it can be applied to almost any context; features are automatically learned, so there is no need for a design of features, producing robust models against natural variations; and due to the possibility of transferring learning, i.e. using the same model for different contexts, it is adaptable to issues with not enough available data [213,214].

The main idea of neural networks is to replicate human brain behaviour. For this reason, the basic computational element of DL models is the neuron which, with some internal parameters, receives external inputs to produce outputs, as shown in Figure 57 [213].

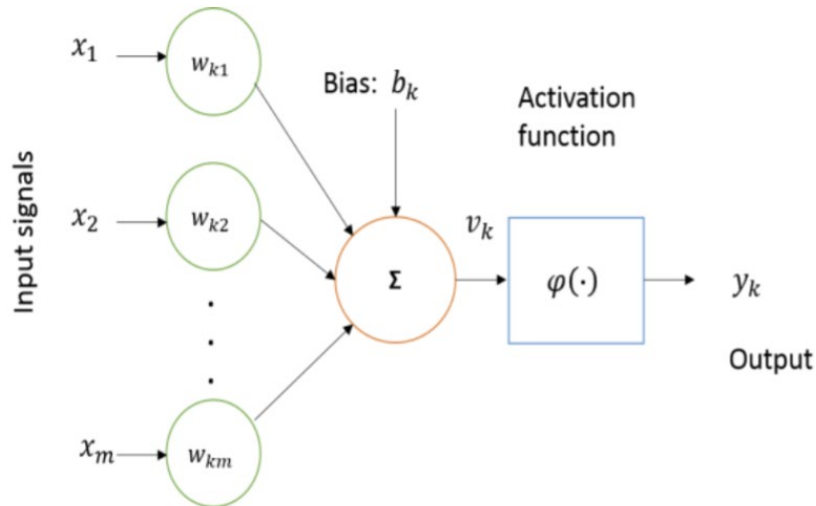


Figure 57. Model of a neuron [213]. x : inputs of the neuron; w : synaptic weights for each input; v_k : linear combination of the inputs; y_k : output; b_k : bias added to the v_k .

Drawing from this neuron unit, neural networks are based on Multi-Layer Perceptron, i.e. a network which contains one or more hidden layers with a determined number of hidden neurons [213]. From this basic architecture, emerge the idea of CNN proposed in 1988 by Fukushima [215]. The overall architecture of those CNNs is shown in Figure 58.

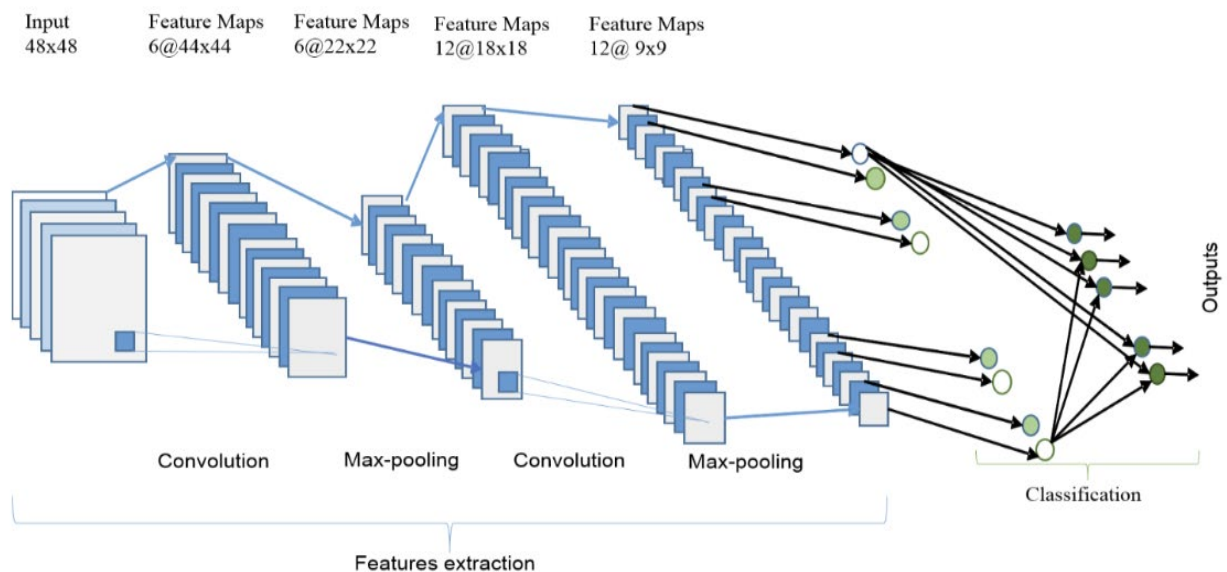


Figure 58. Overall architecture of the CNN with its respective layers: input, convolutional, max-pooling and classification [213].

Although several architectures have been created in the last decades, all of them shared the same characteristics. They are divided in two main parts: feature extractor, composed by extraction layers which receives the output of the previous one as input; and the classifier. The layers of the CNNs are commonly of three types [213,216]:

- Convolutional layer: convolves the output of previous layers with determined kernels. The output goes through an activation function to form the output feature map. The most commonly used activation functions are SoftMax, represented by (Eq. 17) [196], where i are the different classes to distinguish (with a maximum of K), and z_i is the output model value for the class i , guaranteeing a total sum of probabilities of 1 for all the classes; the sigmoid, visually represented in Figure 59a; and Rectified Linear Unit (ReLU), represented in Figure 59b, and the most used for convolutional layers.

$$sm(Z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \text{ (Eq. 17)}$$

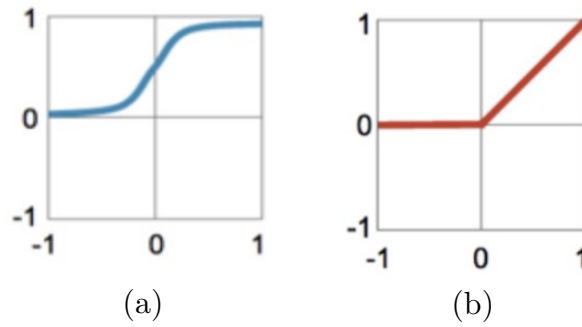


Figure 59. Representation of (a) sigmoid and (b) ReLU activation functions [213].

- Pooling layer: down-samples the input maps. Although the number of input and output feature maps remains the same, their size are reduced, depending on the size of the used kernel. There are two types of pooling layers (Figure 60): average pooling or max pooling. In both cases, the average or maximum of a patch with the size of the kernel, is selected to represent the whole patch in the new reduced feature map. The max pooling layer is more commonly used since it is effective in detecting shape variations.

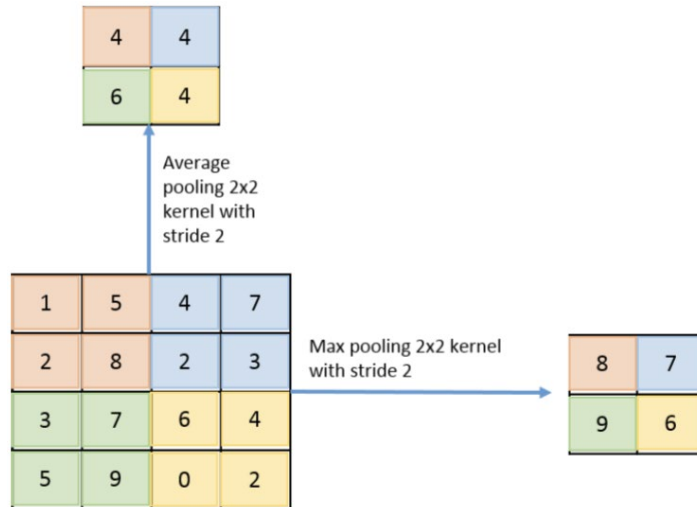


Figure 60. Representation of the average (top) and max (right) pooling operations [213].

- Classification layer: is a fully connected layer which estimates the score of each class from the extracted features obtained from the layers of the feature extractor. Commonly, a SoftMax activation is used. They are expensive in terms of computation.

Following the described architecture and combining differently the explained layers, there are several CNNs architectures which have been proposed and tried in several fields. Due to their capacity of pattern recognition for image quantification, they seem to be an adequate technique for the BM quantification.

4.3.4.2.1 Architectures

As mentioned before, the different combination of the common layers of a CNN, produces different architectures. For the present thesis five different CNNs are used for the classification task. However, all of them are adapted to computational resources. They are explained as follows [213,216]:

AlexNet

AlexNet was presented by Krizhevsky *et al.* [217] in 2012. Compared to the CNNs described until this moment, this architecture achieved the best accuracy performance in visual recognition and classification. Its architecture, shown in Figure 61, is based on two blocks of convolution, max pooling and local response normalization, followed by three convolutional layers with a ReLU activation, and two fully connected layers, finishing with a classification layer with a SoftMax activation. Due to the complexity of its training, the computation is very limited to the computational resources.

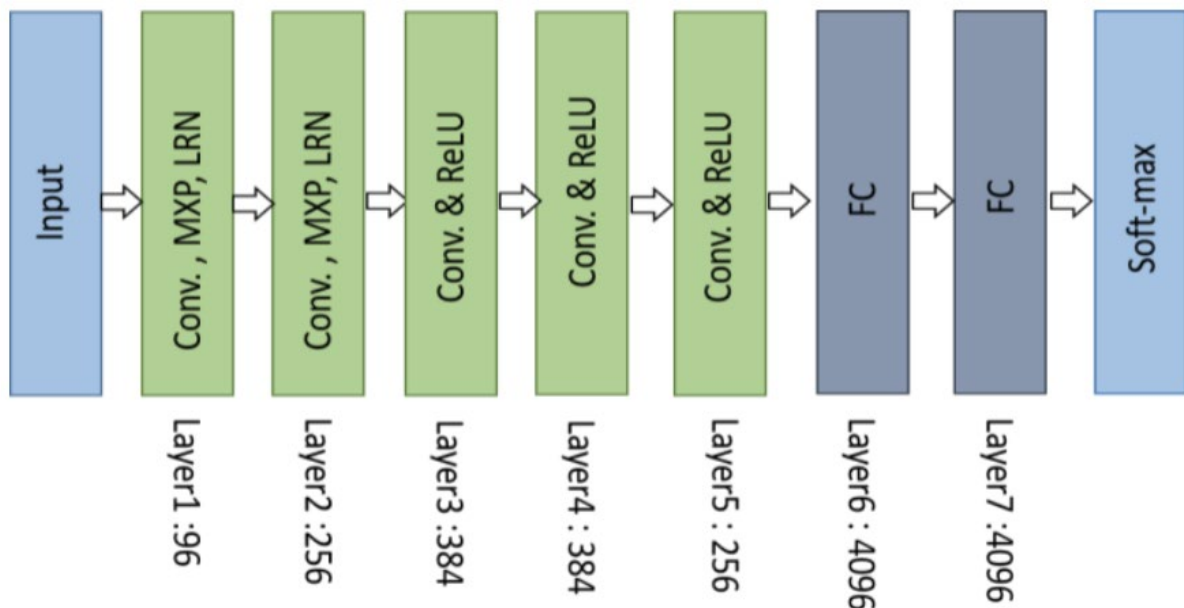


Figure 61. AlexNet architecture [213,217]. Conv.: convolutional layer; MXP: max pooling layer; LRN: local response normalization; FC: fully connected. In the bottom part, the number of filters of each of the feature maps are shown.

VGG16

The Visual Geometry Group (VGG) presented an architecture in 2014 [218] shown in Figure 62, which consists of two consecutive convolutional layers with a ReLU activation function, followed by a max pooling layer, iteratively repeated, and multiple fully connected layers also with a ReLU activation. After these layers, a classification one with SoftMax activation is the end of the architecture. This architecture has many versions, depending on the number of convolution layers, being 13 for the VGG16. The complexity of the architecture, caused by the several layers, can be simplified by reducing the size of the filters.

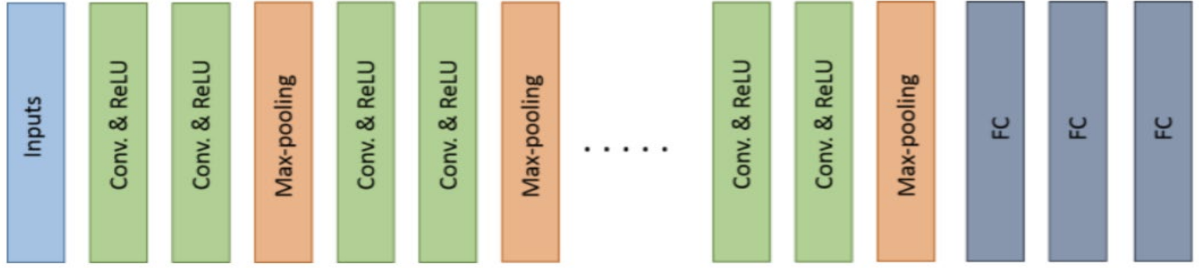


Figure 62. VGG architecture of the features extractor [213,218]. Conv.: convolutional layer; FC: fully connected.

Residual Network

A year later, in 2016, He *et al.* [219] proposed the Residual Network (ResNet) architecture with the intention of designing an ultra-deep network. This ultra-depth was pursued following the theory that the deeper the layer, the more new features the network learns. The main concept of this architecture is the residual block (Figure 63), which allows the network to learn the difference between the desired output and the input, instead of the real output. This difference is named residue, and it is possible to obtain through skipping one or more connections, allowing the original input to be an input for posterior layers without modifications. This way, the model can correct the modifications developed.

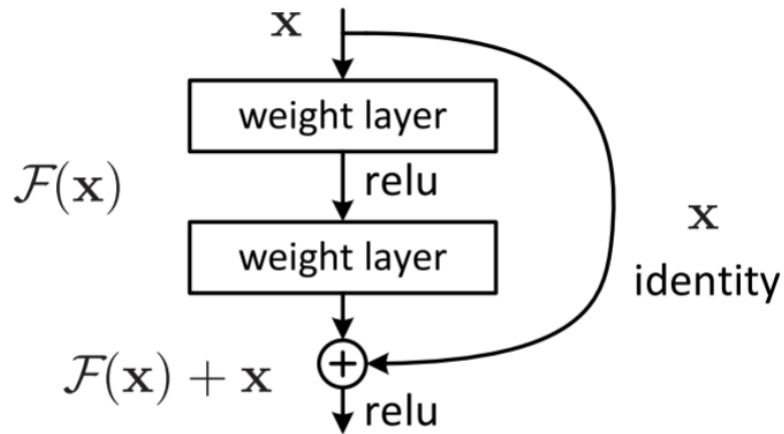


Figure 63. Residual block representation [219].

ResNet is implemented combining several residual blocks, as seen in Figure 64. This combination implies different numbers of layers, being the most popular the ResNet50 version, with 49 convolutional layers and one fully connected layer at the end. In spite of its deepness, this architecture is less computationally complex than many others. Additionally, it showed an error lower than the human eye on the test set of the ImageNet classification challenge (3.57%) [220], against 16.4% and 7.4% of the AlexNet and VGG-16 CNNs, respectively.

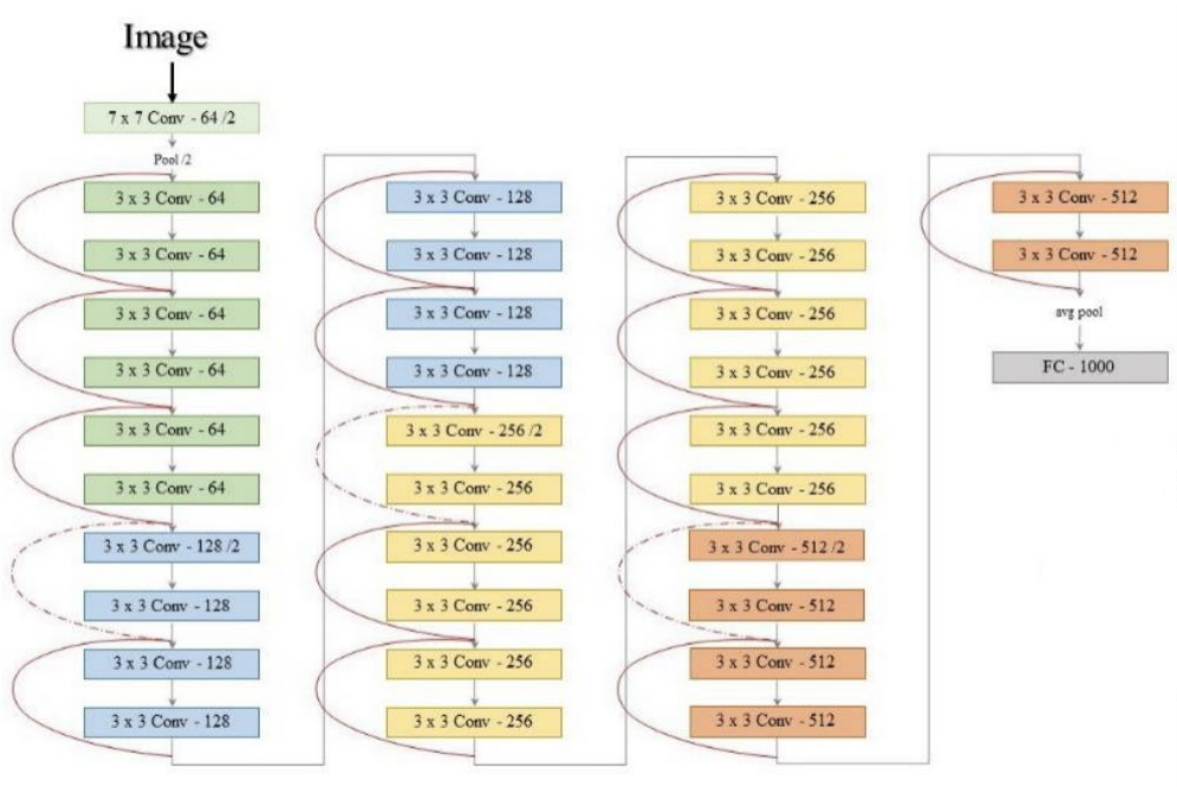


Figure 64. Example of a ResNet architecture [216].

Densely Connected Network

Based also on blocks, in 2017, Huang *et al.* [221] proposed the Densely Connected Network (DenseNet) for avoiding the problem of the rest of deep networks, the possibility of vanishing information. With this aim, a dense block is designed, shown in Figure 65, consist of densely connected layers, each of one receiving as input the output of all the previous ones. The main idea of designing these blocks is to transfer the information between the layers, reusing features and reducing the possible error.

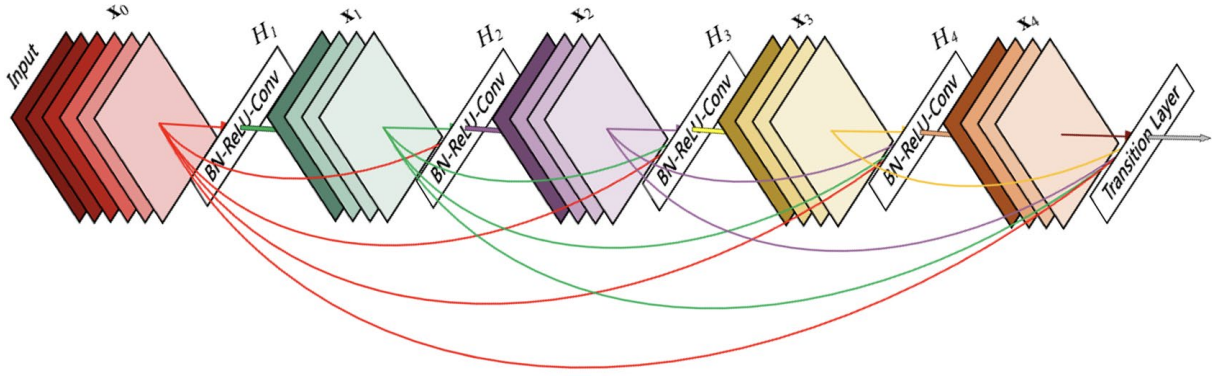


Figure 65. Dense block representation [221]. BN: Batch Normalization.

Then, the different dense blocks are connected by transition blocks between them, as shown in Figure 66. The transition block is composed of a convolution and max pooling layers to control the number of parameters and, consequently, the computational complexity.

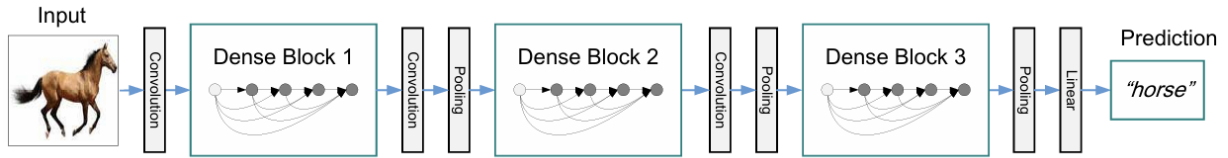


Figure 66. Example of DenseNet architecture representation with 3 dense blocks [221].

3D RPET NET

Besides the mentioned CNNs, which are some of the most known in the DL field, a network proposed by Amyar *et al.* [222] in 2019 is also used for the BM quantification in order to apply a less complex architecture. This 3D CNN was developed for predicting the treatment response in patients with oesophageal cancer after radio-chemotherapy, achieving a maximum AUC of 0.74. As shown in Figure 67, this architecture is also formed by convolutional, pooling and fully connected layers. However, this architecture is simpler since only six layers are forming the network.

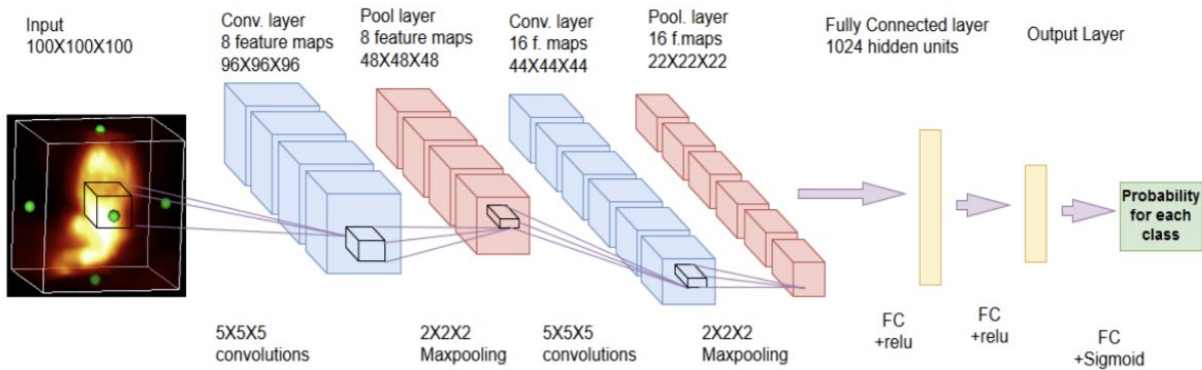


Figure 67. 3D_RPET_NET architecture [222].

4.3.4.2.2 Parameters selection

The original definition of the architectures is respected, including the kernel size and strides of the layers. Only the number of filters or channels of each feature map is change due to limitation on the computational resources. For all of them, a parameter F is defined as the features of the first feature map, and then, the rest of them are computed as multiple of the parameter F, i.e. 2F, 4F, 8F, etc. Following this structure, several values of F are tried: 1, 2, 4 and 8. No higher values are tested due to the mentioned limitation.

In addition to the F parameter, there are other characteristics that must be selected for training the model. First, for optimizing the model, the Adam optimiser is used since, according to Kingma *et al.* [223], the method is “*computationally efficient, has little memory requirement, [...], and is well suited for problems that are large in terms of data/parameters*”. Besides the optimiser, the minimization of the loss function (categorical cross entropy) is desired. Both parameters are selected since they are the most commonly used for CNNs in the optimization during the training.

For obtaining the version with the lowest loss, a checkpoint function is used. This way, for the 1000 epochs of training which are arbitrarily chosen, the model starts training and, when an absolute minimum is found, the resulting model is overwritten. Anytime the loss decreases more than the previous absolute minimum, this new loss is converted into the new absolute minimum and the model is saved again, overwriting the previous one to avoid memory issues. In addition, an early stopping function is also used to avoid training unnecessary epochs, in order to save time. In this way, if the loss does not decrease in 50 epochs, the training stops. Consequently, if a model is incapable of converging, the model training stops early.

Finally, in order to adapt the model to the class imbalance of the databases, a parameter called ‘class weight’ is selected. This parameter allows the model to give more importance to the minority classes. For the present thesis, three different ‘class weight’ definitions are used: (A) none, for learning from the database, without giving more attention to any class; (B) calculating each class weight as the division of the total number of samples by the number of samples of the concrete class as (Eq. 18); and (C) calculating each class weight as the division of the number of samples of the majority class by the number of samples of the concrete class as (Eq. 19).

$$cw(i) = \frac{NT}{Ni} \text{ con } NT = \sum Ni \quad (\text{Eq. 18}) \qquad cw(i) = \frac{Nmax}{Ni} \quad (\text{Eq. 19})$$

4.3.5. Performance assessment

For evaluating the performance of both quantifications, based on ML models and based on CNNs, the used metrics are the same. As shown in the studies included in Table 4 and Table 6, the most used metrics are recall or sensitivity, specificity and AUC. Due to their definition, the recall and specificity are good metrics for evaluating the capacity of the model to predict positive and negative samples, respectively, but are not enough for evaluating the total performance of the model. For this reason, AUC is more typical for the assessment of model performance. However, this metric is inadequate when the database is imbalanced, since it can mislead the classification performance when the positive class is infrequent [224]. For these imbalanced databases, a more informative metric is the one which represents the precision and recall metrics [225], i.e. the F1_score. Drawing from the equation of the F1_score (Eq. 20), and the equations for the calculation of the precision (Pr) and recall or sensitivity (S) in (Eq. 21) and (Eq. 22), respectively, the F1_score can be defined in terms of the parameters of the confusion matrix as shown in (Eq. 23), i.e. TP, TN, FP and FN. It can be seen that the final equation coincides with the equation for the Dice index (Eq. 11), since the segmentation is similar to a binary classification.

$$F1\_score = \frac{2 \times Pr \times S}{Pr + S} \text{ (Eq. 20)} \quad Pr = \frac{VP}{VP + FP} \text{ (Eq. 21)} \quad S = \frac{VP}{VP + FN} \text{ (Eq. 22)}$$

$$F1\_score = \frac{2 \times \left(\frac{VP}{VP + FP} \right) \times \left(\frac{VP}{VP + FN} \right)}{\left(\frac{VP}{VP + FP} \right) + \left(\frac{VP}{VP + FN} \right)} = \frac{\frac{2 \times VP}{(VP + FP)(VP + FN)}}{\frac{VP \times (VP + FN) + VP(VP + FP)}{(VP + FP)(VP + FN)}}$$

$$F1\_score = \frac{2 \times VP^2}{VP \times (VP + FN + VP + FP)} = \frac{2 \times VP}{2 \times VP + FP + FN} \text{ (Eq. 23)}$$

As explained in the subsection 4.2.3. *Evaluation of segmentation models*, these confusion matrix parameters (TP, TN, FP and FN) are defined by the prediction of the classes for the test set and the real classes of the same samples. However, the definition is different for classification compared to the segmentation, since in this case the model separates the samples into different classes. For this reason, the confusion matrix parameters are calculated for each class to predict:

- TP: number of samples for which the predicted class and the actual class coincide with the evaluated class.
- TN: number of samples for which neither the predicted class nor the actual class matches the tested class.
- FP: number of samples for which the actual class coincides with the evaluated one, but no with the predicted one.

- FN: number of samples for which the predicted class coincides with the evaluated one, but no with the actual one.

Although, following the definition, the confusion matrix for two classes is similar to the one explained for the segmentation process (Figure 68a), the confusion matrix for more than two classes must be evaluated individually for each class. Figure 68b shows the example of a confusion matrix centred in the prediction of the class 1. As it can be seen, not every mistaken sample is considered a FP or FN, only if one of the labels, predicted or actual, is the evaluated class.

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

(a)

		Predicted		
		0	1	2
Actual	0	TN	FP	TN
	1	FN	TP	FN
	2	TN	FP	TN

(b)

Figure 68. Representation of the confusion matrix for (a) a binary classification; and (b) a classification of three classes, evaluating the class 1.

Based on both, the F1_score formula and the confusion matrix, the metrics used for evaluating the performance of the classification technique are based on the F1_score for each of the classes present in the corresponding clinical context. For the binary classification, first the calculation of F1_score for class 1 (F1_1) is calculated following (Eq. 23). Then, the F1_score for class 0 is equally calculated (F1_0), assuming as TP what in the original confusion matrix (Figure 68a) were TN, i.e. the samples for which the evaluated, actual and predicted classes coincide. For the classification of three classes, following Figure 68b, the three confusion matrixes are extracted, i.e. one for each class, and then following (Eq. 23), each of the F1_scores are calculated (F1_0, F1_1 and F1_2), changing the values of TP, TN, FP and FN depending on the evaluated class. Having the F1_scores for each of the classes, two metrics are used for assessing the performance:

- F1 score macro (F1_macro): average of F1_scores of each class. Its calculation is described by (Eq. 24), being c the last class to evaluate, i.e. 1 or 2 depending on the type of classification (binary or for three classes); F1_i, the F1_score for class i ; and N_c , the total number of classes, i.e. 2 or 3.

$$F1_{macro} = \frac{\sum_{i=0}^c F1_i}{N_c} \quad (\text{Eq. 24})$$

- F1 score weighted (F1 weighted): weighted sum of the F1_scores of each class, giving more weight to the most frequent class. Its calculation is described by (Eq. 25), being c and $F1_i$ the same as for (Eq. 24); N_i , the number of samples for the class i ; and NT , the total number of samples.

$$F1_{weighted} = \frac{\sum_{i=0}^c N_i \times F1_i}{NT} \quad (\text{Eq. 25})$$

Although both metrics, `F1_score_macro` and `F1_score_weighted` are standardised, the first is hugely affected by the lack of correct predictions of a class, even if the number of samples to predict is remarkably lower than for the rest of classes, while the second is practically not affected. For this reason, another metric, variation of the `F1_score`, is proposed for the classification of more than two classes, named as `F1_score_macro*` (`F1_macro*`). This metric has an intermediate affectation of the imbalanced dataset, since it is calculated as a mean of the `F1_score_macro` of each class assuming a binary problem for calculating the `F1_score` of each class.

First, the `F1_score` for the evaluated class is calculated (`F1_i`) and then the `F1_score` for everything which is not the evaluated class (`F1_noti`), as `F1_0` was calculated for the binary problem. Following the Figure 68b, it can be seen how the TN corresponds to the two classes which are not the evaluated one, i.e. 0 and 2. Then, the mean between `F1_i` and `F1_noti` is calculated as the `F1_score_macro` of the correspondent class (`F1_macro_i`). Finally, the `F1_macro*` is calculated, following (Eq. 26), as the mean of the correspondent `F1_macro_i`. This metric is not used for binary problems because is the same as `F1_score_macro`.

$$F1_{macro*} = \frac{\sum_{i=0}^c F1_{macro\_i}}{Nc} \quad (\text{Eq. 26})$$

5. Results

Following the structure of chapter 4. *Materials and methods*, this chapter presents the results obtained from the applied methodologies. First, the outcomes of the segmentation methodologies are detailed, beginning with the platform developed to address the need for manual mask revisions, followed by the segmentation based on traditional algorithms for both bone and BM. Continuing with the segmentation pipeline, the results of the DL approach for both bone and BM are also presented. Finally in the context of segmentation, a comparative analysis of both algorithms is provided as a discussion section. Then, the BM quantification methodology is divided into two different subsections: one focused on the clinical contexts of MM, and the other on NHL. Due to the limitations of the databases for certain clinical contexts, only the ML approach or, in one case, univariate analysis can be applied. The results of both methodologies, segmentation and quantification, are also compared with the studies previously discussed in chapter 3. **State of art**, highlighting the contributions of the present thesis.

5.1. Bone marrow segmentation

As mentioned in section 4.2. *Bone marrow segmentation*, the initial step involves obtaining a bone segmentation to subsequently use it as a means for deriving the BM segmentation. In addition, as shown in Figure 19, a precise segmentation of both bone and BM is crucial to develop a model capable of generalizing this task across multiple patients. For this reason, prior to implementing the DL model, a semi-automatic segmentation approach composed of two parts, an automatic segmentation based on traditional algorithms followed by a manual revision through a platform specifically developed for the present thesis, are performed to generate the GTs required for training the DL model.

Although the initial segmentation approach relies solely on visual evaluation of the results, since it is intended as a preliminary step for manual revision, for comparing both segmentation approaches, the same test database is used, ensuring the comparison is as realistic as possible. Drawing from the images described in section 4.1. *Materials*, the CT scans are converted into volumes using SPM12, forming the cohort of the present database.

Since a GT is needed to compare the masks resulting from both segmentation algorithms, the masks manually revised for testing the DL model are used as GT. In this way, a total of 77 CT scans from both acquisition protocols (45 whole-body and 32 femur-to-head) are included in the database used for the DL model. Although the model is designed for 2D segmentation, to avoid overfitting, all slices from the same

patient are assigned to the same set, either the training set or the test set. For bone segmentation, the training set includes 65 out of these 77 patients, comprising a total of 21,304 slices. Meanwhile, the remaining 12, with 3,813 slices, constitute the test set.

Regarding BM segmentation, as illustrated in Figure 41 and Figure 45, there are two distinct training and testing phases. For the first phase, the train set is composed of pGT masks, as detailed in subsection 4.2.2.2.1. *Pseudo-ground truth creation*, from 64 out of the 65 CT scans used for the bone training set, while the test set consists of 13 manually revised/segmented masks. In the second phase, where both training and test sets consist of manually revised masks, six scans are included in the training set, and the remaining seven in the test set. A summary of the number of patients, samples, and the origin of the masks for each training and test set is presented in Table 10.

	Train			Test			Total	
	N	S	Origin	N	S	Origin	N	S
Bone segmentation	65	21304	GT	12	3813	GT	77	25117
BM segmentation I	64	21110	pGT	13	4007	GT		
BM segmentation II	6	1749	GT	7	2258	GT	13	4007

Table 10. Division of train and test sets in each training step for both, bone and BM segmentations. N: number of patients, i.e. CT volumes; S: number of samples, i.e. slices of the volumes.

From the Table 10, the test column serves as the reference for quantitatively comparing both segmentation algorithms, with the semi-automatic approach being the initial step prior to manual revision. Since GT is based on the platform developed specifically for the present thesis, it is explained first.

5.1.1. Platform for bone and bone marrow segmentation

As mentioned in subsection 4.2.1.5. *Manual revision*, the masks resulting from the automatic segmentation based on traditional algorithms are manually revised using a platform developed for the present thesis, called ‘Bone Marrow Segmentation App’. Once the DL segmentation models are performed and evaluated, they are integrated into the platform, replacing the traditional algorithm. This way, the platform structure remains practically the same, differing only in certain functions which are explained in detail. The figures provided to illustrate the platform correspond to the version with the DL models, as they reflect the most recent update.

5.1.1.1. Automatic segmentation tabs

To support the full semi-automatic segmentation process, the platform includes three distinct tabs, two for automatic segmentation and one for subsequent manual editing. The difference between the tabs for automatic segmentation depends on the type of CT file available: ‘Pre-processing’, in case of starting from DICOMs as initial files, and ‘Segmentation’, in case of starting from the .mat file of the CT. Then, the ‘Editing’ tab for the manual revision of the masks is provided.

5.1.1.1.1 Pre-processing tab

The first tab, shown in Figure 69, allows the user to perform an automatic segmentation of both bone and BM. Initially, the user should click the ‘Select Subjects Folder’ button, which, at the beginning, is the only clickable option apart from the information button. Upon clicking, a folders explorer is opened, and the user should select the folder which contains a subfolder for each patient. Each of the patient folders, in turn, includes two folders, one for the CT DICOMs and another for the PET DICOMs. Then, the clinical subcontext should be selected from the available options: MRD (MRD detection), MM (MRD prediction), SMM (risk of progression to MM) and NHL (BM status). Once both are chosen, the ‘Select/Create Results Folder’ becomes enabled, and when clicked, the folder explorer appears again, allowing the user to select an empty folder for saving the results. A different subfolder for each of the clinical subcontexts is created within the selected results folder. Finally, before starting the segmentation process by clicking the ‘Segment images’ button, the user can adjust two options via checkboxes, located under the mentioned button: one to re-segment subjects which have been already segmented, and another to skip the manual selection of the first and last slices to take into consideration. This slices selection step is only applied when the DL model is used to segment, since the traditional algorithm is pre-configured to exclude a percentage of slices automatically to remove the head and the slices below the knees based on the total number of slices.

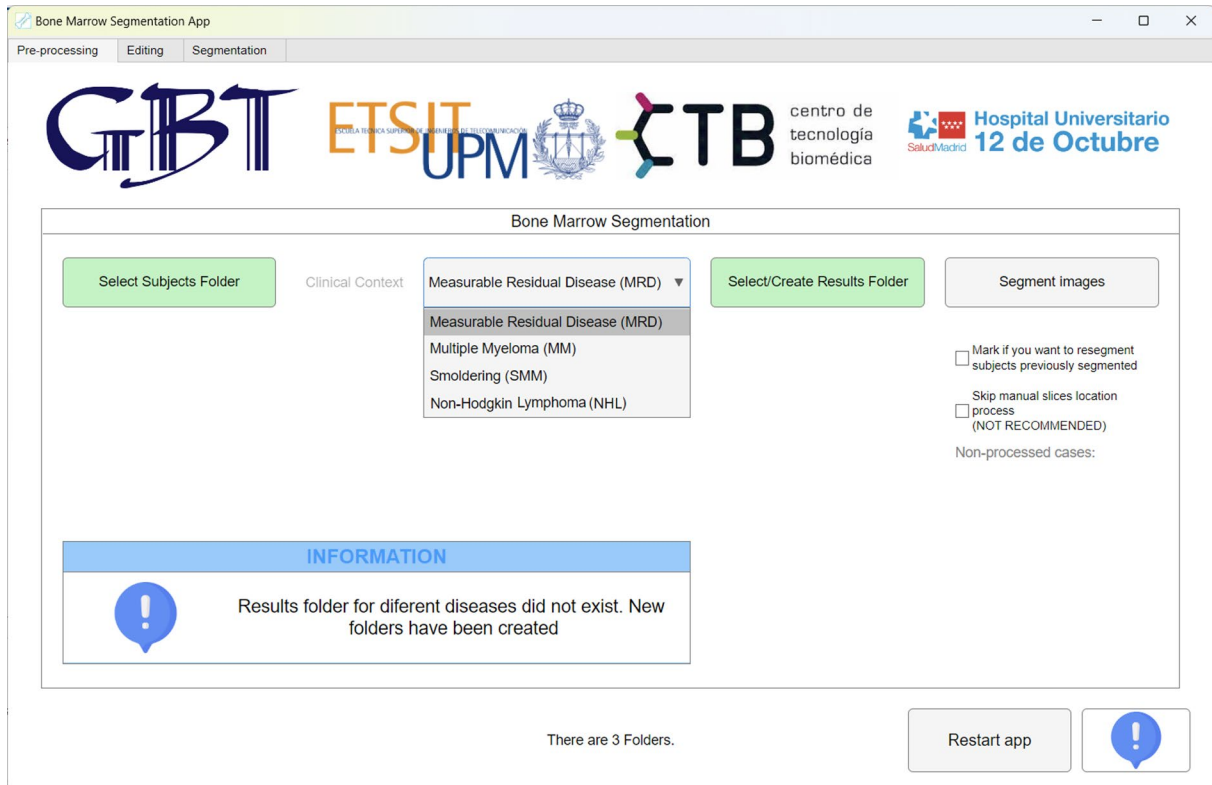


Figure 69. ‘Pre-processing’ tab of the semi-automatic segmentation platform: initiation.

As it can be seen, the tab includes additional elements: an information panel that displays the current step of the process, changing the colour of the message and the icon as illustrated in *Annex B. Figure 107*; a ‘Restart app’ button, which allows the user to restart the application after the segmentation process is complete; and an information button, which opens the instructions specific to this tab. These instructions are included in *Annex B. Instructions of the automatic segmentation tab*.

Once the ‘Segment images’ button is clicked, the platform disables all buttons, and the segmentation process begins with the creation of CT and PET volumes from the DICOM files, being the final step the obtention of the SUVmap from the PET volume (4.3.2. *PET processing*). However, the steps between both the first and final steps, are different depending on the segmentation method. For the segmentation based on the traditional algorithms, the process only concludes when all steps are completed, following the sequence described in subsection 4.2.1. *Semi-automatic segmentation based on traditional algorithms*: stretcher removal, bone segmentation, spinal canal subtraction and compact bone removal for BM segmentation. In the case of the DL segmentation model, the steps are different (4.2.2.1.4. *Pre-processing steps*): stretcher removal, CT pre-processing through a thresholding step, image clipping using the slice selection panel, bone segmentation, and BM segmentation. During this process, as shown in Figure 70, the ‘Segmentation progress’ panel provides information about the step which is being developed in that moment (e.g.,

volume of CT creation, stretcher removal) and displays the percentage of patients processed from the total in the selected folder.

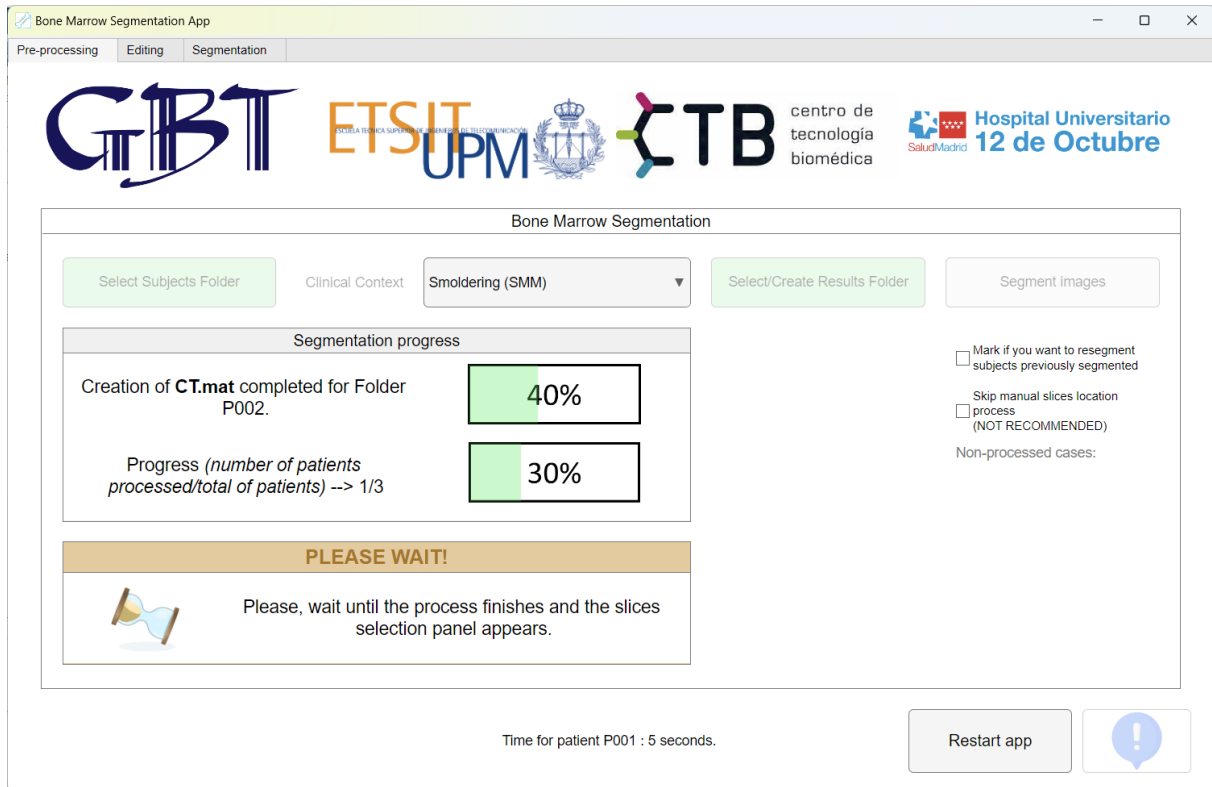


Figure 70. ‘Pre-processing’ tab of the semi-automatic segmentation platform: CT processing.

As mentioned before, an algorithm for slice selection is included for the DL segmentation model. Previous to the manual selection, there is an automatic function for slices selection, which can be used as the definitive if the second checkbox, the last one located under the ‘Segment images’ button, is marked by the user. For this automatic selection, first the volumetric CT is binarised using 98 HU as threshold, then the stretcher is removed from this binarised volume. The first and last slices with mask are located and named z_0 for the lower limit, i.e. close to the femurs or feet as shown in Figure 71a; and z_F for the upper limit, i.e. near the head as shown in Figure 71b.

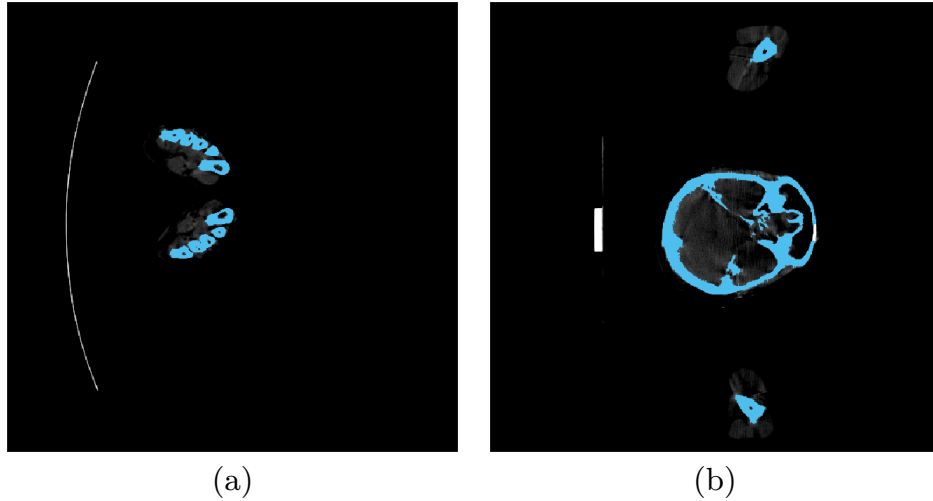


Figure 71. Slices location I: first approximation for (a) the lower slice, i.e. close to the femurs, and (b) the upper slice, i.e. close to the head, showing the binarised CT (blue) over the thresholded CT scan between 0 and 300 HU for a better visualization.

Then, the binarised volume is divided into isolated subvolumes, selecting the largest one, i.e. the one with the highest number of voxels. From the largest volume, each slice is evaluated starting from zF and moving downwards. This way, the last slice which includes the chin is tried to be located. To achieve this, the area including the entire mask must (1) be at least 10 pixels in size in the column-to-chin direction and (2) be smaller than 4000 pixels, avoiding head slices selection. The first slice which does not meet these requirements, moving from head to feet, is the one selected as the final slice of the volume, i.e. the last before the head. In Figure 72ac are two examples of the last slice which fulfils both requirements, and in Figure 72bd are the correspondent two first slices which do not, i.e. the ones selected as the final slice of the volume.

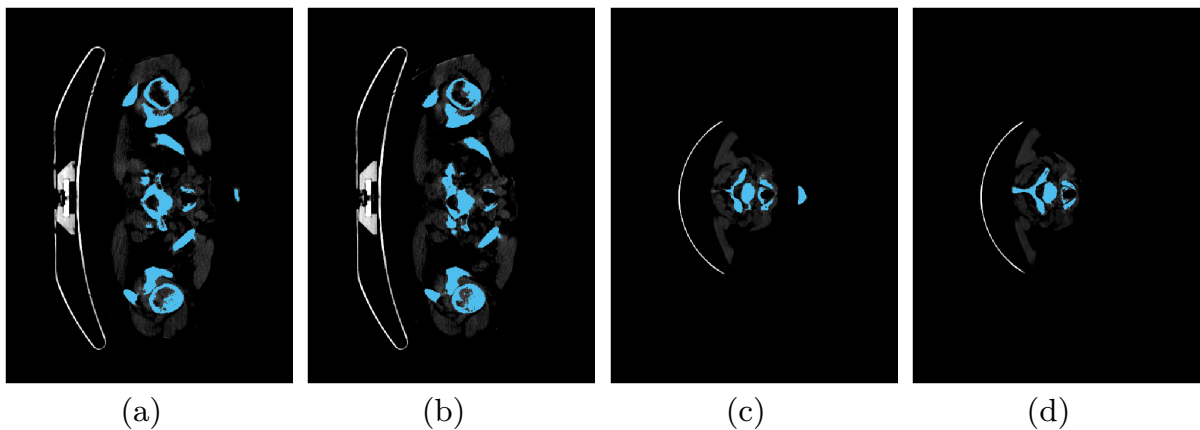


Figure 72. Slices location II: last slice location. Two examples of (a and c) location of the last slice including the chin, and (b and d) slice directly below to (a) and (c) respectively, showing the binarised CT (blue) over the thresholded CT scan between 0 and 300 HU for a better visualization.

For the first slice location, the main objective is to find the slice closest to the knees, directly above. With this aim, moving from the slice selected as the last one (Figure 72bd) to the ones which represent the lower part of the body, the slice with the largest area including the entire mask is located and considered as the pelvis one (Figure 73a). Then, considering that the whole-body acquisition protocol requires more than 70 slices from the pelvis to the feet due to the slice thickness, 70 is subtracted from the pelvis slice index, evaluating two possibilities: (1) the subtraction results in a number equal to or lower than 0, in which case the slice 1 of the total volume is defined as the first one to be selected (Figure 73b); and (2) the subtraction is higher than 0. In this last case, if the pelvis is correctly located, the resulted index corresponds to the femurs. Then, the next slice approaching the feet that has an area considerably larger than the slice directly above it is considered to be the one corresponding to the knees, selecting the directly above one as the first slice of the volume (Figure 73c).

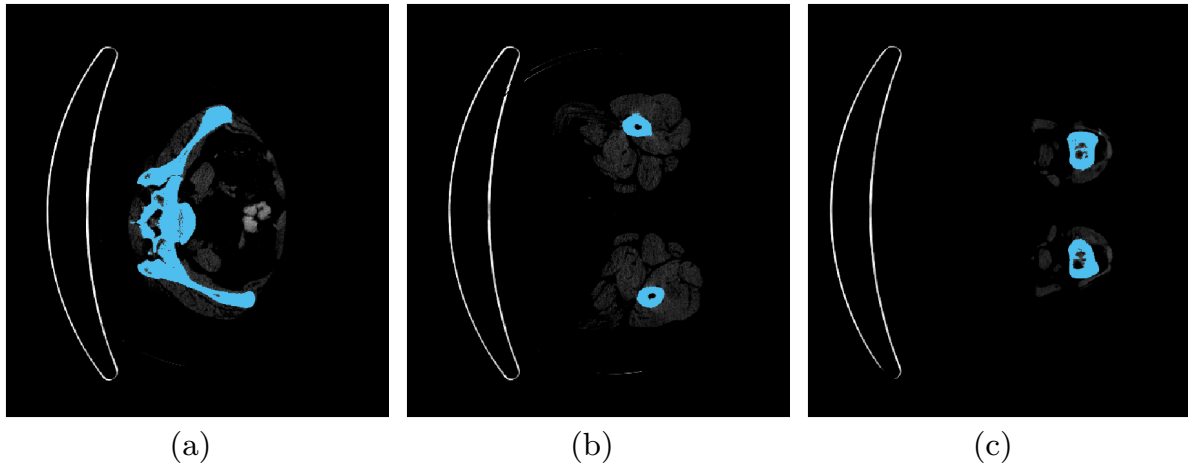


Figure 73. Slices location III: first slice location. Examples of (a) pelvis location as the largest area, (b) first slice for CT starting from femur and (c) first slice for CT of whole-body protocol, showing the binarised CT (blue) over the thresholded CT scan between 0 and 300 HU for a better visualization.

Although this slices location process works for the majority of cases, some of them, due to the binarization and anomalous uptake in soft tissues, have problems with pelvis location (Figure 74a) implying a wrongly first slice location (Figure 74b). For this reason, the slice location panel is included in the platform, in order to segment the ROI of the CT correctly.

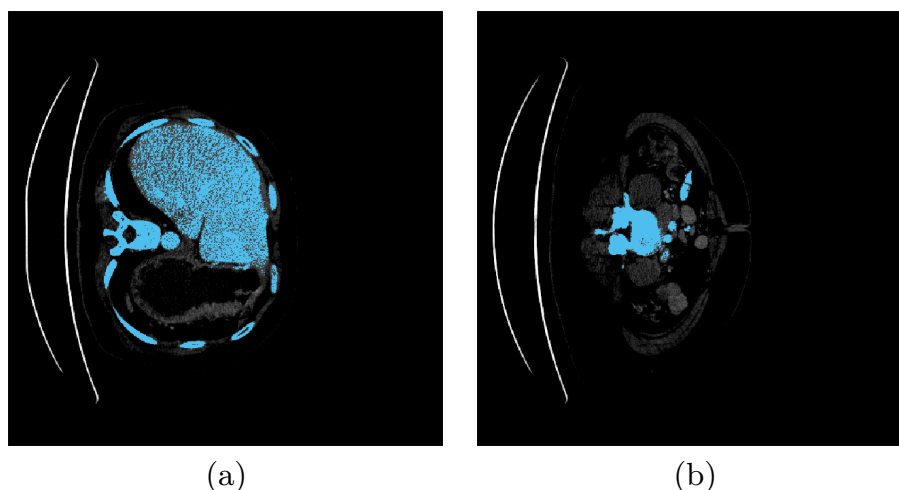


Figure 74. Slices location IV: pelvis slice wrongly located. Examples of (a) pelvis location as the greatest area mistaken by a high HU in some internal organs, (b) slice defined as first slice due to the mistake in (a), showing the binarised CT (blue) over the CT thresholded between 0 and 300 HU for a better visualization.

In Figure 75, the ‘Slice selection’ panel is shown, where the user can select the first and last slice of the CT between which the bone and BM are segmented. To select these slices correctly, the platform displays the CT slice corresponding to the slice index, allowing the user to evaluate which slice is the first one containing the femur and the last one that does not include the head. Although this process is simple, fast and allows a more precise segmentation, it could be skipped for automatic segmentation.

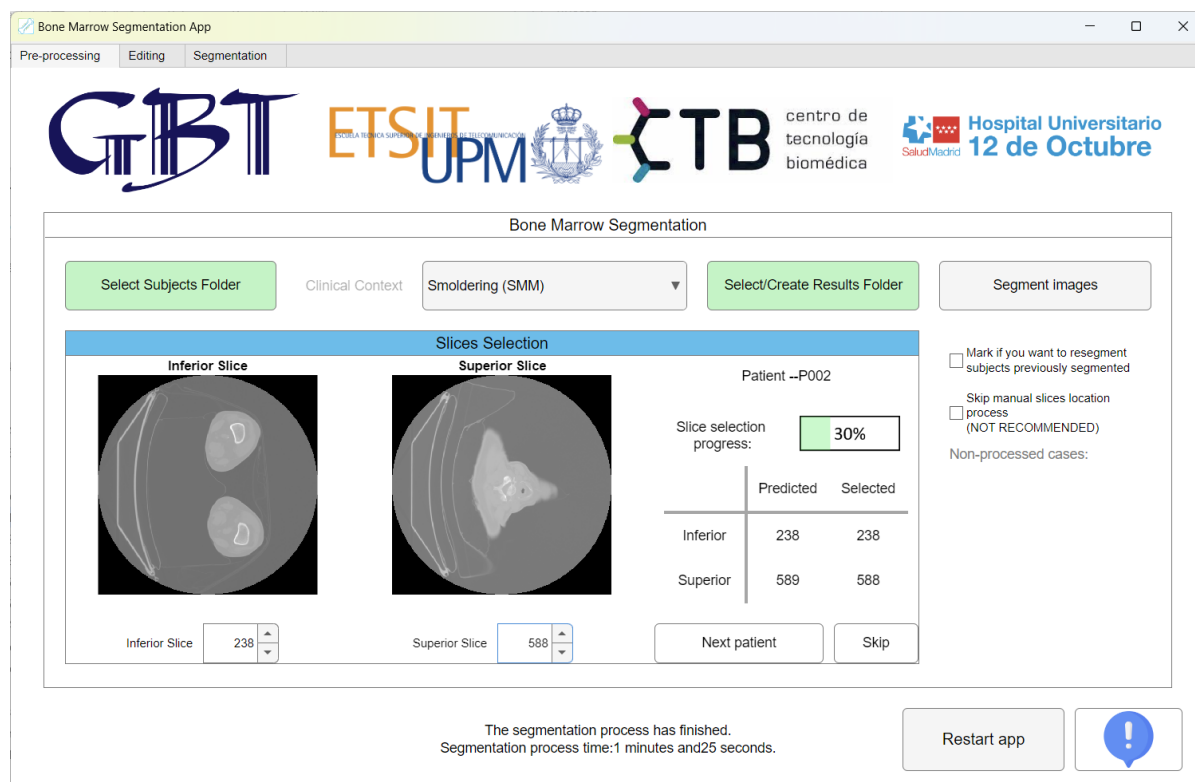


Figure 75. ‘Pre-processing’ tab of the semi-automatic segmentation platform: slices selection.

5.1.1.1.2 Segmentation tab

In case only CT volumes are available, nor the original images or PET images, the ‘Segmentation’ tab allows to obtain the same bone and BM segmentations (Figure 76). In this case, the folder selected by clicking ‘Select Patients Folder’ must contain only a subfolder for each of the patients with their CT volume in .mat format, which is one of the results of the ‘Pre-processing’ pipeline. The internal process is exactly the same as the ‘Pre-processing’ one described, without PET or CT obtention. However, it must be taken into account that the bone and BM masks are saved in the same folder where the CT is found, in order to have the same results structure than the one obtained for the ‘Pre-processing’ tab. The information button allows to open the instructions, similar to those of the ‘Pre-processing’ tab, included in the *Annex B. Instructions of the automatic segmentation tab.*

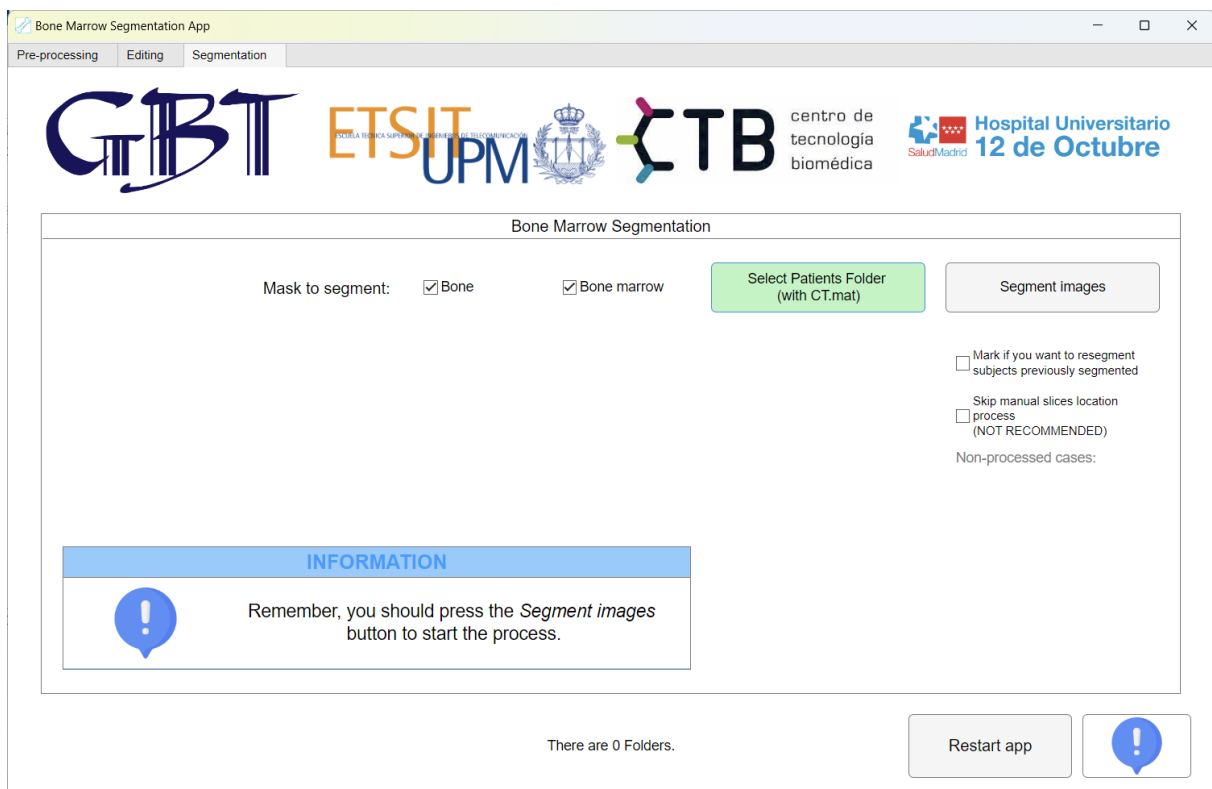


Figure 76. ‘Segmentation’ tab of the semi-automatic segmentation platform: initiation.

5.1.1.2. Editing tab

Once the automatic segmentation is performed, the user can move to the ‘Editing’ tab. This tab is composed by four different panels, as shown in Figure 77: ‘Files functions’ for selecting and loading the case; ‘Edition functions’ for modifying the segmentation; ‘Visualization functions’ which allows the user to get a better visualization of the mask and the complete volume; and ‘Extra functions’ which includes a variety of functions detailed below. Figure 77 shows the initial interface

of the platform once the case is loaded. Previous to case loading, only the ‘Bone’, ‘Load case’ and information buttons are enabled.

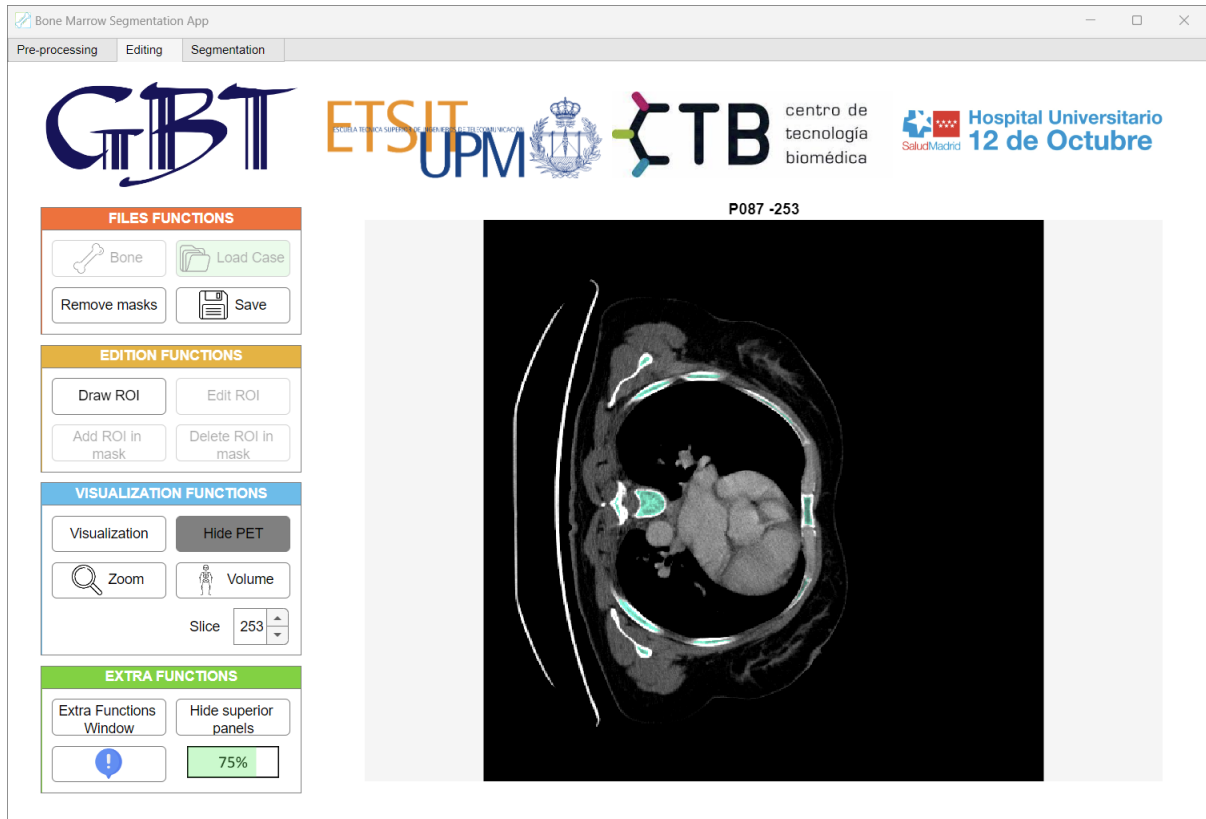


Figure 77. ‘Editing’ tab of the semi-automatic segmentation platform: initiation.

In addition to the panels, there are functions applied when certain keys are pressed. Some functions are applied both by clicking one button and pressing one key to make the platform as useful as possible, since segmenting can be faster using the keys than pressing the buttons. However, there are some functions exclusive for keyboard, avoiding including more buttons in the main window in order to maximise the display of what is desired to be segmented.

5.1.1.2.1 Files functions panel

This panel is formed by four buttons essentials for the correct loading and saving of the masks. First, the user should click the ‘Bone’ button if, and only if, the edition of the mask is for the bone segmentation, since some of the more complex editing functions are adapted to bone or BM depending on this selection. Then, clicking the ‘Load case’ button, a folder explorer is opened, and the user should select mask to edit, which should be saved in the patient subfolder from the folder selected in the automatic process when ‘Select/Create Results Folder’ button is clicked. The platform loads the CT, the SUVmap and the selected mask, nonetheless only CT thresholded and the overlayed mask are shown, since only the anatomical information is useful for the segmentation. If any of these volumes are not available,

an emergent window showing a warning message appears, indicating the impossibility of loading the case.

The other two buttons are important for finishing the segmentation. The ‘Save’ button allows the user to save the edited mask, same as the ‘Control’ key. When clicked and saved, the platform shows the emergent window shown in Figure 78. This window displays the pixels which have been changed with regard to the original mask or the new mask. In addition, the total time of edition and the corresponding last saving date and time are included in this emergent window.

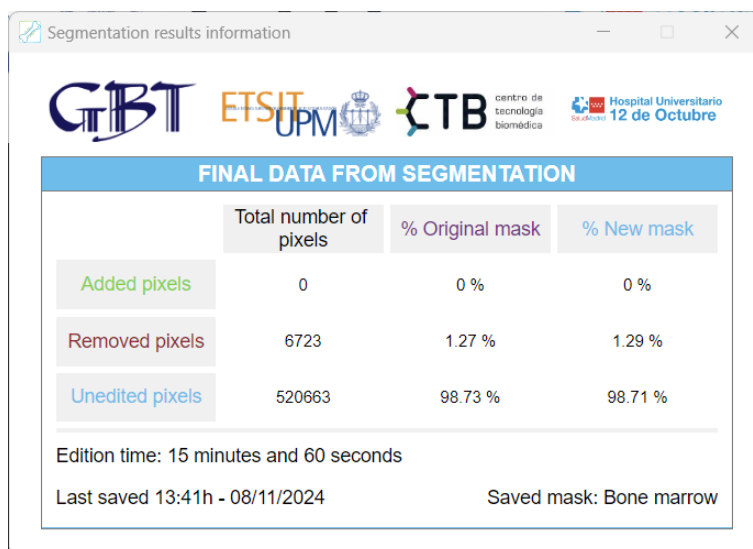


Figure 78. ‘Editing’ tab of the semi-automatic segmentation platform: saving information.

The ‘Remove masks’ button is used for removing all the masks, allowing to load another new case. When clicked, the emergent window shown in Figure 79 appears, assuring the saving of the masks before removing them, since there is no possibility of saving after the removal.

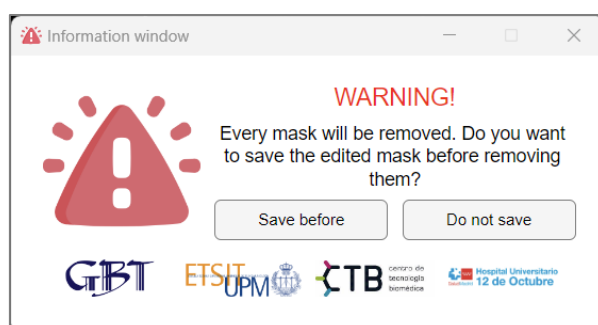


Figure 79. ‘Editing’ tab of the semi-automatic segmentation platform: Removing masks emergent window.

5.1.1.2.2 Visualization functions panel

If the user wants to see also the SUVmap when the case is loaded, the ‘Hide PET’ button should be clicked (Figure 80). Although this option could be useful for removing parts of the mask with a high SUV due to liver or any artefact, instead of

the real functional activity of the BM, it is not recommended to obtain a precise bone or BM segmentation, since the overlaid SUVmap can obstruct a correct visualization of the anatomical information of the CT.

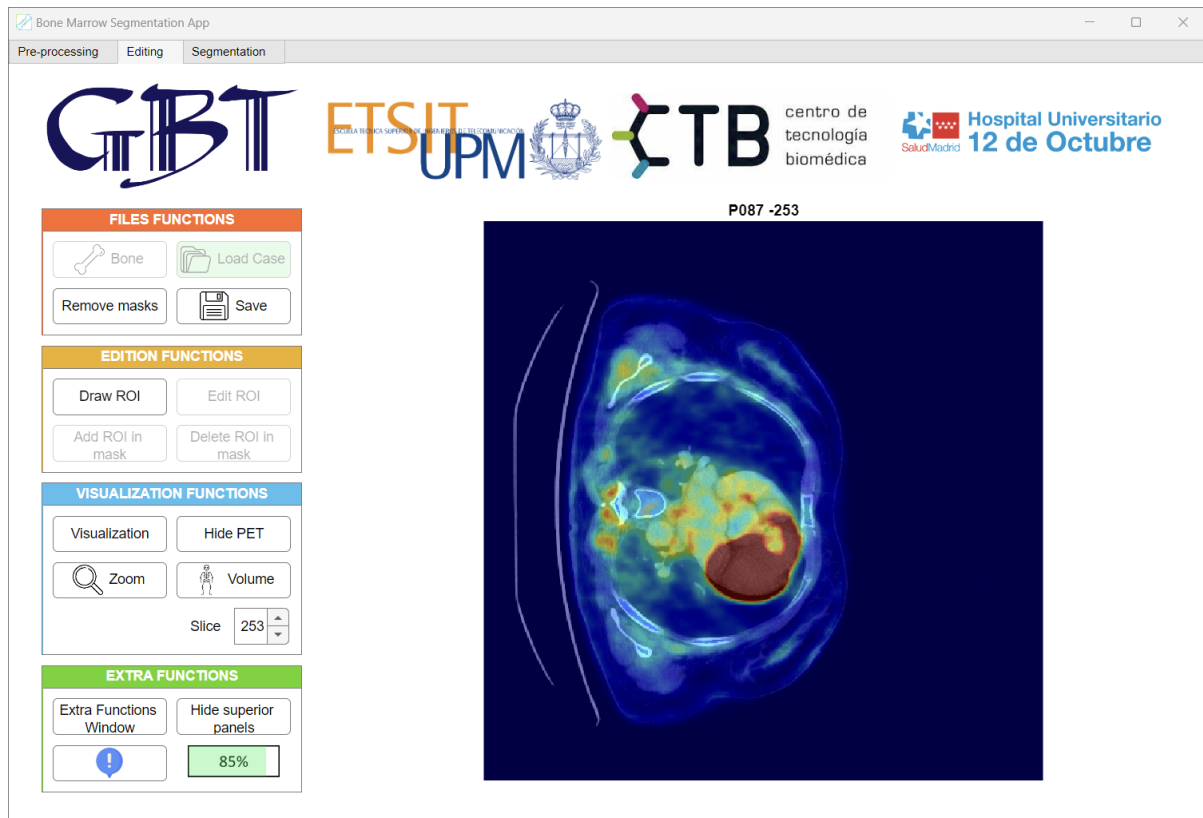


Figure 80. ‘Editing’ tab of the semi-automatic segmentation platform: including SUVmap.

For this reason, the visualization settings of every layer, i.e. CT, mask to edit and SUVmap, can be changed with the ‘Visualization’ button, trying to enhance the visualization and consequently the precision in the segmentation. Figure 81 shows the window which is opened when the ‘Visualization’ button is clicked. As it can be seen, the maximum value of the display range of the SUV can be changed, as the colour of the mask drawing from the basic colours in buttons and with a change of the brightness to make them clearer and opaque, or obscurer and transparent. It must be mentioned, that for editing the mask with the more complex functions, the created ROI is drawn in a contrasting colour: if the mask is green, the ROI is pink; if the mask is purple, the ROI is orange, etc. In addition, the mask could be completely binary or transparent with the correspondent buttons ‘Binary mask’ or ‘Hide mask’. In case of not desiring any of them, the user must have both buttons unclicked. Finally, the display range of the CT can be change, changing the minimum and maximum, and there are two ‘Presets’ buttons depending on the ROI, trying to enhance its visualization. There are presets for two anatomical structures: BM or soft tissues.

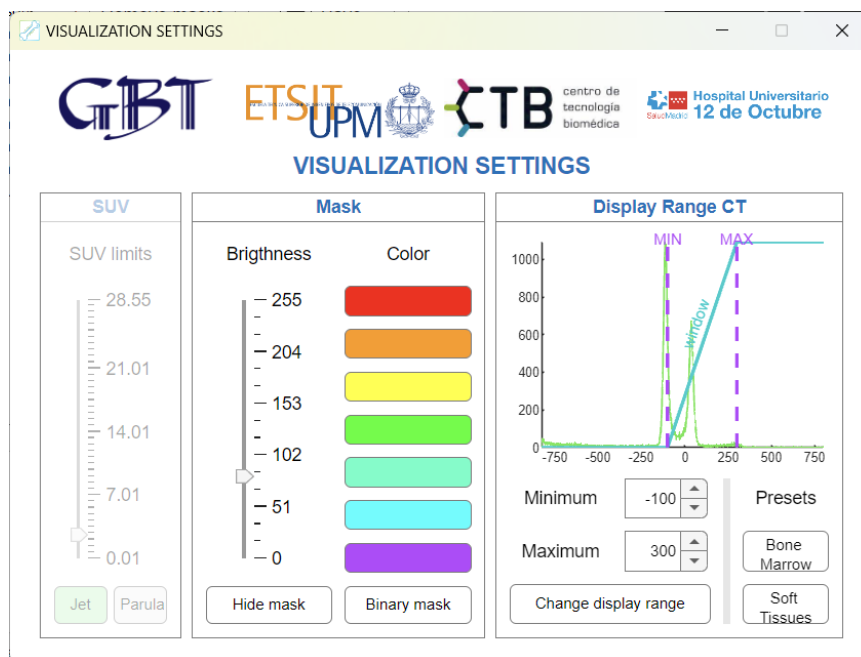


Figure 81. 'Editing' tab of the semi-automatic segmentation platform: 'Visualization settings' window.

The 'Zoom' button opens the emergent window in Figure 82. It allows the user to select the region which want to be zoomed, introducing manually the horizontal and vertical indexes. An illustration to explain the correct positions of the zoom is included in the window. Once the limits are changed, the user must click the 'Apply' button to get the zoom. Finally, there are a 'Reset zoom' for coming back to the original limits. Although it is a useful window, there are other ways to change the zoom easier to be interpreted, or at least faster, during the segmentation which are explained posteriorly.

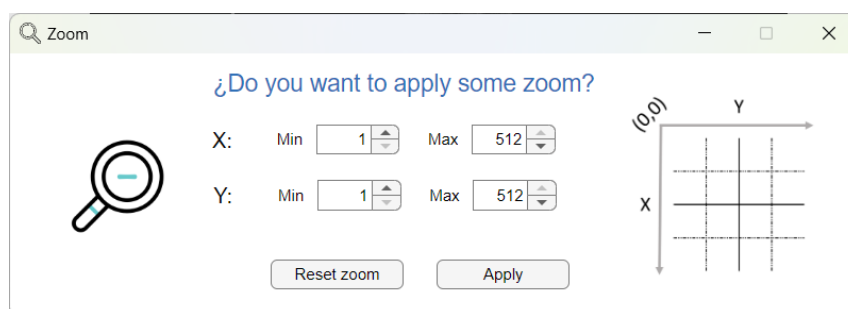


Figure 82. 'Editing' tab of the semi-automatic segmentation platform: 'Zoom' emergent window.

The 'Volume' button allows users to visualise the complete volume as shown in Figure 83. This way, the user can appreciate if there are regions not segmented or small structures to segment. It can be manually moved by the user using the mouse. In addition, the SUVmap in the volume regions can be visualised clicking the 'Show PET' button (*Annex C*. Figure 109).

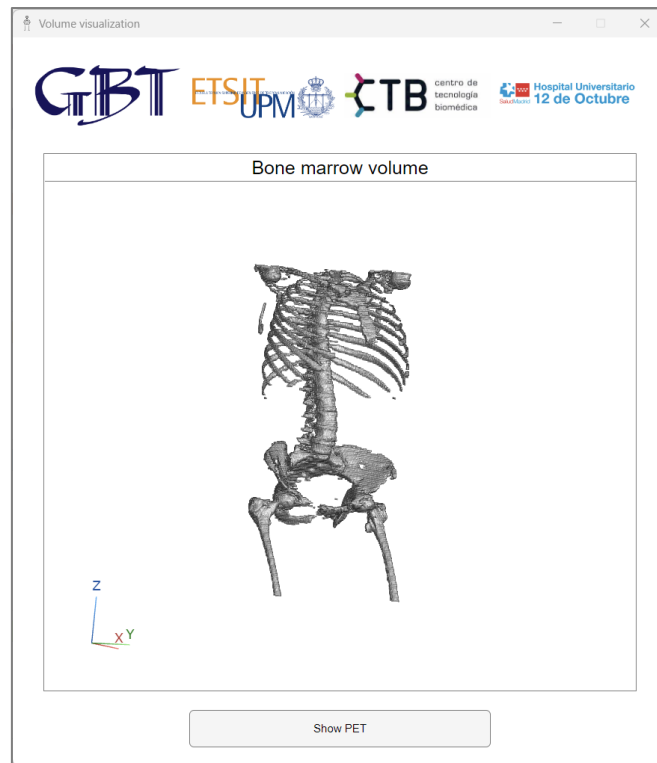


Figure 83. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Volume’ emergent window.

Finally, the slice to be visualised in the main window can be changed manually with the arrows next to ‘Slice’. This same function can be applied with the key ‘T’ for showing the directly above slice, i.e. closer to the head, or ‘G’ for the directly below, i.e. closer to the knees or feet. In case the slice selected is not included in the correct range, i.e. between 1 and the number of slices for both CT and PET, the user is warned with an emergent window (*Annex C*. Figure 110).

5.1.1.2.3 Edition functions panel

This panel includes the most basic functions for the mask edition. In Figure 84, an example of this function is included, with a zoomed version of the CT in order to help the explanation. Clicking the ‘Draw ROI’ button or pressing the ‘W’ key, the user can draw over the CT with the mouse, with any restriction except for the horizontal and vertical limits. The result of this drawing is the blue ROI. Then, clicking the ‘Edit ROI’ button or pressing the ‘E’ key, a circular auxiliary structure appears. This structure appears in red if the majority of the circular area is outside the ROI, and removes every pixel coincident with the ROI, and in green if the majority of the circular area is inside the ROI, adding any pixel which is not included in the ROI. When the user assumes the ROI as the final one, it could be added to the mask with the ‘Add ROI in mask’ button or the ‘A’ key, or it could be removed from the mask with the ‘Delete ROI in mask’ button or the ‘D’ key. These last two

functions are used for adding or removing, respectively, the ROIs generated with any editing functions, not only the basic ones.

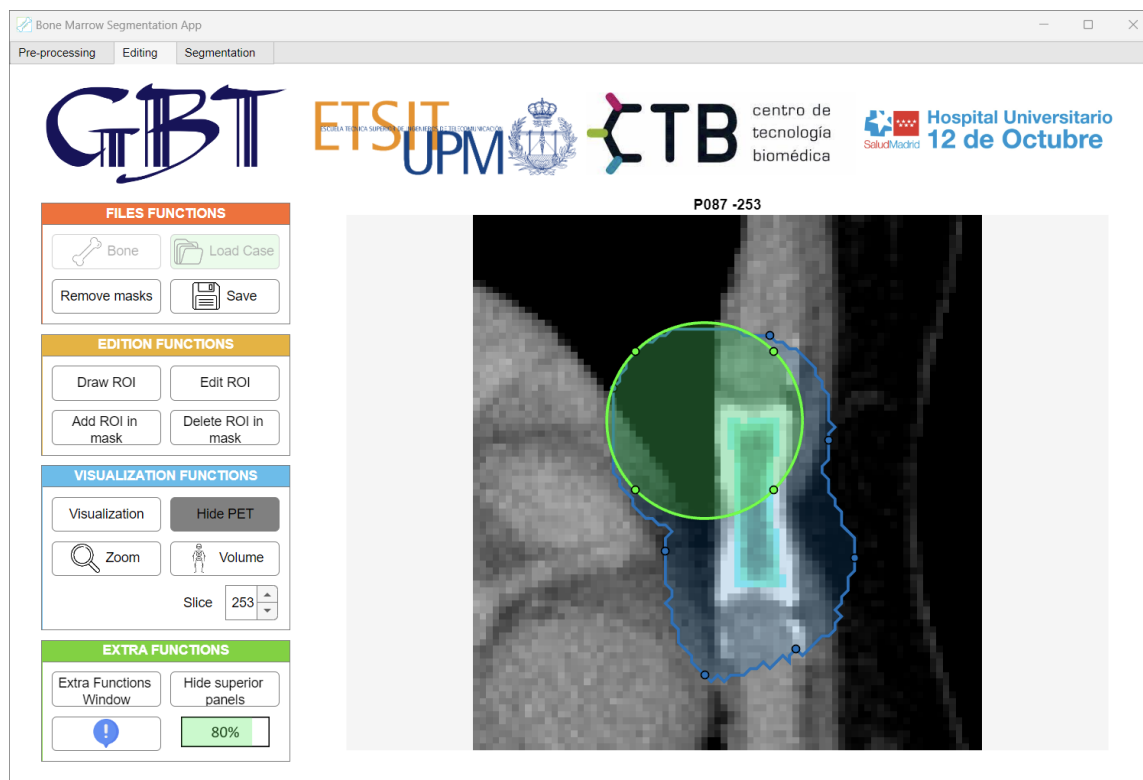


Figure 84. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Draw ROI’ and ‘Edit ROI’ functions.

5.1.1.2.4 Extra functions panel

The last panel is the ‘Extra functions’ panel, which includes a variety of functions not included in the previous panels. The first button, the ‘Extra Functions Window’ opens the window shown in Figure 85. These functions are divided into two subgroups: functions related to visualization, and functions related to edition.

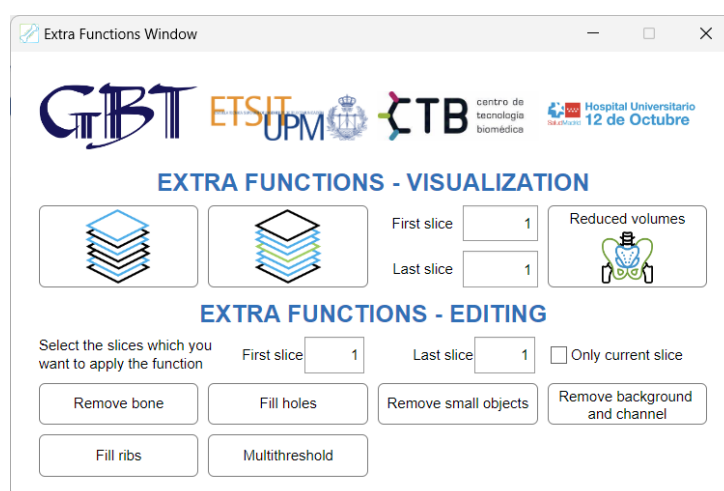


Figure 85. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Extra Functions Window’.

The first button of the visualization functions is represented with squares, where each of them illustrates a slice of the volume. Some of them are marked in blue, since this function opens a window (*Annex C. Figure 111*) with 8 equidistant slices from the first to the last slices with mask. The second button opens the same window, but in this case, the window shows the four directly above and below of the current slice, i.e. the one shown in the main window. For this reason, the squares marked in blue are the ones closer to the green one. The last button opens the ‘Volume’ window showing the volume between the marked slices with the zoom of the main window (Figure 86), instead of the complete volume. This way, the user can visualise a concrete region with more detail.

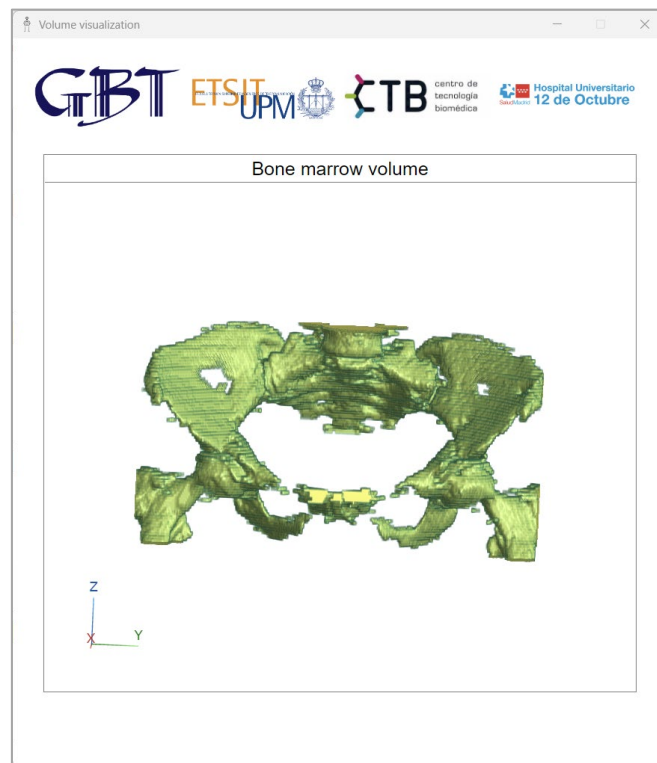


Figure 86. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Volume’ window for a concrete region selected in the ‘Extra Functions Window’.

All the editing functions from this ‘Extra Functions Window’ can be applied only in the current slice, i.e. the one shown in the main window, or in a selected range (Figure 82). In addition, everything is applied in the selected slices but only in the zoom or horizontal and vertical limits of the main window. The editing functions are:

- Remove bone: removes any pixel of the mask with a grey intensity level superior or equal to 300 HU.
- Fill holes: completes the small holes of the mask, also applied in the current slice by clicking ‘H’ and changing the radius of the holes to be filled with ‘9’ and ‘0’ for reducing and increasing it, respectively.

- Remove small objects: removes any structure smaller than 10 pixels in the slice.
- Remove background and channel: removes the low-grey-level parts of the mask.
- Fill ribs: finds ribs in the selected or drawn volume or area and includes them in the segmented mask in case of bone selection. In case of BM selection, only the inside of ribs is included. This function is also applied in the current slice with the 'R' key.
- Multithreshold: this button or pressing the 'M' key, opens the window shown in Figure 87. It allows to apply a concrete threshold over multiple slices, selecting: the range of slices to be applied; if a post-processing function to fill holes must be applied; inversion of the threshold, i.e. maintaining the pixels which are below the threshold instead of over the threshold; and the dilation of the mask, using a structural element of disk shape with the selected radius. Once everything is chosen, the user must click the 'Preview' button, and the ROI is drawn in a colour with the maximum contrast to the mask.

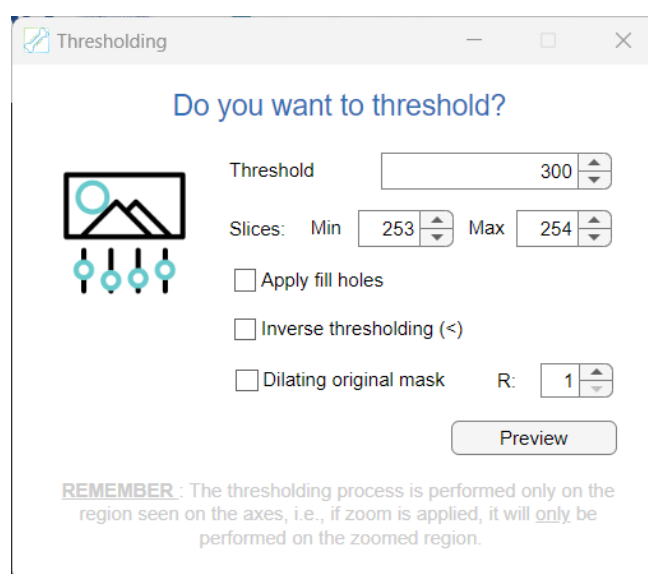


Figure 87. 'Editing' tab of the semi-automatic segmentation platform: 'Thresholding' window.

The next button of the panel, i.e. 'Hide superior panels', changes the superior panels of the main window for a general visualization of the current slice, marking in yellow the zoom shown in the main axes (Figure 88). This way, the user has a better interpretation of what is shown in the context of the whole slice.

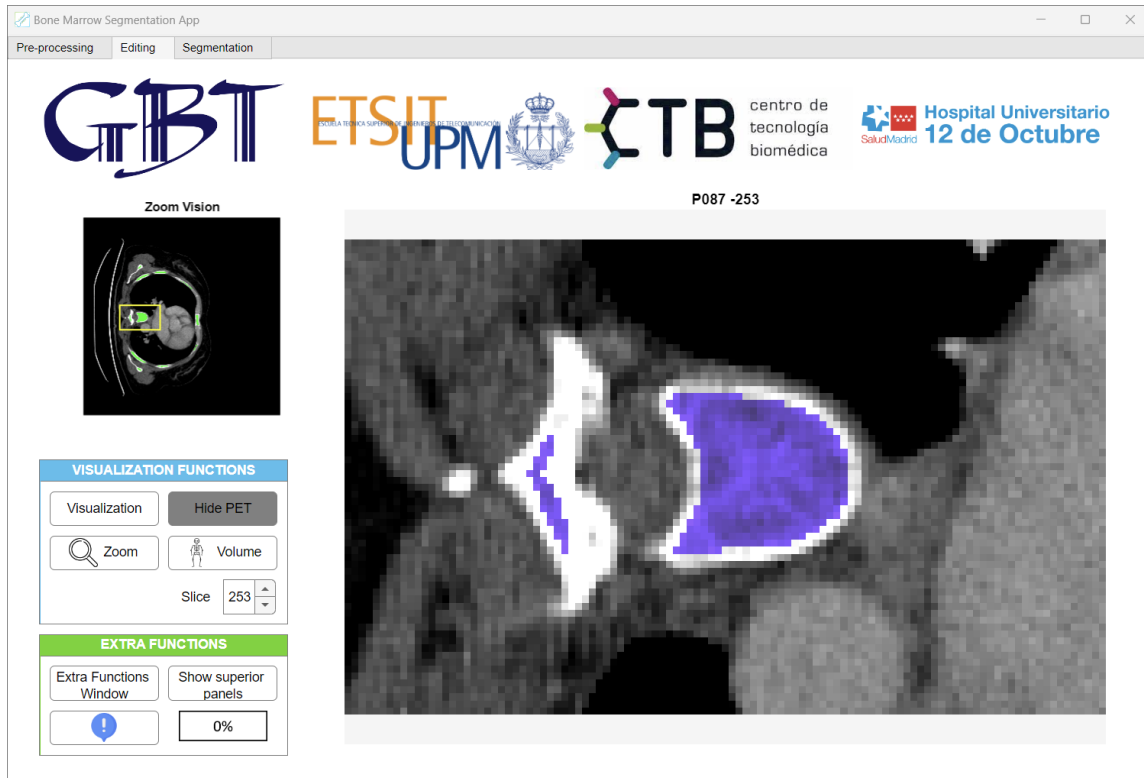


Figure 88. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Hide superior panels’ button.

The last row of this panel includes an information button and the percentage of slices of the CT revised, since the platform shows the slice number 1 at the beginning. Clicking the information button, the window shown in Figure 89 is opened. This window is composed by four buttons, each of them opens a window with the instructions for using the functions included in: the buttons of the interface panels; concrete instructions for the buttons of the panels which open other windows, i.e. the ‘Visualization’ and ‘Extra Functions Window’ buttons; and functions related to the keyboard. The instructions mentioned can be consulted in the *Annex D. Instructions of the Editing tab*.

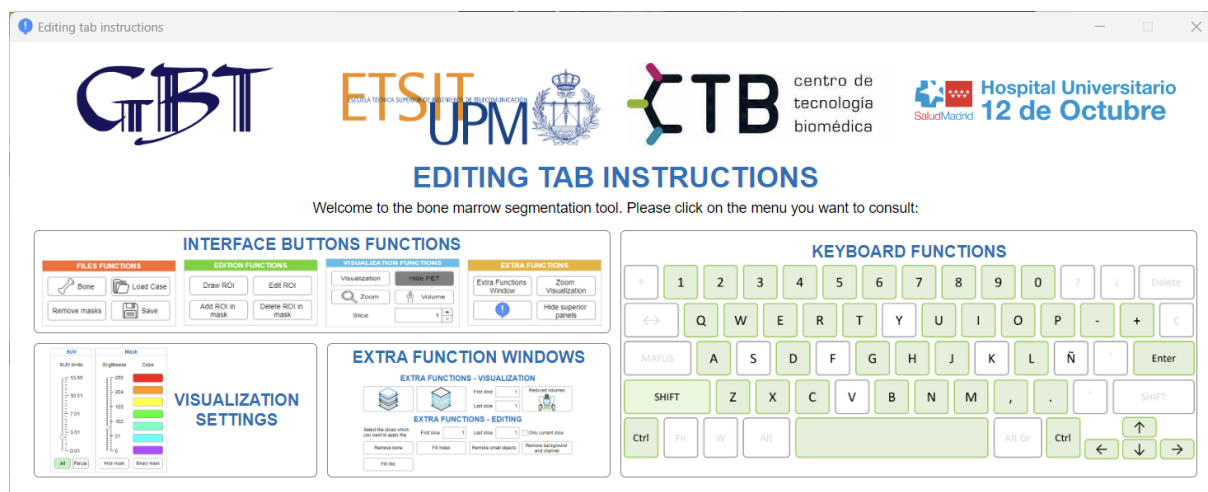


Figure 89. 'Editing' tab of the semi-automatic segmentation platform: 'Editing tab instructions'.

5.1.1.2.5 Keyboard functions

For the keyboard functions, their instructions are shown in Figure 90, since the explanation is easy to follow. The keyboard functions are divided into four different groups using different colours: visualization functions (blue), basic edition functions (yellow), information functions (red) and advanced edition functions (green).

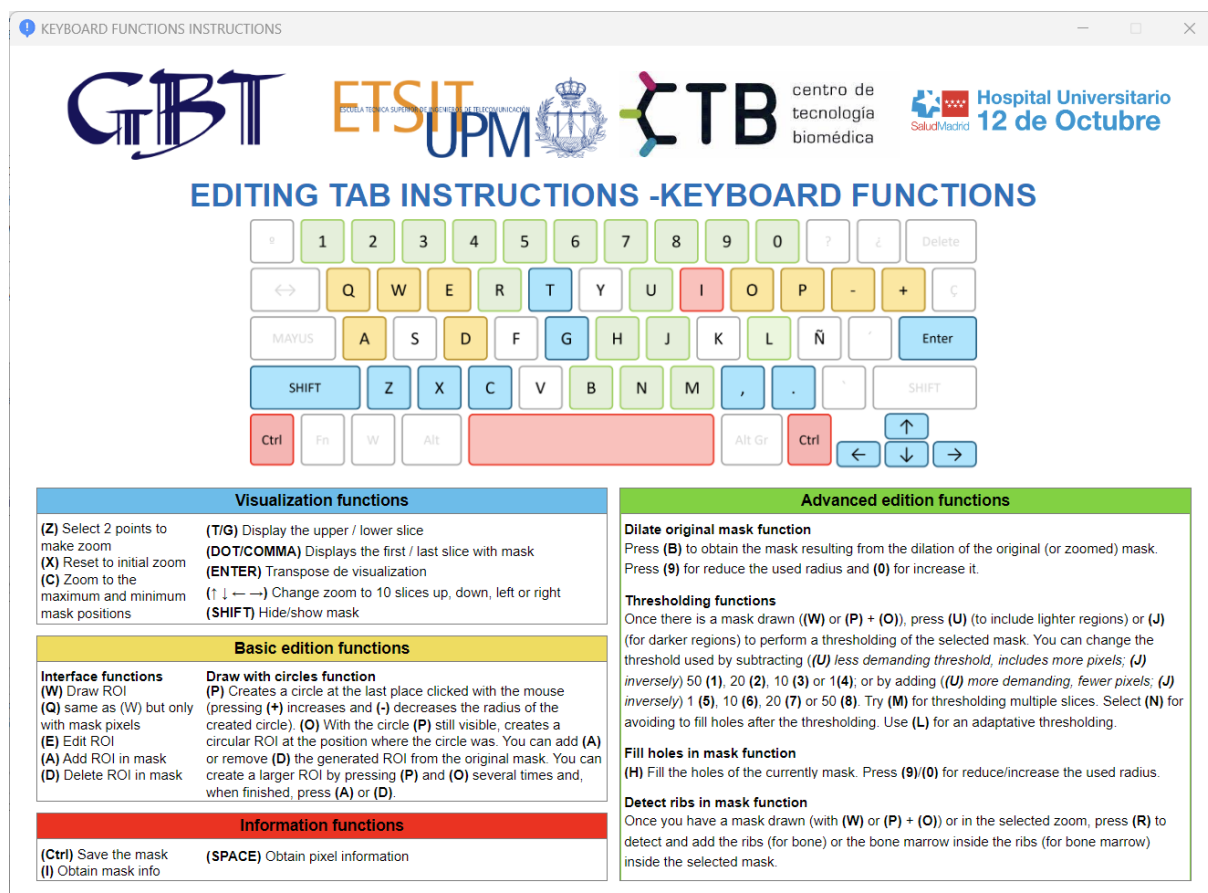


Figure 90. 'Editing' tab of the semi-automatic segmentation platform: 'Keyboard functions instructions'.

The visualization functions include three keys related to zoom: ‘Z’ for selecting two opponent corners of the rectangle to be zoomed; ‘X’ reset zoom to the original one, i.e. the limits of the slice in horizontal and vertical directions; and ‘C’ zoom to the mask limits in horizontal and vertical directions for all the volume. In addition, the key arrows ‘↑↓←→’ allow moving the zoom 10 pixels to the selected direction. This way, the user can change the ideal visualization without the need for resetting the view and selecting again the zoom manually. In addition to these functions, other which are applied in the current slice are a transposition of the visualization by pressing ‘ENTER’ or hiding/showing the mask by pressing ‘SHIFT’. For changing the visualised slice, the ‘.’ key changes the current slice for the first, i.e. the slice with mask closer to the feet, and the ‘,’ key changes the current slice for the last, i.e. the slice with mask closer to the head. As mentioned before, ‘T’ and ‘G’ allow changing the current slice for the next above or below, respectively.

The basic edition functions first include the same functions as the ones explained in subsection 5.1.1.2.3. *Edition functions panel*. In addition, the key ‘Q’ allows creating a ROI selecting a concrete area but including only the mask pixels, not everything drawn. For creating a ROI, there are other function, associated to the ‘P’ key, which creates a circle in the last place where the mouse has clicked. The circle can be made bigger by clicking ‘+’ or smaller by clicking ‘-’. Once the circle is the correct size and position, since it could be also moved with the mouse, by clicking ‘O’ key, the ROI on this circle position is created. For adding or removing those pixels, again the keys ‘A’ or ‘D’, or their respective buttons, are used.

The information functions are the ones shown in the red panel, and they include: information of the mask by clicking ‘I’ key and information of the last clicked pixel by clicking the ‘SPACE’ key. In addition, and as mentioned before, the user can save the mask by clicking the ‘Control’ key.

The advanced edition functions are developed to provide solutions to problems encountered through the use of the initial version of the implemented platform: dilating and thresholding the mask, filling holes or detecting ribs. First, the ‘B’ key allows dilating the mask present in the current slice, with the selected zoom, in order to use it as ROI, useful for example for a posterior thresholding. This dilation is based on a structural element of shape disk with a radius of 1 pixel, which can be changed pressing ‘9’ key to reduce the radius and ‘0’ to increase it, moving only between integer values. Secondly, for thresholding the mask, once a previous ROI is created, the ‘U’ key allows to select from the ROI the pixels of lighter regions, i.e. with an intensity value higher than the determined threshold, and ‘J’ the darker regions (as shown in Figure 91), i.e. with an intensity value lower than the threshold. The original threshold is 300 HU, and it could be changed to lower values by removing 50, 20, 10 or 1 HU to the threshold with the keys ‘1’, ‘2’, ‘3’ and ‘4’,

respectively, or by adding 50, 20, 10 or 1 HU with the keys ‘8’, ‘7’, ‘6’ and ‘5’, respectively. The value of the threshold is shown in the ‘Visualization functions’ panel next to the selected slice. If the ‘N’ key is clicked before the ‘J’ or ‘U’ ones, there is not a smoothing function applied after the thresholding. Another way of thresholding is creating a ROI and then clicking the ‘L’ key, which implements an adaptative threshold, based on the mean value of the ROI. Finally, the extra functions to fill holes or ribs previously mentioned with ‘H’ and ‘R’, respectively.

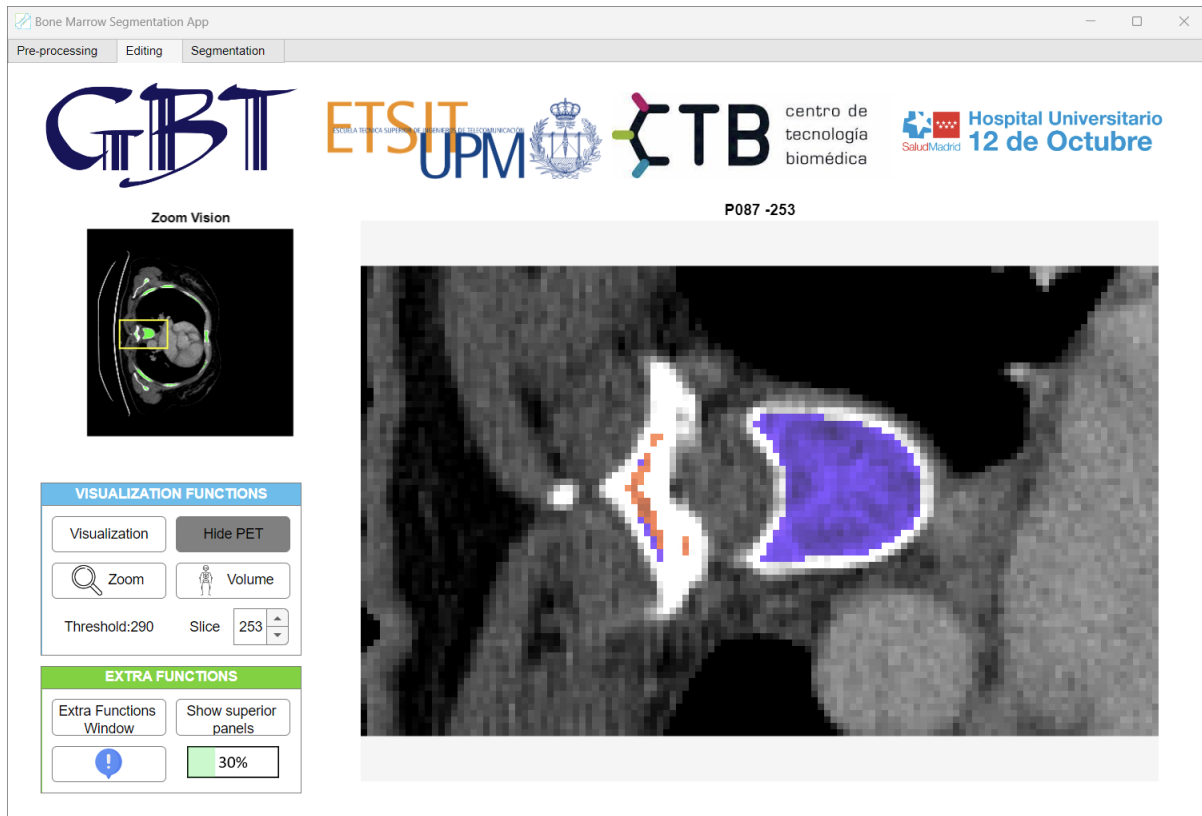


Figure 91. ‘Editing’ tab of the semi-automatic segmentation platform: thresholding example. Purple: mask; orange: ROI created in the thresholding process.

5.1.2. Semi-automatic segmentation based on traditional algorithms

As mentioned above, the first version of the platform allows the segmentation based on traditional algorithms, before manual revision of the masks. As shown in figures included in subsection 4.2.1. *Semi-automatic segmentation based on traditional algorithms*, bone and BM segmentations are visually successful, although those images are from only the automatic part. As mentioned before, the manual revision of bone and BM masks for using them as GT of the DL model, are also used for evaluating quantitatively the performance of this segmentation based on traditional algorithms.

5.1.2.1. Bone segmentation

In order to obtain a fairer comparison between segmentation approaches, the bone segmentation is considered after the spinal canal removal step. Figure 92 and Figure 93 show examples of bone segmentations based on traditional algorithms with satisfactory and unacceptable performances, respectively. As seen, for satisfactory results cases, there are no false negatives pixels, i.e. the algorithm fails adding more pixels than those corresponding to bone, but there are no missing bone pixels. However, even in these satisfactory examples there are some mistakes, especially in the elimination of low grey level intensity regions between compact bones, such as regions between pelvis bones (Figure 92be). In addition, the spinal canal removal fails mainly in the pelvis region (Figure 92abde), having a good location in regions near to the ribs (Figure 92cf). Then, there are some edges which fail to be recognised, but it does not imply a significant mistake (Figure 92df).

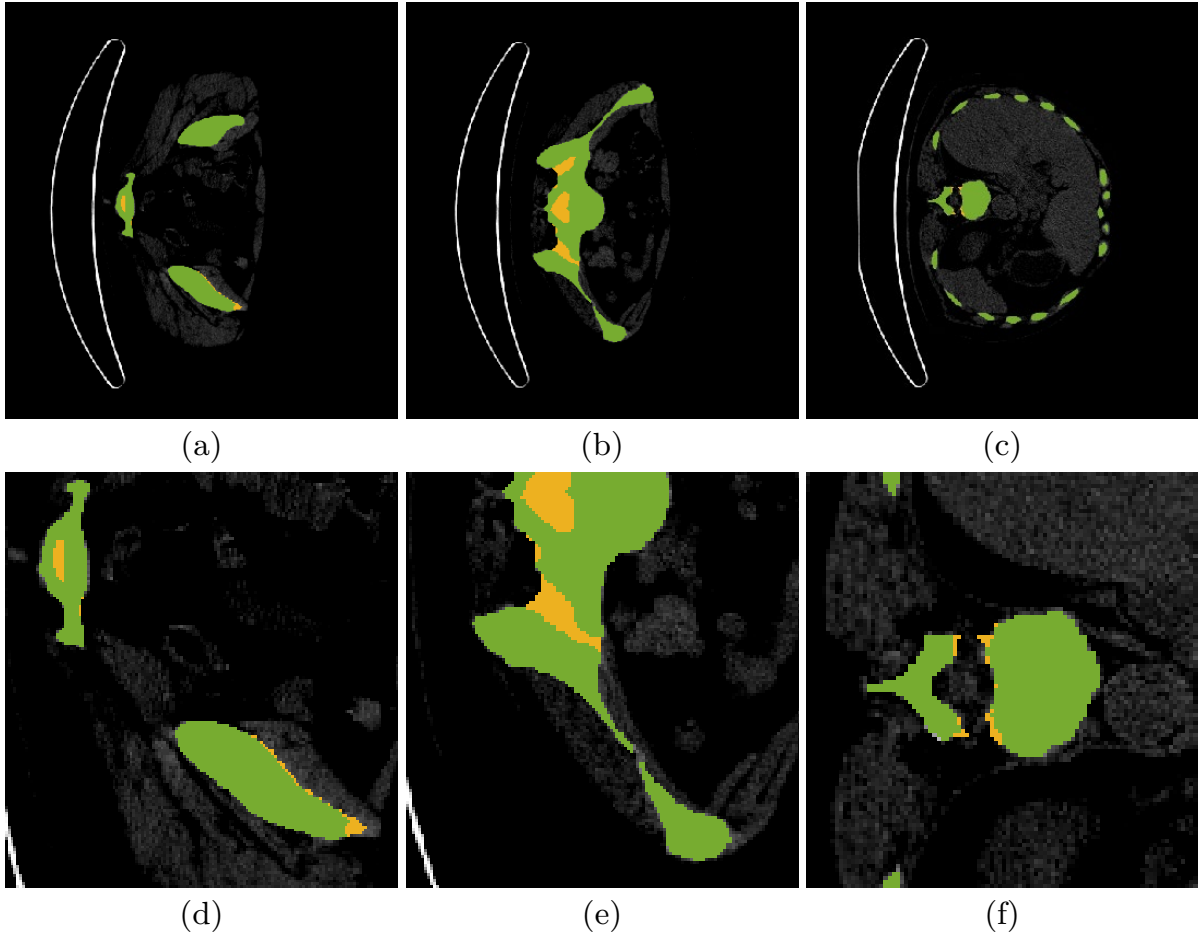


Figure 92. Examples of a satisfactory bone segmentation based on traditional algorithms: (a and b) pelvis region and (c) ribs region, (d), (e) and (f) are zoom of (a), (b) and (c), respectively. (Green) true positive pixels; and (yellow) false positive pixels.

On the other hand, in the cases of unacceptable performance, the main mistakes are the definition of soft tissue as bone due to a high grey intensity level, such as in the

liver (Figure 93ad), or heart (Figure 93be). In addition, there are incorrect location of humeri when patients locate their arms above their head, which imply an exclusion of the humeri and a failing in localising the shoulders. Finally, as shown for this patient, there are some spinal canal locations which are impossible for the algorithm, generating a complete inclusion of this region, or missing small regions such as the shoulder blades (Figure 93cf).

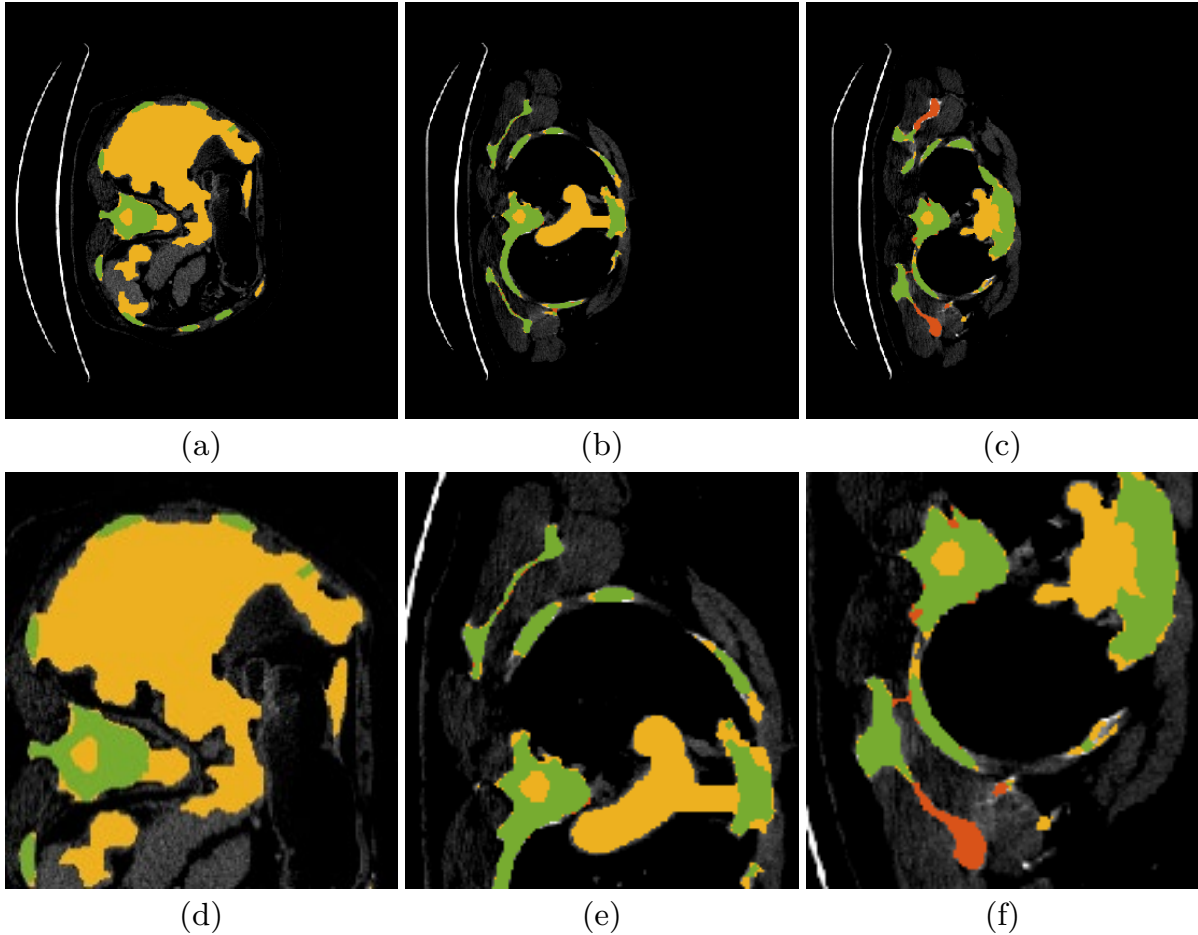


Figure 93. Examples of a satisfactory bone segmentation based on traditional algorithms: (a) lower torso, (b) mid-chest torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.

As seen thanks to these examples, the segmentation of the bone based on traditional algorithms can result in outstanding results (Figure 92cf), but some of them fail in many regions (Figure 93). It must be considered that this approach is developed as an unsupervised segmentation evaluated only in some of the patients. For these reasons, the results of Jaccard and Dice fit perfectly with the visually evaluated results with values of 0.811 and 0.904 for IoU (mask) and IoU (mean), and 0.896 and 0.947 for Dice (mask) and Dice (mean), respectively.

5.1.2.2. Bone marrow segmentation

As for the bone segmentation assessment, Figure 94 and Figure 95 represent examples of BM segmentations based on traditional algorithms with satisfactory and unacceptable performances, respectively. Nonetheless, BM segmentation is more challenging to obtain than bone segmentation. For this reason, even segmentations with the best performances have some mistakes. As illustrated in Figure 94, although satisfactory segmentations obtain the vast majority of the BM region correctly segmented, there are several mistakes in the edges, i.e. in the differentiation of compact and trabecular bone. This mistake is acceptable while less pixels are taken (red) and the segmented mask is truly the BM (green), i.e. for the quantification it is better to select as mask where every pixel is BM though missing some from the edges than taking part of the compact bone. However, pixels of compact bone interpreted as BM mask are also present in this segmentation (yellow), obtaining an inaccurate segmentation even for the patients with a satisfactory segmentation.

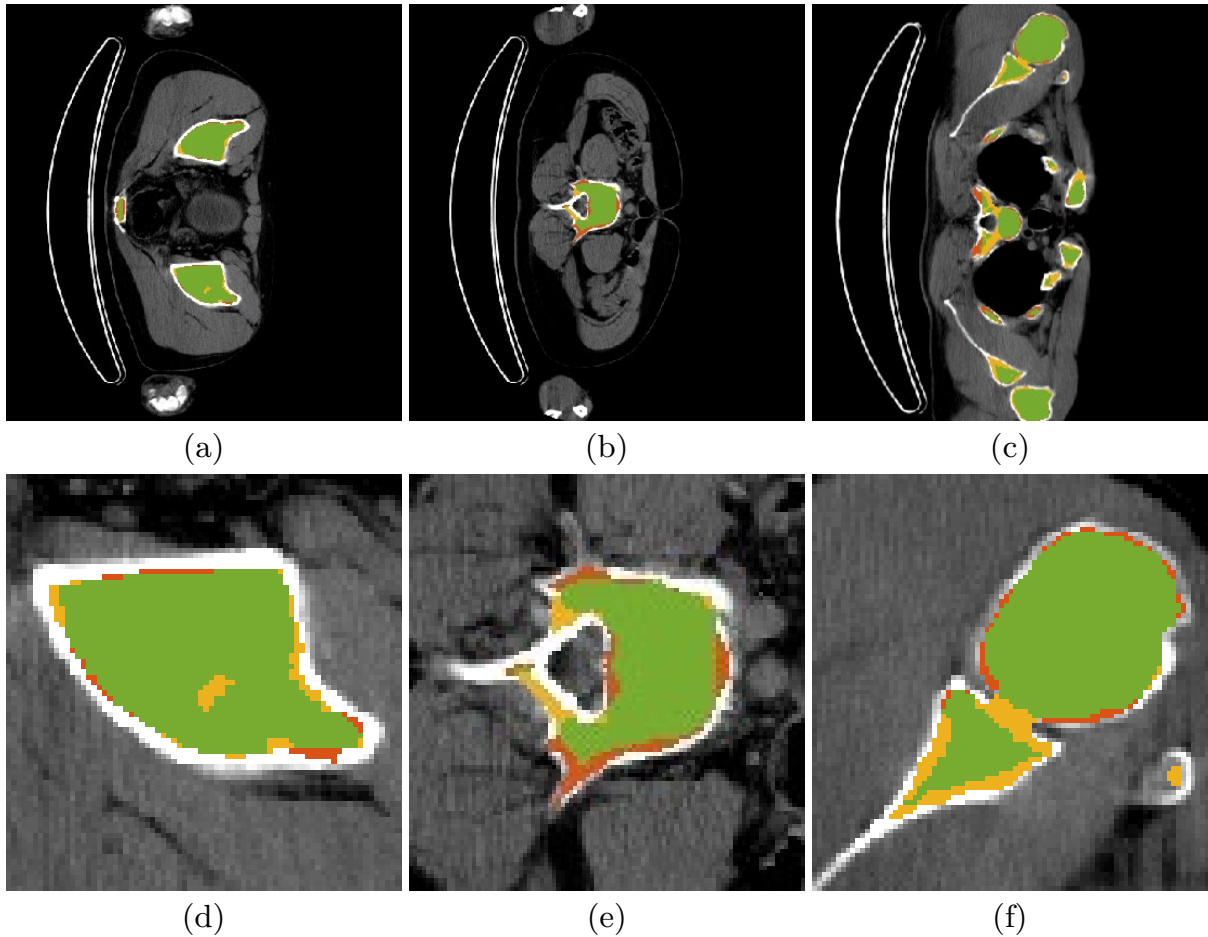


Figure 94. Examples of a satisfactory BM segmentation based on traditional algorithms: (a) pelvis, (b) lower torso and (c) shoulders region, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.

In addition to those patients with good or satisfactory segmentations, there are many of them with an unacceptable segmentation, such as the patient included in Figure 95. In this case, other mistakes are observed. First, as a consequence of an incorrect location of compact bone (Figure 92ad), the algorithm interprets the space with a low grey level value between compact bone separated structures as BM, same as spinal canal when it is not previously removed during compact bone segmentation (Figure 95ad). This is also due to the low grey level intensity of the thin compact bone of the iliac crest or breastbone, which are difficult to distinguish from the BM. Secondly, due to their size, there are small regions composed of only some pixels, which are not correctly located (Figure 95be), such as spaces in the vertebrae or the ribs. Finally, due to the fact that the ribs lie diagonally to the axial slices, ribs are visualised as structures for which the compact bone does not completely surround the BM region, complicating their location (Figure 95bcf).

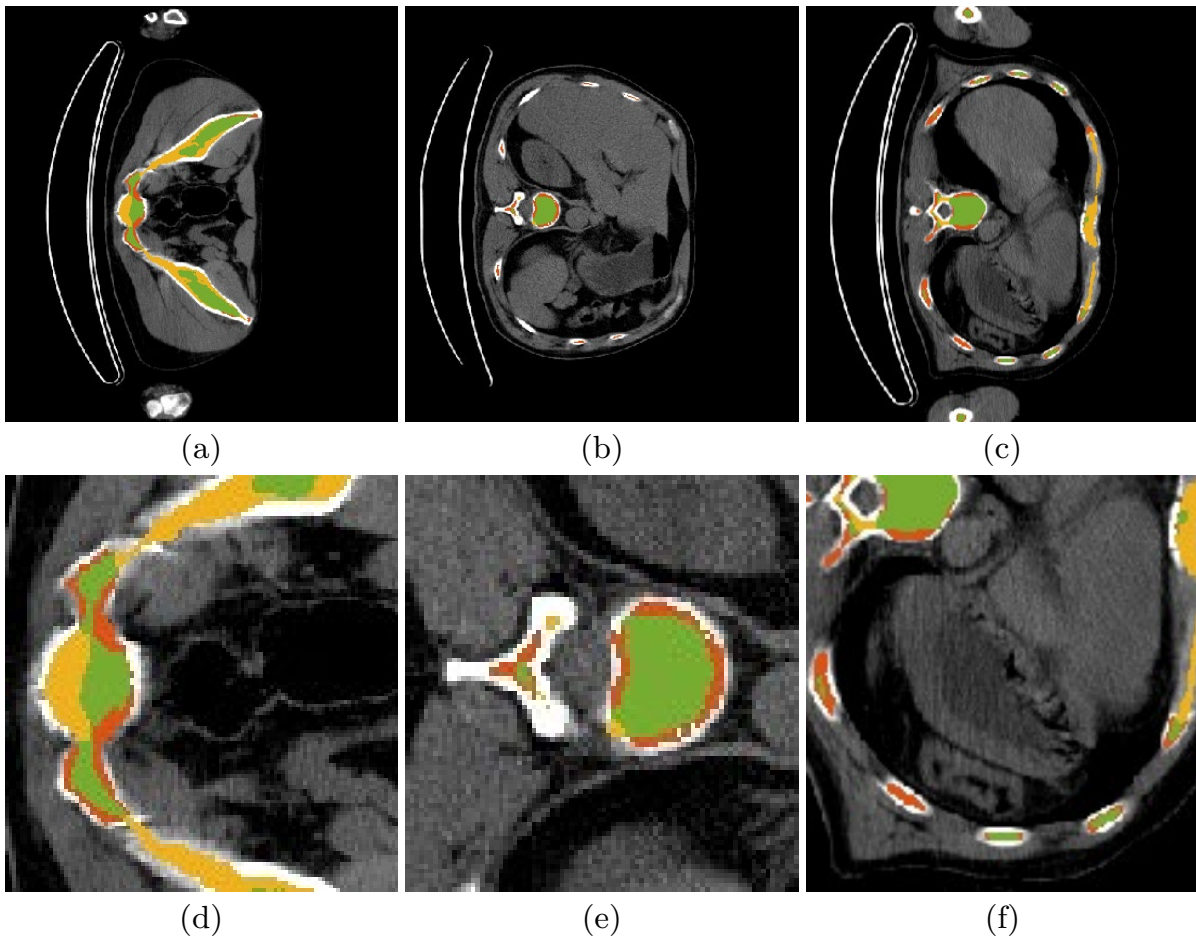


Figure 95. Examples of an unacceptable BM segmentation based on traditional algorithms: (a) pelvis, (b) lower torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.

In order to make fairer the comparison of BM segmentation approaches, as for bone segmentation, segmentation based on traditional algorithms is tested with the same

patients of the DL model test set, although it is developed as an unsupervised segmentation. The metrics of this test are an IoU (mask) of 0.496, IoU (mean) of 0.734, Dice (mask) of 0.650 and Dice (mean) of 0.818.

5.1.3. Automatic segmentation using Deep Learning models

In this subsection, the bone and BM segmentation based on DL models results are detailed. Since the bone segmentation trained with unprocessed data is only a mean to obtain the pre-processing steps, its results are included in *Annex E. Unprocessed bone segmentation*.

5.1.3.1. Bone segmentation

The adapted U-Net architecture is created, trained, tested, and assessed for every set of values: F, epochs, and training order. Figure 96 shows the model results of the training order with the highest value of performance metric for each pair of model parameters F and epochs. As seen, all models except the model with F=2 and 5 epochs obtain a value over 0.9 for all the indexes. In terms of epochs variation, comparing models with same value for F parameter, all models with 5 epochs obtain worse performance than models with a higher number of epochs. However, the performance results for models varying from 10 to 20 epochs does not follow a stable tendency, i.e., a higher value of epochs does not imply a better performance. In terms of the F parameter, models with F=2 obtain worse results than the rest of tested models. However, models with the same number of epochs but differing between F=4 and F=8 obtain similar results.

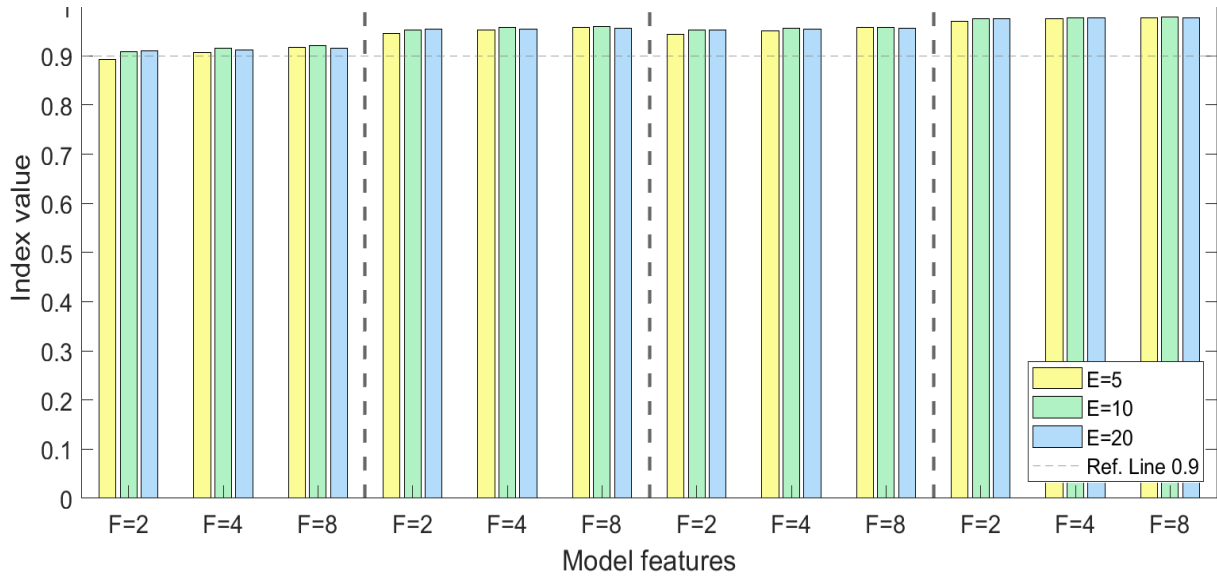


Figure 96. Comparison of IoU and Dice metrics for different combination of parameters F, E and training order. F: number of channels in each feature map; E: number of epochs in the training process.

For comparing all the models, and not only those with the best Dice for the pair F and number of epochs, Figure 97 shows a heatmap with the mean Dice index for all the combinations of parameters F, number of epochs and training sets order. Some combinations of parameters obtain a model which classifies every pixel as background, i.e., the model fails to converge into a segmentation. Although the simpler (lower values of F) and less trained (lower number of epochs) the model, the greater the chances of obtaining a convergent model, the results are worse than those for more complex and trained models (higher values of F and number of epochs). Considering all the training orders, a total of 62 (69%) convergent models are produced. In terms of training sets orders, some such as 2, 4 and 10, achieve convergent models for more combinations of F and number of epochs, with a total of eight convergent combinations in all three cases. Moreover, the best mean Dice index is obtained for one of these training sets orders, concretely the tenth order, with a value of 0.979 for F=8 and 10 epochs. The rest of metrics for this optimal model are, in terms of IoU index, 0.920 and 0.959 for mask and mean between mask and background, and 0.958 for Dice (mask).

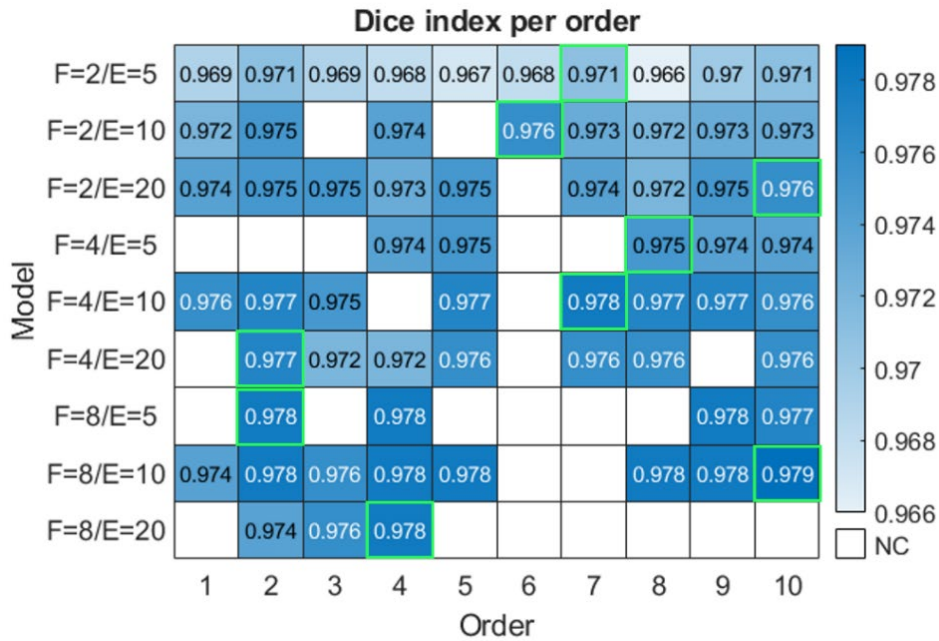


Figure 97. Comparison of training orders for each bone segmentation model, i.e., different combination of F parameter and number of epochs (E). NC: Non-Convergent. In lighting green are shown the best result for each pair of feature models: the best Dice Index; if same value of Dice Index, the best Dice (mask) Index; if same value of Dice (mask) Index, the model with higher value of TP.

In terms of visual assessment, bone segmentation based on U-Net architecture obtains in the majority of slices a satisfactory segmentation, in line with the quantitative results. Figure 98 shows examples of a satisfactory segmentation for this model. The main mistake of this segmentation method is the inclusion or exclusion

of pixels on the edges, as seen in Figure 98ad, especially in small bones which are usually characterised by lower grey level values in the compact bone, such as ribs (Figure 98be), collarbone or scapula (Figure 98cf).

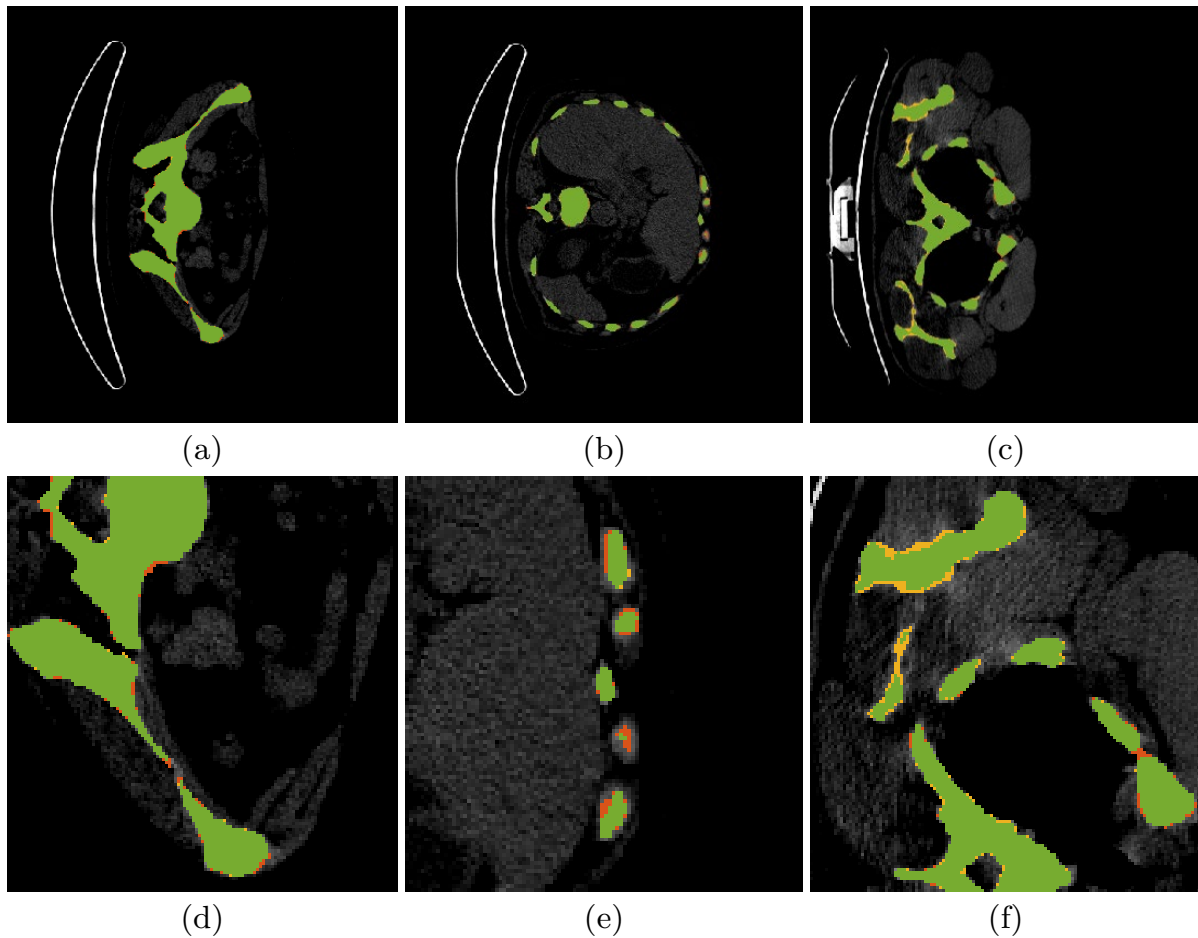


Figure 98. Examples of a satisfactory bone segmentation based on DL model: (a) pelvis, (b) mid-chest torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.

Although segmentation is mainly precise, there are mistakes not related to the edges. First, there are some misleading structures due to the high grey level intensity in the CT images, such as the intestine in Figure 99ad. Secondly, there are some missing areas inside bone which are mainly correctly segmented as bone, as seen in Figure 99be where the interior part of the shoulder is missing. Finally, there are some regions, especially in the upper torso near to shoulders and breastbone, with a medium grey level of intensity in its compact structure, which are completely missing (Figure 99cf).

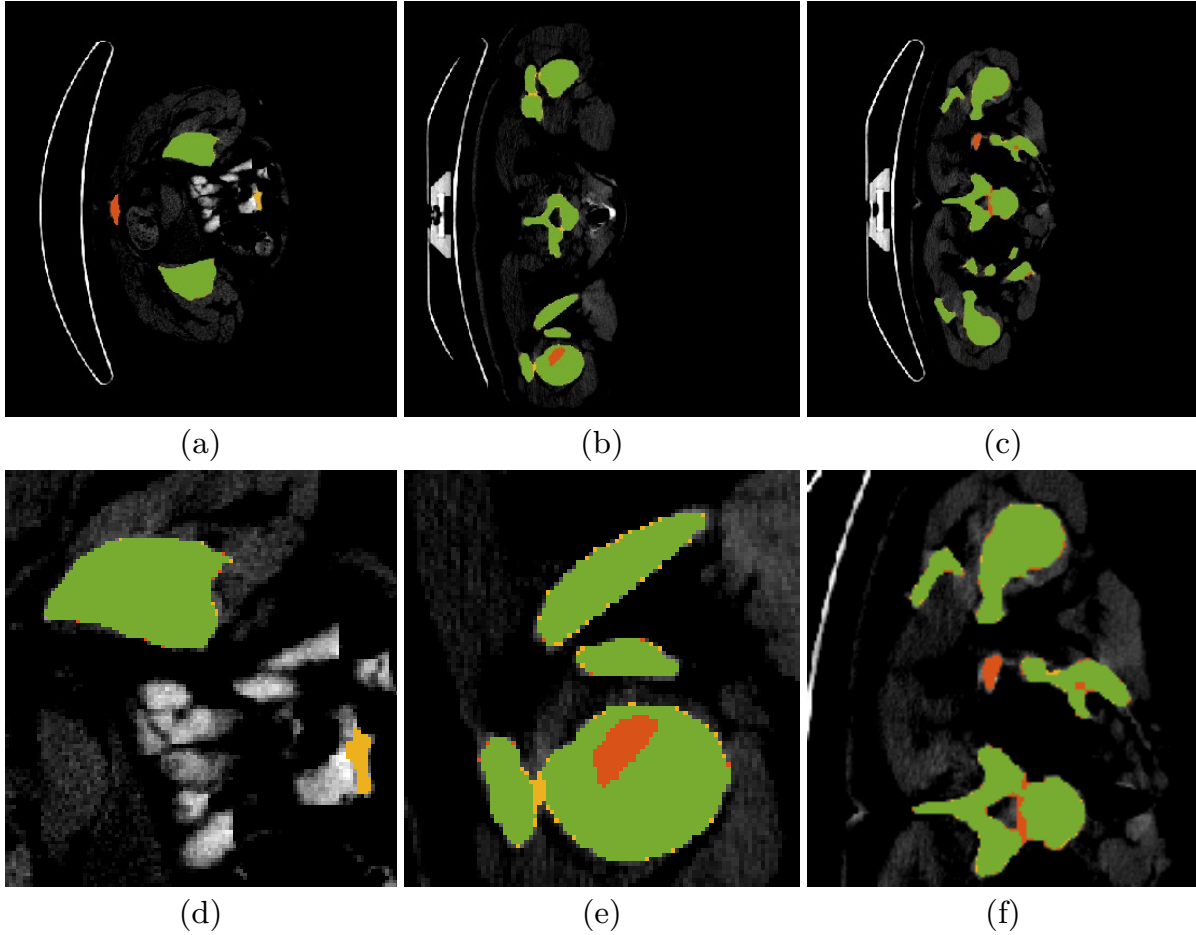


Figure 99. Examples of main mistakes of bone segmentation based on DL model: (a) misleading areas in pelvis region, (b) missing areas inside correctly segmented structures in upper torso and (c) missing structures in upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.

5.1.3.2. Bone marrow segmentation

Once the optimal model for bone segmentation is selected, the different possibilities of BM segmentation model are tested. However, there are some considerations to be taken account before detailing BM segmentation results: the pGT creation and the pre-processing adaptation.

As mentioned in subsection 4.2.2.2. *Bone marrow segmentation*, since manual segmentation of BM requires an excessive amount of time to get enough samples to train a model, a pGT creation process is implemented. Figure 42 and Figure 43 show the examples of the pGT, which are visually satisfactory despite the inclusion of spaces between compact bone which are not BM and the inclusion of thin compact bone as BM. It is also quantitatively measured in order to provide a fairer comparison with the models. However, the comparison is made before the model training, so it is tested in the 13 manually segmented patients, obtaining IoU and Dice index of 0.776 and 0.874 for the mask, and 0.887 and 0.912 for the mean. It must be considered

that, using as original mask the ones manually revised, there are no mistakes in the bone masks. However, if pGT mask creation would be used as the methodology of BM obtention, the masks will have two biases, one from the pGT creation and the other from the bone segmentation model mistakes. Furthermore, the obtention would depend on the accuracy of the bone segmentation.

In terms of the pre-processing adaptation, the stretcher removal step results are shown in *Annex F. Stretcher removal model results*. Once all combinations of parameters are tested, the optimal stretcher removal model has the following metrics: IoU(mask) of 0.977, IoU(mean) of 0.989, Dice(mask) of 0.989 and Dice(mean) of 0.992.

Once the steps previous to the training are defined, as mentioned in subsection 4.2.2.2. *Bone marrow segmentation*, three different approaches, three different image collections and two training phases are considered. In Table 11, results for the different combinations of approaches, collections and training phases are shown in terms of IoU and Dice indexes. As seen, there are some image collections, B and C, those in which pGT is obtained from bone masks computed by the bone segmentation model, where the U-Net model cannot converge, at least for approaches II and III. This impossibility to converge can be caused by the fact that those bone masks are obtained with the model and, although providing a precise segmentation, it still performs worse than the manual process. In line with this result, the convergent models for B and C, i.e. the ones in the approach I, are worse than the results for A. On the other hand, approaches II and III obtain exactly the same metrics, with less FP in the approach III and less FN in the approach II. However, the approach III does not depend on a manual revision of the bone mask for obtaining these metrics, while the approach II does. It must be taken into account, that for all the possible combinations, the second training phase implies an improvement of the performance metrics, achieving a maximum performance for the database A and approach III. For this optimal model, the values for IoU are 0.866 and 0.933 for mask and mean, respectively, and for the Dice they are 0.928 and 0.964, respectively.

Approach	DDBB	Phase I				Phase II			
		IoU		Dice		IoU		Dice	
		Mask	Mean	Mask	Mean	Mask	Mean	Mask	Mean
I	A	0.770	0.884	0.870	0.935	0.818	0.909	0.900	0.950
	B	0.621	0.809	0.766	0.882	0.774	0.886	0.872	0.936
	C	0.620	0.808	0.765	0.882	0.741	0.870	0.851	0.925
II	A	0.786	0.892	0.880	0.940	0.615	0.806	0.762	0.880
	B	NC	NC	NC	NC	X	X	X	X
	C	NC	NC	NC	NC	X	X	X	X
III	A	0.786	0.892	0.880	0.940	0.866	0.933	0.928	0.964
	B	NC	NC	NC	NC	X	X	X	X
	C	NC	NC	NC	NC	X	X	X	X

Table 11. Results of BM segmentation models, varying the approaches, image collections and training phase. Approaches based on: (I and II) untrained U-Net model using as input (I) the pre-processed CT scan; and (II) the pre-processed CT scan in the regions of the bone mask; (III) U-Net model trained for the bone segmentation using as input the CT pre-processed scan, as (I). Image collections: (A and B) same cohort used for the bone segmentation model and (C) larger cohort than the one used for the bone segmentation model, using as bone masks (A) those manually segmented for being the GT of the bone segmentation model, and (B and C) those obtained from the bone segmentation model. NC: non-convergent, X: not applicable since the previous model is NC. The model with the best performance is marked in green.

In terms of visual evaluation, Figure 100 shows examples of satisfactory BM segmentation with the optimal model previously defined. As observed, the majority of mistaken pixels are from the edges of the BM mask, similar to bone segmentation. However, in this optimal segmentation, there are only a few pixels, being most of the mask correctly segmented.

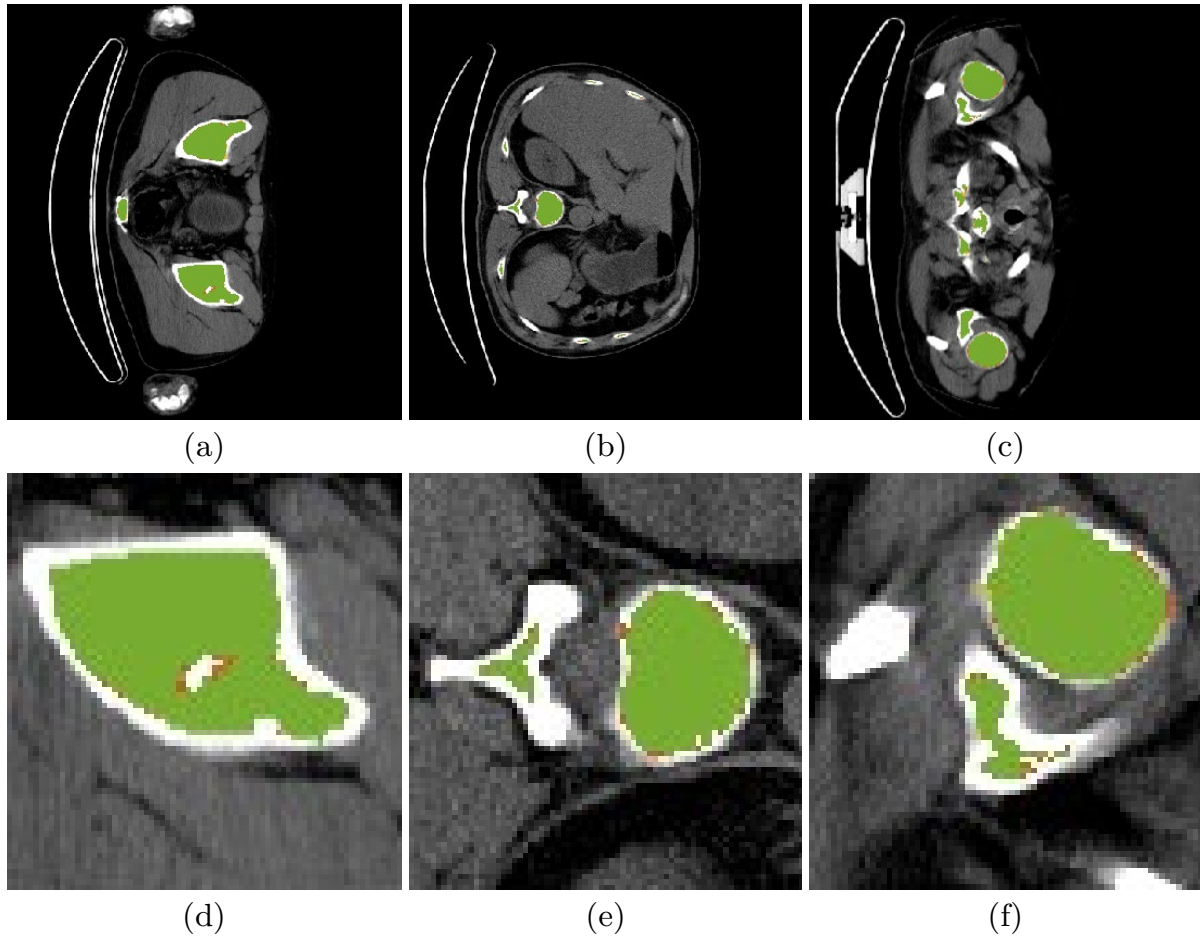


Figure 100. Examples of a satisfactory BM segmentation based on DL model: (a) pelvis, (b) mid-chest torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.

Although this model obtains precise segmentations for some patients, there are some considerable mistakes which appear in other patients. First, some regions corresponding to thin compact bone, such as in the pelvis region (Figure 101ad) or the breastbone (Figure 101be), are interpreted as BM. Also in the pelvis, some BM areas in the central part are interpreted as compact bone, due to its high intensity (Figure 101a), or some regions which are only surrounded by a thin compact bone are considered background, since there is not a high intensity closed region which isolates the interior BM (Figure 101becf). Finally, in the shoulders area, the visualization of the final part of bones is delicate, since it appears to be diffuse when it is a thin layer of compact bone, making the model interpret them as BM.

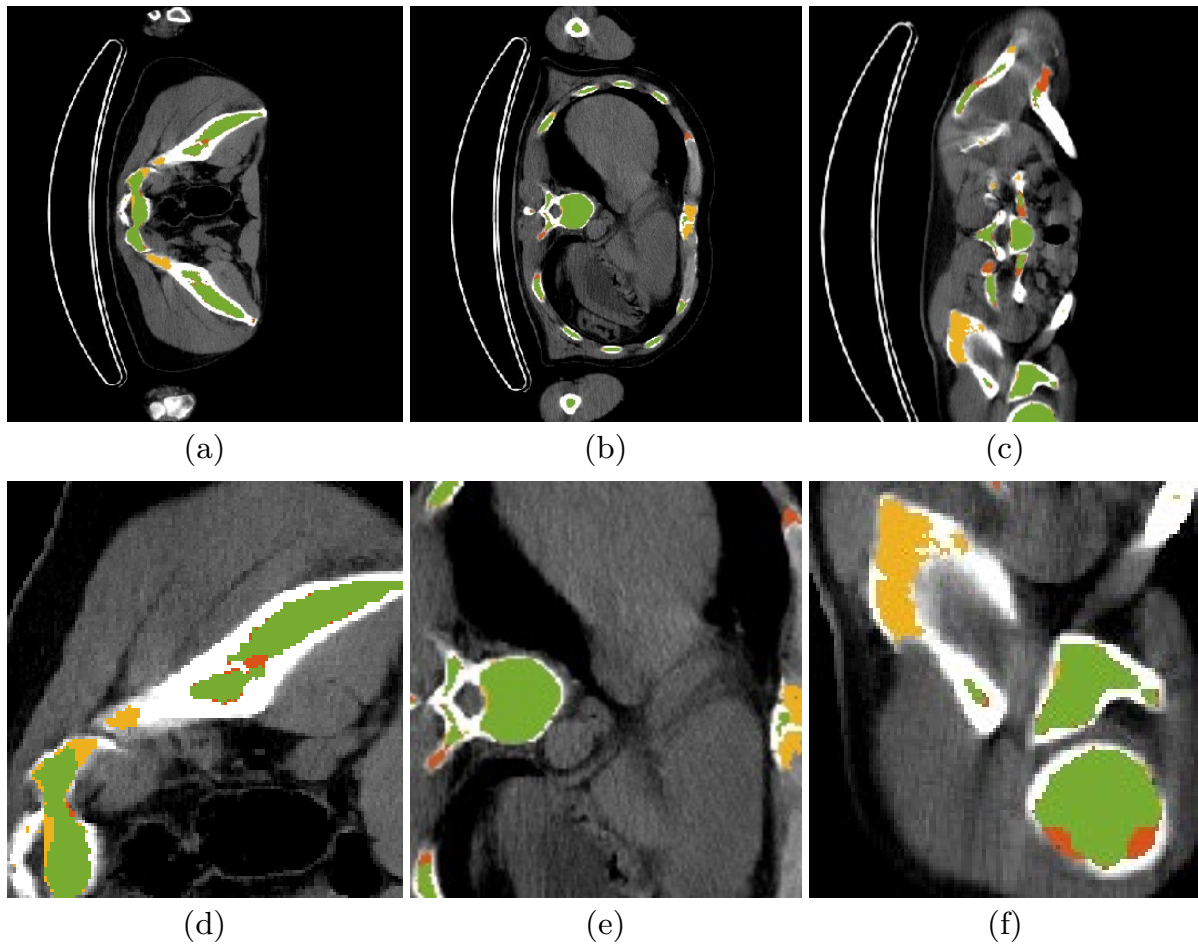


Figure 101. Examples of the main mistakes of BM segmentations based on DL model: (a) pelvis, (b) mid-chest torso and (c) upper torso, (d), (e) and (f) are zoom of (a), (b) and (c) respectively. (Green) true positive pixels; (yellow) false positive pixels; and (red) false negative pixels.

5.1.4. Discussion

An adequate quantification of the BM is only possible with a precise segmentation. Although there are multiple studies which try to segment this region [82–85,107,108,112,113], none of them perform a segmentation for the whole-body, and the only one focused on the complete region, does not evaluate the algorithm performance. Most of them develop the segmentation based on a previous bone segmentation. In addition to the use for a posterior BM segmentation, an accurate bone segmentation can be useful for multiple clinical applications such as bone metastases assessment, evaluation of oncological therapies or bone dosimetry measurement.

Although a manual segmentation requires a huge amount of time, a platform for manually segmenting the masks which are posteriorly used as GT for training the automatic models is developed. This platform is also composed of a tab for an automatic segmentation, adapted to the best segmentation methodology obtained in

the present thesis. Then, the tab for manually editing the masks includes a great variety of functions to help the user in the segmentation process. Although the basic functions are clearly indicated in the main window, the most concrete ones are included in the keyboard or auxiliar windows. This variety can complicate the understanding of the platform, tending to not use its whole potential. However, there are multiple information windows which can be consulted by the user in any moment of the segmentation process. In addition, there are many visualization settings to be adapted to the segmentation problem. In spite of the limitations of this platform, especially its complexity, it is a complete platform to be used for both bone and BM segmentation. Moreover, this platform allows not only BM or bone segmentation, but also any organ manual segmentation based on CT imaging visualization, since the ‘Editing’ tab can be used for segmenting any organ of interest. It also guarantees a visualization of PET imaging over CT imaging, allowing the visualization settings to be changed, helping the user to evaluate the functional processes in the body or a concrete ROI. The aims of future works for this platform are focused on its implementation in the clinical practice for allowing physicians to segment any interest organ with research purposes. With this purpose, many adaptations of the platform could be applied, such as other organs segmentation.

In the present thesis, in addition to a manual segmentation through the platform, two different automatic segmentation approaches for bone and BM are proposed. The first is based on the combination of multiple traditional image processing algorithms, such as thresholding or morphological operations. Then, using the segmentations obtained from this first approach and the platform developed in the present thesis, for a manual revision of the masks, a DL model based on the U-Net architecture is trained as the second approach. For an easier understanding Table 12 summarises the performance metrics for the different segmentation methods implemented in the present thesis, differentiating between bone, BM test set composed of the 13 manually revised masks, i.e. first phase DL model training, and BM test set for second phase DL model training, i.e. the one composed only by 7 out of the 13 manually revised masks. As it can be seen, DL models exceed the performance of the traditional approach for both segmentations.

		IoU			Dice		
		Mask	Bg	Mean	Mask	Bg	Mean
TRAD.	Bone	0.811	0.997	0.904	0.896	0.999	0.947
	BM I	0.454	0.972	0.713	0.603	0.985	0.794
	BM II	0.496	0.973	0.734	0.650	0.986	0.818
pGT	BM I	0.776	0.999	0.887	0.874	0.999	0.912
DL	Bone	0.920	0.998	0.959	0.958	0.999	0.979
	BM I	0.786	0.998	0.892	0.880	0.999	0.940
	BM II	0.866	0.999	0.933	0.928	0.999	0.964

Table 12. Summary of the segmentation results for both, bone and BM. TRAD: segmentation based on traditional algorithms; BM I: test set for phase I of DL training, with the 13 manually segmented masks; BM II: test set for phase I of DL training, with 7 out of the 13 manually segmented masks; Bg: background.

Although the first approach for bone segmentation obtains satisfactory performance, with a Dice (mean) and Dice (mask) of 0.947 and 0.896, the developed algorithm is sensible to changes in new CT images. First of all, the algorithm is adapted to the original database, since the percentage of slices considered head or below the knees depends on the total number of slices. However, if acquisition parameters were changed or cases with a different total number of slices were included, there would be even more mistakes in head or limbs removal. In addition, the spinal canal removal process is not good enough in the region of the pelvis, interpreting it first as bone and, consequently, with more probability then as BM. Furthermore, any region with a high intensity level is considered bone, such as liver (Figure 93ad), heart (Figure 93be), or intestine, rendering inadequate the segmentation, due to the high extension of this regions. Despite these mistakes, the results of this segmentation approach are competitive with other studies shown in subsection 3.1.1. *Bone segmentation* which also applied traditional algorithms [77,98–100] and achieve Dice (mask) between 0.8 and 0.948, but only for some bones, none of them for the whole-body skeleton. Thus, the algorithm presented in the following thesis is an improvement over those focusing on only one region of the skeleton. Moreover, considering that the performances of the previous studies are analysed in specific regions, their better performance can be understood by the homogenisation of the segmented structures, compared to the whole-body segmentation of this thesis.

On the other hand, the first approach for BM segmentation obtains an automatic segmentation of the whole-body, not found in any other study based on traditional algorithms of the subsection 3.1.2. *Bone marrow segmentation*. There are also no

studies based on traditional algorithms which quantitatively evaluate the performance of the BM segmentation, since all of them are unsupervised. Although this first approach supposes a breakthrough in the BM segmentation field, it is still not precise enough to be used for BM quantification, achieving a Dice(mask) of 0.650 and Dice(mean) of 0.818. Moreover, observed in Figure 94 and Figure 95, even the best segmentations include large parts of compact bone as BM, and regions with low grey level intensities such as spinal canal or spaces between compact bone.

In the DL approach, the robustness of the proposed methodology for bone segmentation, including both model building based on the U-Net architecture described in Ronneberger *et al.* [102] and pre-processing definition based on unprocessed training, is demonstrated. This robustness is considered demonstrated based on two factors: the similarity of the performance of the models (Figure 97) and the improvement in the performance compared to unprocessed training (*Annex E. Unprocessed bone segmentation*). The variation of the F parameter demonstrates the greater the number of channels of the initial feature map, and consecutively the rest of feature maps that make up the network, the higher the performance, i.e., the more complex the model due to the increase of the number of channels or filters for feature maps, the better the performance. On the other hand, the variation of the number of epochs demonstrates that for all values of F, the results obtained with 5 epochs are lower than those obtained with 10 and 20 epochs, observing minimal changes between these last two values. For this reason, the test of this model is limited to these values for number of epochs with the aim of avoiding an overfitting in the model. The best performing model is the one obtained for $F = 8$ and 10 epochs, in line with the trends observed in the individual study of each parameter. The performance of this model achieves a IoU(mask) of 0.920, IoU(mean) of 0.959, Dice(mask) of 0.958 and a Dice(mean) of 0.979. The performance surpasses the first approach proposed for the segmentation based on traditional algorithms, also observed in Figure 98 and Figure 99, where even the main mistakes are visually less remarkable than those usually found in the first approach. The main reason for this improvement is due to the fact that the first approach applies image processing based solely on morphological processing based on grey intensity level, whereas the DL approach is based on more complex features.

Comparing this model with other studies of bone segmentation based on CNNs included in Table 2, the developed DL model narrowly exceeds the results obtained in Klein *et al.* [78], study with a Dice (mask) of 0.95. In spite of using the same architecture, the study by Klein *et al.* implements a more complex version, with a higher value of F. However, their lower results may be due to (1) the robustness of the dataset compared to the present study, since they used only 18 patients; and (2) the absence of pre-processing steps, specially the one for stretcher removal. On the

other hand, the study by Noguchi *et al.* [79] results in a model with a better performance, possibly due to the robust data augmentation methodology implemented. Furthermore, this study uses a network architecture also based on U-Net less complex than the study by Klein *et al.*, but more than the present thesis.

For the optimal or best performing model, the following limitations are observed in Figure 98 and Figure 99. First, segmentation of misleading structures (Figure 99ad). Structures with a high level of grey intensity such as internal organs (e.g. intestine, heart, liver) in rare cases, and medical devices (catheters, prostheses, etc.), are segmented as part of the bone mask. However, it is still less sensible to these regions than the first approach (Figure 93abde). Secondly, segmentation with missing structures (Figure 99abde). This mistake is observed in two different variants. The less common case, when a structure is completely missed, as the sacrum in Figure 99a, or small regions next to collarbones in Figure 99c. These regions are mainly missed due to their size, since they are small, i.e. composed by only a few pixels. This lack of the complete structure is also observed for humeri, since the GT for the model training includes bones only from the elbow even when the arm appears completely in the CT, i.e. when it is not covered or partially covered by the scanner structure. As a result, the model can overlook the humeri as part of the bone mask. Conversely, the distal bones below the elbow may also be included as part of the skeletal mask when the patient places the arms over the body. On the other hand, the most common case is when only some regions are missed, which causes incomplete structures such as ribs near to breastbone in Figure 98b or an incomplete shoulder as in Figure 99be. In this case, the main causes are a medium intensity level of these structures, such as ribs near to breastbone since they are mainly composed of cartilage, and the absence of a thick layer of compact bone in its perimeter and a BM region with very low levels of grey intensity which causes an overlook of some low-grey-level regions, such as ribs due to their diagonal location or shoulders in their internal region.

In terms of BM segmentation, the DL approach also obtains a satisfactory performance. It must be noted that different approaches based on input, model to train, and image collections are proposed for this aim. Despite the different possible combinations (Table 11), the performance of the models trained with the collection obtained from the manually segmented bone masks (A) is outstanding for all types of training, since it implies zero bias caused by the bone segmentation model. The other two collections cause a non-convergence of the models with a higher frequency. In terms of approaches, the approach I demonstrates a better convergence with worse results than the other two approaches, similar to the bone segmentation model, since the same bases are applied. In the first training phase, the remaining two approaches, i.e. II and III, obtain the same performance. However, the approach II depends on

the bone mask precision while III does not. In this way, for the cases where the bone is segmented with the bone obtained through the DL model instead of a manual segmentation, the BM mask would not be as good as the metrics this model represents. Due to the limitation of using a pGT for the training of the model, i.e. the model is trained with no realistic outputs, in the second phase these models are trained with manually segmented masks. For this reason, a remarkable improvement of the performance is observed, especially for the optimal model, i.e. the one with the best performance. This model achieves metrics of IoU (mask) of 0.866, IoU (mean) of 0.933, Dice (mask) of 0.928 and Dice (mean) of 0.964, importantly surpassing the results of the approach based on traditional algorithms for the BM segmentation. The visual assessment of the model results demonstrates that most pixels are correctly segmented, failing mainly in the edges of the BM regions. However, there are a few important mistakes. Same as for the bone segmentation, the model fails to consider BM the regions which are not surrounded in their totality by a clear layer of compact bone. For example, in Figure 101be some regions of the ribs appear in red, i.e. not interpreted as BM, due to their diagonal location in each slice. In addition, some regions from shoulders (Figure 101cf), marked also in red, are not interpreted as BM since the layer of compact bone is too thin or represented by medium grey intensity levels. This way, the BM inside these regions is interpreted as background by the model, creating missing structures or missing areas inside structures. Similarly, due to the intensity level of the cartilage which composes ribs near to breastbone, the model can over-segment BM. Despite these mistakes, overall segmentation is satisfactory, and substantially superior to the performance of the approach based on traditional algorithms, with a Dice (mask) of 0.928 vs 0.650 of this previous methodology.

Comparing this BM segmentation model with those included in Table 3, only the study of Wennman *et al.* [112] obtains a BM segmentation for the majority of bone structures with a nnU-Net achieving a Dice between 0.76 and 0.97 depending on the region to be segmented. The rest of the studies only segment some concrete regions and with unsupervised segmentation, i.e. there is not a quantitative evaluation of the segmentation algorithm. However, the study of Wennman *et al.* uses MRI instead of CT imaging as input, and they include not enough information to make a fair comparison of the performance metrics, since they divide the segmentation performance into different skeletal regions without mentioning the number of slices evaluated in each case, the way to divide each region, etc.

In short, the approach based on DL model, concretely in the adapted U-Net architecture, obtains a satisfactory bone and BM segmentation. However, some limitations must be mentioned. In terms of database, despite utilizing a total of 25,117 samples for bone segmentation, as a 2D segmentation is performed, only 77

volumetric images are considered, all acquired with the same scanner, acquisition protocols and in the same hospital centre. Furthermore, considering that only 12 patients, with 3,813 samples, of the whole dataset are used for testing the model, the test is also importantly limited. Related to BM segmentation, this limitation is even worse, since not only the limitations for bone segmentation are presented, but also masks used for the first training phase are obtained from the pGT creation algorithm which is, in turn, dependent on bone segmentation. Although this algorithm has an acceptable performance, with a Dice (mask) of 0.874, it must be considered that the model is being trained to predict pGT masks, and not GT. Despite the fact that in the second training phase, the model is trained with real GT, the number of samples is very limited for both training and testing, and all of them come from the same 6 and 7 volumes, respectively. Additionally, for both, bone and BM segmentation, the lack of data augmentation processes can be considered as another limitation in the methodology, since it is the main difference with the study by Noguchi *et al.* [79].

Then, there are some limitations related to the methodology. First, due to limited computational resources (GPU-VRAM 6GB), the maximum value for the parameter F is 8, i.e. it is not possible to develop a U-Net network architecture with a higher value of F , not even using the original architecture. The methodological limitation is not only given by the computational resources, but also by the architecture used. Although the U-Net network is used in this thesis since it is demonstrated to be useful, since it handles small datasets better than other complex architectures, and concretely in the bone segmentation [78,79], many other architectures could be used for the same purpose: DeepLab [226], Fully Convolutional Network [227] or variants of the U-Net itself [228–231]. In the case of the optimal BM segmentation model, the U-Net model is selected for the possibility of transfer learning from the bone segmentation model, a factor which limits the possible use of other architectures.

Overall, the bone and BM segmentation models implemented in the present thesis achieves a satisfactory performance, even outstanding, with Dice(mask) of 0.958 and 0.928, respectively. In addition, both depend only on the CT as input, decoupling BM segmentation from adequate previous bone segmentation. Future works lead to improving model performance by experimenting with other architectures, such as nnU-Net as in [112] or other U-Net variants, and including a data augmentation process to increase the robustness of the methodology and the generalization of the segmentation model. Moreover, the study aims to use a higher variety of samples, not only patients from the same scanner, acquisition protocols and centre of acquisition.

5.1.5. Conclusions

For an adequate BM quantification, a precise segmentation is crucial. Although the importance of this region in numerous clinical contexts, there are not many studies which develop an adequate segmentation, limiting its quantification. Nonetheless, there is a high variety of studies which segment the bone instead. Being the bone the union of BM and compact bone, i.e. they are adjacent regions, since there is no BM without compact bone, some studies benefit from the bone segmentation for obtaining the BM mask.

In the present section, basing the approaches on the robust bone segmentation studies found, both bone and BM segmentations are proposed. The potential of DL models against traditional algorithms is demonstrated in both the state of the art and the present study. Despite database and methodological limitations, the resulting models obtain satisfactory segmentations, providing the first step for the quantification of the BM. Moreover, thanks to the bone segmentation model, other clinical contexts can be evaluated such as bone density or weakness. The optimal segmentation methodology, i.e. bone and BM segmentation based on DL models, is implemented in a platform which allows not only applying the automatic segmentation, but also developing a manual revision of the masks. This platform lays the foundation to develop a clinical decision support system, implementing the first step, the segmentation of the interest organ, for a posterior quantification.

5.2. Bone marrow Multiple Myeloma quantification in ^{18}F FDG PET/CT

As mentioned, BM segmentation represents the first step for BM quantification in multiple clinical contexts. In chapter 4. *Materials and methods*, different subcontexts in the MM field are proposed to be quantified. First, a MRD detection using ^{18}F FDG PET/CT imaging acquired for the assessment of the treatment response in patients diagnosed with MM is proposed and implemented. This is followed by the prediction of this same condition from ^{18}F FDG PET/CT imaging acquired for the MM diagnosis. This way, the potential of the quantification methodologies is tested not only for detecting but also for predicting the probability of relapse. Finally, an SMM risk to progression to MM, using ^{18}F FDG PET/CT imaging of the SMM diagnosis, is also quantified. Although this quantification is not related to MRD, it could be useful for experts, since it helps to define the treatment strategies.

5.2.1. Measurable Residual Disease detection

For the MRD detection, the cohort is composed of 150 ^{18}F FDG PET/CT images of the treatment relapse assessment in MM patients with their corresponding visual

assessments by experts in nuclear medicine. In addition, for 148 patients, a BM biopsy evaluation is also available. As seen in Table 13, there is a proportion of PET+/PET- (MRD+/MRD- by PET visual assessment) of 34%/66%, and MFC+/MFC- (MRD+/MRD- by biopsy and its haematological MFC assessment) of 36.5%/63.5%. It must be mentioned that 37.2% of patients obtain discrepant results for PET and MFC evaluation. For this reason, as mentioned in subsection 4.3.1. *Pseudo-ground truth labelling*, not only the complete cohort is used for prediction, but also the portion of PET- patients. This way the potential of the quantification methods to distinguish pattern difficult to see for experts is evaluated. For this cohort of 97 patients in pGT labelling, the proportion of MRD+/MRD- is 30%/70%.

		PET assessment	
		MRD-	MRD+
Biopsy assessment (MFC)	MRD-	68	26
	MRD+	29	25
	X	2	0

Table 13. Confusion matrix for PET visual assessment and MFC assessment of MRD. X: no biopsy available.

For comparing quantification methodologies with [^{18}F]FDG PET visual assessment, performance metrics using MFC as GT are obtained, achieving the following values: sensitivity = 0.463, precision = 0.490, F1_score (MRD-) = 0.712, F1_score (MRD+) = 0.476, F1_score_macro = 0.597 and F1_score_weighted = 0.630.

5.2.1.1. Quantification by Machine Learning models

The first quantification methodology by ML models based on radiomics features extracted from [^{18}F]FDG PET imaging in the BM region, the results shown in Table 14 are obtained. The included metrics are those corresponding to the model with the best F1_macro of all the combinations of group features (ORIG, R, OS and OS_R) and labelling (PET, MFC and pGT). Since there are 11 ML models and 4 BWs for each pair of group and labelling, 1 out of the 44 possibilities is expressed in Table 14. *Annex G. Performance of ML models MRD detection* includes F1_macro and F1_weighted for the 44 combinations of each pair group and labelling. It must be remarked that no significant radiomics features result from the features selection described in subsection 4.3.3.2.1. *Univariate analysis* for both labelling MFC and pGT.

From Table 14, in terms of PET labelling, the outstanding values, i.e. higher than 0.85, for both F1_macro and F1_weighted, should be mentioned. Although both metrics obtain their highest value for the OS model, it could be the result of an

overfitting of the model. For this reason, the results for the ORIG group are considered optimal. For this model, a F1_macro of 0.860 and F1_weighted of 0.872 are obtained. As seen, there is no benefit from reducing the number of features through the univariate analysis. For the MRD detection based on MFC labelling, models obtain worse but acceptable (higher than 0.7) results, with a F1_macro of 0.760 and F1_weighted of 0.775. Finally, from the pGT test, again the ORIG model obtains acceptable values, with a F1_macro of 0.708 and F1_weighted of 0.726.

		S	Pr	F1 (MRD-)	F1 (MRD+)	F1_m	F1_w
PET	ORIG	0.875	0.778	0.897	0.824	0.860	0.872
	RED	0.625	1.000	0.909	0.769	0.839	0.860
	OS	0.750	1.000	0.938	0.857	0.897	0.910
	OS_R	0.750	0.857	0.903	0.800	0.852	0.867
MFC	ORIG	0.750	0.667	0.815	0.706	0.760	0.775
	OS	0.875	0.583	0.750	0.700	0.725	0.732
pGT	ORIG	1.000	0.500	0.750	0.667	0.708	0.726
	OS	0.500	0.667	0.857	0.571	0.714	0.775

Table 14. Performance metrics for the ML model with the best F1_macro in MRD detection. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted. Outstanding values, i.e. above 0.85, are marked in bold.

From these results, conclusions can be drawn. First, F1_score of the negative class (MRD-), i.e. the majoritarian, is higher than F1_score of the positive class (MRD+), since there are not enough samples from this minoritarian class, limiting the learning of the model. Second, PET labelling F1_scores are, in any group, better than the rest of labelling F1_scores, since the radiomics features are directly extracted from the same image used for distinguishing between classes. There is no influence of clinical parameters. Finally, pGT labelling obtains worse results than MFC although both use the same GT. The main reason is the fact that the database is being reduced, from 148 to 97 samples, and that the clearest cases, i.e. those where both labelling methods, MFC and PET, are MRD+, are not included. This implies a worse generalization of the model.

In order to compare the potential of the different ML models, and the way the discretization process affects to models results, Table 15 shows model and BW for the models whose performance is included in Table 14. As seen, not a clear tendency of improvement depending on the BW is observed. However, BW of 0.0625 and 0.25 are repeated 3 times, while 0.5 and 0.125 only appear once. With the exception of MLP2 model, the simpler models (NB and k-NN) find their optimisation with larger BWs (0.5 and 0.25) and the more complex models (GB) with smaller BWs (0.0625). Meanwhile, SVM models (RBF and SIG) do not follow any pattern.

		BW	MOD
PET	ORIG	0.125	RBF
	R	0.0625	GB
	OS	0.25	NB
	OS_R	0.0625	GB
MFC	ORIG	0.25	SIG
	OS	0.5	MLP2
pGT	ORIG	0.0625	SIG
	OS	0.25	k-NN

Table 15. BW and model type for the ML model with the best F1_macro in MRD detection.

In terms of ML models appearance, there are two models twice repeated (GB and SIG), four models which appears only once (RBF, NB, MLP2 and k-NN), and five models which are not the optimal one in any of the combinations (DT, LIN, RF, LR, and MLP1). It must be noted that NB is the simplest implementation model, since features are binarised to its training. Considering that it is the best model for the OS group, it could be due to the overfitting of the model due to oversampled samples after binarization. On the other hand, GB and MLP2 represent some of the most complex ML models implemented. In the case of GB, it is the best model for the reduced groups, indicating its potential with an optimal database. Moreover, this model obtains optimal performance for the smaller BW, highlighting its potential with discretised image composed of a higher amount of intensity levels, i.e. more complex patterns. In the case of MLP2, only the OS group for MFC labelling shows the potential of this architecture, emphasizing the complexity of detecting MFC through PET radiomics features. Contrarily, k-NN represents one of the simplest models, since it is instance-based, and it is the one with a superior performance for the most complex approach, i.e. distinguishing PET images which are MRD- for visual assessment.

Although these tables (Table 14 and Table 15) represent the results for F1_macro optimization, the table for F1_weighted optimization is exactly the same except for the ORIG group in the pGT labelling. In this case, another model with a F1_weighted of 0.736, but a sensitivity of 0.25 is obtained. For this reason, this is not evaluated.

5.2.1.2. Quantification using Convolutional Neural Networks

For the second quantification methodology, the one based on CNNs, Table 16 shows the performance metrics for the combination of parameters, i.e. architecture, F and class weights, with the best F1_macro for each combination of labelling (PET, MFC and pGT), model input (3D, pRGB and MIP) and application of data augmentation

(applied or not applied). This way, only 1 model is shown for the 60 possibilities (4 values of F, 5 architectures and 3 class weights). The rest of metrics can be found in *Annex H. Performance of CNNs in MRD detection*. Only in two cases, the optimal combinations of parameters for F1_weighted optimization differs from the F1_macro ones (3D and MIP, both with data augmentation for pGT). In both cases, the F1_macro is remarkably unacceptable, so they are not shown.

In terms of PET labelling, non-outstanding results are obtained, being acceptable for 3D and pRGB inputs. The most remarkable one is characterised by a 3D input and application of data augmentation, with a F1_macro of 0.817 and F1_weighted of 0.830. For the MFC labelling, the result for the combination pRGB input and data augmentation application highlights for its performance, with a F1_macro of 0.856 and F1_weighted of 0.865, being both outstanding values. In the case of pGT labelling, pRGB input is also the one with the optimal performance, with F1_scores of 0.792 and 0.850 for F1_macro and F1_weighted, respectively. The coincidence of pRGB as the best model adaptation for both, MFC and pGT, is in line with their definition, since pGT is originated from MFC labelling.

		DA	S	Pr	F1 (MRD-)	F1 (MRD+)	F1_m	F1_w
PET	3D	X	0.750	0.667	0.828	0.706	0.767	0.785
		✓	0.875	0.700	0.857	0.778	0.817	0.830
	pRGB	X	0.750	0.667	0.828	0.706	0.767	0.785
		✓	0.750	0.667	0.828	0.706	0.767	0.785
	MIP	X	0.500	0.667	0.813	0.571	0.692	0.729
		✓	0.500	0.667	0.813	0.571	0.692	0.729
MFC	3D	X	0.500	0.800	0.839	0.615	0.727	0.758
		✓	0.500	0.500	0.714	0.500	0.607	0.636
	pRGB	X	0.750	0.545	0.720	0.632	0.676	0.688
		✓	0.875	0.778	0.889	0.824	0.856	0.865
	MIP	X	0.250	1.000	0.824	0.400	0.612	0.670
		✓	0.250	1.000	0.824	0.400	0.612	0.670
pGT	3D	X	0.750	0.600	0.857	0.667	0.762	0.806
		✓	0.500	0.400	0.762	0.444	0.603	0.677
	pRGB	X	0.500	1.000	0.917	0.667	0.792	0.850
		✓	0.500	0.667	0.870	0.571	0.720	0.790
	MIP	X	0.250	0.500	0.833	0.333	0.583	0.700
		✓	0.500	0.500	0.818	0.500	0.659	0.733

Table 16. Performance metrics for CNNs architectures with the best F1_macro in MRD detection. DA: Data Augmentation; S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted. Outstanding values, i.e. above 0.85, are marked in bold.

From the results shown in Table 16, some statements can be made. First, the use of MIP as input of any CNN architecture is pointless in any labelling method. Despite the advantage of its use, since it needs less computational resources, it seems to provide less information than the 3D image. By contrast, the pRGB adaptation of the model for reducing the needed computational resources by generating a convolutional block at the beginning of the architecture, is demonstrated to be useful, obtaining the best performance for 2 out of the 3 possible labelling methods. Second, there is no combination which obtains a better MRD+ F1_score than MRD-F1_score, due to disbalanced database, being even more notorious in the pGT labelling, for the maximum disbalance (30%/70%). Thirdly, although the best performance is obtained for MFC labelling, PET detection is steadier, i.e. it obtains acceptable or at least near to acceptable results for most of the combinations, while MFC obtains acceptable results for one combination and outstanding for the other one, being the rest unacceptable results. Finally, there is no pattern that indicates that the models improve when using data augmentation, in fact, there are 3 cases where it improves, 3 cases where it stays the same and 3 cases where it gets worse. No trend can be determined.

Table 17 shows the combination of parameters for the performance metrics shown in Table 16. In terms of F parameters, it could be seen the abundant appearance of value 4 compared to the rest of values. However, studying a pattern related to the increase or decrease of complexity due to the change in F value while using models of different level of complexity is useless. Related to the CNN architecture, there is a clear tendency towards the simpler models, since 3D_RPET_NET, proposed by Amyar *et al.* [222] in 2019, is the simplest model and it is the best performance architecture in 8 out of the 18 cases. Followed by AlexNet, the next simplest one, and VGG16, which, although has more layers than AlexNet, is simpler than ResNet and DenseNet, since it does not include extra measurements such as residues. Both, AlexNet and VGG16, appear 4 times, meanwhile ResNet appears only twice, and DenseNet does not appear. However, this observation is made over the total of models, which only indicates that the simpler the architecture, the more probabilities to converge. Considering only combinations with acceptable F1_macro, the medium complexity level represented by VGG16 seems to be optimal for 4 out of the 9 parameters combination with acceptable performances, varying the F value from 1 to 4. In addition, AlexNet with $F = 4$, seems to nearly follows VGG16, being the optimal 3 times. ResNet and 3D_RPET_NET appear only once, both with a lower value of F (1 and 2, respectively). Finally, observing the three different possibilities for class weights, A and B seem to be the most adequate ones, since both appear 10 and 7 times for the total of combinations, and 4 and 5 for the ones with acceptable performance metrics, respectively. A remarkable tendency cannot be defined. However, the superiority of both against C is demonstrated.

		DA	Architecture	F	Class weights
PET	3D	X	AlexNet	4	B
		✓	AlexNet	4	A
	pRGB	X	AlexNet	4	B
		✓	VGG16	2	A
	MIP	X	3D_RPET_NET	4	A
		✓	3D_RPET_NET	4	A
MFC	3D	X	VGG16	1	B
		✓	3D_RPET_NET	1	C
	pRGB	X	AlexNet	8	A
		✓	VGG16	4	B
	MIP	X	ResNet	8	B
		✓	3D_RPET_NET	4	B
pGT	3D	X	ResNet	1	B
		✓	3D_RPET_NET	2	A
	pRGB	X	3D_RPET_NET	2	A
		✓	VGG16	4	A
	MIP	X	3D_RPET_NET	8	A
		✓	3D_RPET_NET	1	A

Table 17. Architecture, F value and class weights for the CNN with the best F1_macro in MRD detection. DA: data augmentation. DA column is marked in green for combinations which obtains an acceptable value of F1_macro.

5.2.1.3. Discussion

For MM follow-up, MRD is a biomarker with prognostic ability to identify groups with longer responses and better survival outcomes. BM biopsy combined with ^{18}F FDG PET/CT imaging are the most common techniques to evaluate MRD. However, both provide different information, increasing the discrepancies between their results, and consequently, hesitation in therapeutic decision making. Furthermore, they have their own limitations. Despite the high sensitivity of BM biopsy, this technique extracts a small sample of BM. Considering the heterogeneous and patchy pattern of MRD, this sample can be unrepresentative of the whole BM. For ^{18}F FDG PET, there are many cases which are doubtful for experts in nuclear medicine, due to their low ^{18}F FDG uptake. In this section, a cohort of 150 ^{18}F FDG PET images with their correspondent BM masks is used to develop different quantification methodologies in order to detect MRD, using as GT both visual assessment of ^{18}F FDG PET and MFC evaluation of BM biopsies.

The first quantification methodology, which consists in ML models based on radiomics features extracted from ^{18}F FDG PET, demonstrates the potential of these

features for MRD detection, especially for PET labelling with a F1_macro of 0.860. For both labelling methods based on MFC assessment, i.e. MFC and pGT, the performance of these models is lower but still acceptable, with F1_macro of 0.760 and 0.708, respectively. Although both PET and pGT labelling methods obtain better performance for OS than for the ORIG group, due to the recommendation of not using oversampling since synthesised samples could not really belong to the minority class in real-life [232,233], the possible overfitting of models to these synthesised samples and the similar performance between them, the ORIG group is prioritised over OS, since these models are expected to generalised better.

For the second quantification methodology, which consists in CNNs architectures trained directly with [^{18}F]FDG PET imaging in BM regions, the performance metrics also demonstrate its potential, especially for MFC labelling, with a maximum F1_macro of 0.856. From this result, more complex features than radiomics can be considered to detect patterns that can predict the outcome of the MFC assessment. Considering that pGT labelling is based on MFC labelling, similar performance is observed for this labelling method, with a maximum F1_macro of 0.792. Finally, the most surprising result is the one for PET labelling, since using directly the same image that is visually assessed by experts, a better results than for the rest of labelling methods is expected, while the F1_macro maximum is 0.817. It must be mentioned that, in spite of down-sampling the 3D images resolution due to computational resources limitations, CNN architectures based on 3D inputs demonstrate a great potential.

Comparing both methodologies some statements can be made. First, PET labelling seems to need simpler features to be predicted, since the performance for the methodology based on radiomics features is notably better than the one for CNNs methodology. Meanwhile, both biopsy-based labelling methods, i.e. MFC and pGT, need more complex features for prediction, obtaining better performance for the CNNs methodology. Secondly, although oversampling methods are not recommended for ML models approach, data augmentation is still considered for CNNs approach. The main reason is that synthetised images are created through random operations over the real images, creating only possible images, changing: the position of the patient in the acquisition process (rotation, translation and flip) and the size of the person (zoom). Thirdly, and also related to data augmentation, it must be mentioned again, the impossibility of creating reproducible results for CNNs quantification approach, since data augmentation is applied on-the-fly. This way, models for different labelling methods are not trained exactly the same way making the comparison between them less fair, while ML models are completely reproducible and, consequently, steadier, since databases are stationary. Finally, both methodologies demonstrate to have a great potential to detect MRD, since both

obtain acceptable metrics for all the labelling methods. Finally, comparing quantification methodologies to ^{18}F FDG PET visual assessment to classify images based on MFC evaluation, both achieve a remarkable better performance.

Comparing both methodologies with the state of the art related to MRD detection summarised in Table 4, no studies apart from this thesis are found. Only Takahashi *et al.* [129], compares both images, pre- and post-treatment, with their own metric without considering residual tumour burden. This way, two studies within the framework of this thesis [234,235] and the methodologies just explained represent almost the complete MRD detection based on ^{18}F FDG PET research line. These two first studies are a proof-of-concept for the methodologies included in the present thesis. The performance of the models implemented in these two studies obtained F1_score values higher than the ones of the mentioned methodologies only for the case of the oversampled database. Considering that these studies have more patients created by the SMOTE technique than real patients for the MRD+ class (8 real and 23 synthetic), a high overfitting of the models is assumed. For the original PET-labelled database, much higher values are obtained in the present methodologies, with an F1_score of 0.860, compared to a maximum of 0.620 and 0.400 in [234] and [235], respectively. In addition, viable models are obtained for the MFC labelling, which was not feasible in previous studies.

The novelty and great performance of these methodologies is not exempted from some limitations. In terms of database, there are multiple limitations. The number of samples used in this thesis is very limited, even more for a CNN approach, since the maximum number of samples is 150 for PET labelling while the minimum is 97 for pGT labelling. In addition, a clear disbalance between classes is observed, especially for the pGT labelling with a proportion of 30%/70% between MRD+/MRD- classes. However, the main limitation is that the classification label corresponds to the ^{18}F FDG PET visual assessment or the biopsy assessed by MFC, but there is no real GT analysed as patient follow-up. In terms of methodology, limitations are divided depending on the methodology. For methodology based on ML models, the number of extracted features is limited to 125, since not wavelets or other filters are applied to the images previous to feature extraction. In addition, features selection is based on univariate analysis, overlooking the potential information based on features combinations. Regarding ML models, separation between classes threshold is fixed to 0.5, while other possibilities could obtain better performances considering the disbalanced between classes. On the other hand, regarding the methodology based on CNNs architectures, limitations are mainly based on computational resources. First, 3D inputs resolution is down-sampled due to the impossibility of training with large files. Second, only architectures with a maximum of 8 filters or channels in their first feature map (F) are possible to be

implemented. Third, all CNN models are not reproducible due to the use of data augmentation on-the-fly. Finally, there are some limitations not related to resources, such as parameters selection fixed, or MIP input limited to one of the three planes.

From these limitations, multiple future works can be considered. For the whole quantification, the obtention of a wider and more balanced database based on follow-up GT, in order to obtain models with a better generalization of the problem. In terms of feature obtention, the extraction of a larger number of radiomics features to find features that suit the problem better, rather than using even more features for training. Furthermore, using other techniques for dimensionality reduction instead of univariate analysis, such as Principal Component Analysis (PCA) [236,237], for avoiding the elimination of essential information due to features combination. Then, for ML models, optimising the separation between classes threshold based on the training set, for compensating class imbalance. For CNNs inputs, considering no computational limitations, using the 3D image in its original resolution, and using three MIP images, one for each plane (axial, coronal and sagittal). Finally, implementing a CNN parameters optimization, varying learning rate, type of optimiser, loss function, etc.

In addition to future works based on limitations, two new future lines are proposed. First, based on the mentioned study [235], where only biopsy sites are analysed to detect MRD, and a conference proceeding [238] also developed in the framework of the present thesis, where a DL model is used for classifying skeleton regions into 8 different anatomical structures, using the mentioned quantification methodologies to detect MRD drawing from different BM regions. This way, more important regions in the MRD pattern could be determined, helping clinicians to focus their visual assessment. Second, combining both methodologies by using radiomics features as CNNs inputs. Finally, both can be combined, since the extraction of radiomics features individually for each anatomical structure provides a huge increase in the dimensionality of the database, which can be used to train CNNs.

As detailed, the quantification methodologies implemented in the present thesis not only demonstrate their potential for MRD detection, but also lay the foundation for multiple future works related to this field. In this way, by improving the models with higher sample training and lower class imbalance, it will be possible to detect the most visually doubtful cases. From this reflexion, it must be noted that models based on pGT labelling represent the most important contribution to MRD research, since the possibility of distinguishing MRD+ cases (assured by the high sensitivity of the biopsy), which cannot be differentiated by visual assessment of [^{18}F]FDG PET, is demonstrated. Moreover, these models obtain acceptable performance being trained with a limited database, in terms of number of samples (97 against 148 for MFC) and disbalance (30%/70% of MRD+/MRD-), demonstrating the potential of this

methodology. This thesis is the beginning of a line of research to generate clinical decision support systems for nuclear medicine experts, especially in the most dubious cases.

5.2.1.4. Conclusions

In spite of the importance of MRD detection for the follow-up of MM patients, the current state of the art is scarce. The potential of BM biopsy and ^{18}F FDG PET visual assessment for MRD detection has been previously demonstrated. However, there is still a high percentage of discrepancies between both results, implying hesitation in therapeutic decision making. In the present thesis two different BM quantification methodologies are proposed to classify patients between MRD+ and MRD- based on ^{18}F FDG PET visual assessment and MFC evaluation of BM biopsy.

From the obtained results, the potential of both quantification methodologies to detect MRD based on any labelling method is demonstrated. In addition, an experiment based only on patients classified as MRD- by ^{18}F FDG PET visual assessment using as GT the MFC is implemented. Its acceptable metrics, demonstrate the potential of both methodologies to be implemented as a clinical decision support system for helping nuclear medicine experts with the most difficult cases, reducing discrepancies between both assessments and the number of patients wrongly diagnosed as MRD- by the PET visual assessment.

5.2.2. Measurable Residual Disease prediction

For MRD prediction, i.e. trying to predict which patients are more probable to suffer MRD by using their MM diagnostic images, a cohort of 69 ^{18}F FDG PET/CT images is used, of which 68 has also a BM biopsy available. Table 18 shows both labelling methods, i.e. ^{18}F FDG PET visual assessment and MFC evaluation of BM biopsy, for all the patients. Both labelling methods are disbalanced, with a proportion of 29%/71% and 29.4%/70.6% for MRD+/MRD- in PET and MFC labelling, respectively. As seen, 26 out of the 68 with both assessments, represent discrepancies between them, achieving a total of 38.2%.

		PET assessment	
		MRD-	MRD+
Biopsy assessment (MFC)	MRD-	35	13
	MRD+	13	7
	X	1	0

Table 18. Confusion matrix for PET visual assessment and MFC assessment of MRD. X: no biopsy available.

Due to the limited number of samples for the present database, only the quantification based on ML models trained with radiomics features is implemented. For this quantification, the whole process is applied, including univariate analysis to reduce the number of features. It must be taken into account that this quantification is performed to predict MRD, this way, classification is based on the visual assessment and MFC evaluation of the moment of treatment response assessment, i.e. between $[^{18}\text{F}]$ FDG PET image from which radiomics features are extracted and the moment of labelling obtention there is a period of time comprising the treatment and some time to observe the response to this treatment.

5.2.2.1. Quantification results

Table 19 shows performance metrics of ML models quantification applied to MRD prediction. Only the best model for each pair of labelling method and feature group is included in the table, i.e. 1 out of 44. The rest of them can be consulted in *Annex I. Performance of ML models in MRD prediction*. As seen, there is no reduced groups from these features, i.e. there are no features with significant differences between MRD+ and MRD- classes. For the ORIG and OS groups in PET labelling, performance metrics are exactly the same, with a F1_macro of 0.762 and F1_weighted of 0.8. In the case of MFC labelling, the results are slightly better for OS than for the ORIG group, with a F1_macro of 0.762 against 0.688. However, it could be due to overfitting of the model.

		S	Pr	F1(MRD-)	F1(MRD+)	F1_m	F1_w
PET	ORIG	0.670	0.670	0.857	0.667	0.762	0.800
	OS	0.670	0.670	0.857	0.667	0.762	0.800
MFC	ORIG	0.330	1.000	0.875	0.500	0.688	0.763
	OS	0.670	0.670	0.857	0.667	0.762	0.800

Table 19. Performance metrics for the ML model with the best F1_macro in MRD prediction. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted. Outstanding values, i.e. above 0.85, are marked in bold.

Since only 15% is used for testing, only 10 samples are tested with only 3 MRD+ cases, for both PET and MFC labelling. This way, models can obtain acceptable metrics classifying only 2 out of the 3 MRD+ cases correctly. This can be easily observed since any value of F1_score for MRD+ is higher than 0.667, i.e. there is not enough MRD+ samples in any labelling to obtain a generalised model. However, since the majority of cases are correctly classified in spite of the limited database, the possible potential of this quantification methodology to predict MRD from diagnostic images is demonstrated.

Table 20 includes the BW and ML model for the metrics shown in Table 19. As it can be seen, the optimal results are repeated multiple times, due to the reduced number of samples of the test set. From all of these combinations, it should be noted that the simplest models, i.e. NB, DT and k-NN are the most repeated ones, since there are not enough samples to train correctly a more complex model. In terms of BW, no tendency can be observed.

		BW	MOD
PET	ORIG	0.25, 0.125	SIG, NB
	OS	0.125 ,0.125, 0.0625	DT, k-NN, NB
MFC	ORIG	0.5, 0.125, 0.0625	GB, k-NN, k-NN
	OS	0.5	NB

Table 20. Bin Width (BW) and model type for the ML model with the best F1_macro in MRD detection.

5.2.2.2. Discussion

As mentioned above, MRD is an essential biomarker for MM follow-up. Not only its detection during response evaluation could be useful in clinical practice, but also its prediction since MM diagnosis moment. This way, physicians could take other therapeutical decisions, such as less spaced follow-ups. For this reason, ML models are trained to predict both MFC and PET labelling from radiomics features extracted from [^{18}F]FDG PET imaging.

Although a limited database is used, the potential of ML models can be observed, with acceptable values for both F1_macro and F1_weighted, being their maxima of 0.762 and 0.8, respectively, in both labelling methods. However, in both cases, a poor F1_score for the minoritarian class against an outstanding F1_score for the majoritarian is observed, due to the notable database disbalance. Avoiding the use of oversampling methodologies due to their disadvantages, PET obtains better performance than MFC, since radiomics features are directly extracted from the image, like in MRD detection.

Since mentioned above, studies about MRD are very limited. Consequently, none has been found about MRD prediction from MM diagnostic images. The most similar studies are focused on survival prediction [122–125]. From these, the most comparable studies are Jamet *et al.* [123] and Ni *et al.* [124]. Both studies are focused on survival prediction without mentioning MRD, but survival is partially related to this condition. Since both are evaluating survival instead of MRD and with different performance metrics, no direct comparison can be made. In general terms, Jamet *et al.* obtain a robust model, while Ni *et al.* obtain an acceptable one. Considering the

number of samples for both, 139 and 98 respectively, compared to the present one, 69, the potential of the present quantification methodology is demonstrated.

Despite the novelty and the possible usefulness in clinical practice of this research line, this quantification is characterised by several limitations. Equally to the MRD detection, the main limitation is the database, since it is limited, imbalanced, discrepant between classification methods, i.e. PET and MFC, and based on these evaluations instead of the follow-up of the patient. In terms of methodology, the same limitations as in MRD detection related to the number of radiomics features extracted, features selection method and classes separation threshold are observed. However, due to the complexity of the prediction compared to the detection, another observed limitation is the lack of clinical features for predicting, like in other studies [124,125].

Focusing on these limitations, some future works can be proposed. First, database ampliation, using follow-up as GT and including an equivalent proportion of both classes. If acceptable results are obtained with the actual database, robust ones are expected with a representative one. Second, including clinical features for MRD prediction, based on the diagnosis parameters: percentage of plasma cell in BM biopsy, level of paraprotein in serum or urine, haematological conditions such as hypercalcemia or anaemia, number of focal lesions, etc.

5.2.2.3. Conclusions

Although many studies focus on survival outcomes, none of them try to predict MRD. Prediction models can provide patterns for differentiating patients with a higher probability of suffering MRD, helping experts in the therapeutic decision making. In the present subsection, ML models to predict both, PET and MFC labelling are proposed. Due to their acceptable results, in spite of limited database, the potential of this quantification methodology is demonstrated, laying the basis for developing a research line focuses on this prediction. This way, clinical decision support systems for helping experts to detect patients with more probability of relapse can be implemented.

5.2.3. Risk assessment in Multiple Myeloma Smouldering

Also related to MM field, not regarding MRD, the quantification of ^{18}F FDG PET imaging is implemented for evaluating risk of progression to MM in SMM patients. Table 21 shows the number of patients classified as any combination of the three different levels of risk and the two different assessment methods, i.e. IMWG and PETHEMA. The cohort is composed of 45 ^{18}F FDG PET images, with 19 cases discrepant between assessment methods, i.e. 42%. For IMWG labelling, there is a

proportion of low/intermediate/high level of risk of 11.1%/35.6%/53.3%, while for PETHEMA it is 33.3%/17.8%/48.9%.

	Class	PETHEMA		
		Low	Intermediate	High
IMWG	Low	3	2	0
	Intermediate	9	4	3
	High	3	2	19

Table 21. Confusion matrix for IMWG assessment and PETHEMA assessment of risk of progression to MM in SMM patients.

Due to the limited database, not only in terms of total samples, but also in terms of samples per class, only the first part of univariate analysis is applied for this quantification. This way, every test from Figure 53 before multi-testing correction by means of the Benjamini-Hochberg procedure is applied individually for both assessment methods.

5.2.3.1. Quantification by univariate analysis

In Table 22, the radiomics feature with significant differences between classes for BW of 0.5SUV, i.e. with a p-value lower than 0.05, for the IMWG labelling is included with the p-values between each pair of classes. This feature represents the proportion of low grey levels zones in the image. Considering p-values for pair of classes, a significant difference of this feature between low- and high-risk groups is observed.

Group	Variable	Bin Width = 0.5			
		P-value	Low-Inter	Low-High	Inter-High
GLSZM	LowGrayLevelZoneEmphasis	0.030	0.570	0.029	1.000

Table 22. P-value lower than 0.05 for ANOVA or Kruskal-Wallis tests for univariate analysis to assess the risk of progression to MM in patients with SMM in IMWG labelling. P-value corrected by Benjamini-Hochberg correction is not shown, as its values is higher than 0.05. P-values of post hoc analysis for each pair of classes are also included, shown in bold those below 0.05.

For a better understanding of the feature, Figure 102 shows the boxplot for different classes. As it can be observed, the lower the risk, the higher the value of this feature. It indicates that a higher proportion of zones of a lower intensity grey level are directly related to less risk of progression.

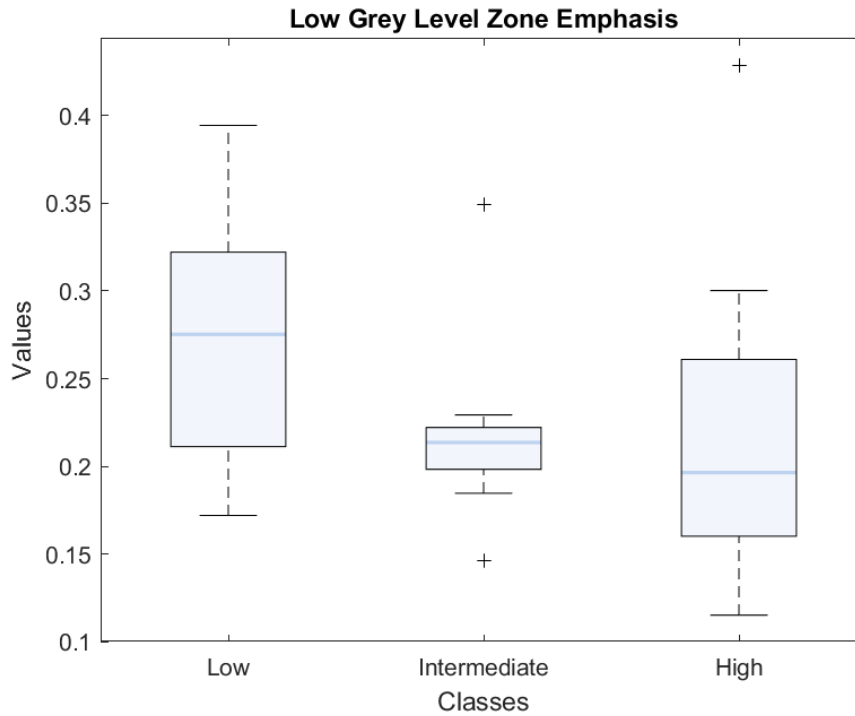


Figure 102. Boxplot for Low Grey Level Zone Emphasis for IMWG labelling of SMM patients, differing between low-, intermediate- and high-risk of progression to MM.

For PETHEMA labelling, the same information shown in Table 22 for IMWG is divided into two different tables: Table 23 for p-values of ANOVA and Kruskal-Wallis tests for the three classes and Table 24 for the p-values of each pair of classes through post hoc tests. First, it should be noted that the first three variables represent the same information, the minimum of the $[^{18}\text{F}]\text{FDG}$ PET in the BM region. This way, they obtain the same p-value even for the different bin widths. For the rest of variables included in both tables, Imc2, MCC and LowGrayLevelZoneEmphasis are directly related with homogeneity, i.e. the higher the feature value, the higher the image homogeneity. Finally, DependenceNonUniformityNormalized is inversely related to the image homogeneity, while SmallAreaLowGrayLevelEmphasis is directly related to heterogeneity. Imc2 obtains a better p-value the smaller the BW is, just the opposite of SmallAreaLowGrayLevelEmphasis. The remaining features seems to work better with concrete BWs. MCC and DependenceNonUniformityNormalized only obtains significant differences between classes for BW of 0.25SUV, while LowGrayLevelZoneEmphasis for the two biggest BWs, being better for 0.25SUV. From all of the features, it should be mentioned that the ones which obtain a better p-value are LowGrayLevelZoneEmphasis with 0.013 and Imc2 with 0.019 for BWs of 0.25SUV and 0.0625SUV, respectively.

Group	Feature	Bin Width			
		0.5	0.25	0.125	0.0625
Descriptors	Mask_interpolated_Minimum	0.045	0.045	0.045	0.045
	Mask_resegmented_Minimum	0.045	0.045	0.045	0.045
First order	Minimum	0.045	0.045	0.045	0.045
GLCM	Imc2	0.386	0.035	0.021	0.019
	MCC	0.190	0.045	0.057	0.210
GLDM	DependenceNonUniformityNormalized	0.236	0.044	0.467	0.413
GLSZM	LowGrayLevelZoneEmphasis	0.030	0.013	0.166	0.268
	SmallAreaLowGrayLevelEmphasis	0.035	0.135	0.206	0.388

Table 23. P-values lower than 0.05 (shown in bold type), for any of the bin widths, obtained in ANOVA or Kruskal-Wallis tests for univariate analysis to assess the risk of progression to MM in patients with SMM in PETHEMA labelling. P-values corrected by Benjamini-Hochberg correction are not shown, as none obtain values which means significant differences between classes.

Significant differences between some pair of classes are observed in three radiomics features (Table 24), achieving none of them significant differences between all the pairs. While LowGrayLevelZoneEmphasis and SmallAreaLowGrayLevelEmphasis find significant differences between intermediate- and high-risk classes, DependenceNonUniformityNormalized find differences between low- and high-risk classes. No features with significant differences between low- and intermediate-risk classes are found.

Gr.	Variable	Label pair	Bin Width			
			0.5	0.25	0.125	0.0625
Image and ROI descriptors	Mask_interpolated Minimum	Low-Inter	1.000	1.000	1.000	1.000
		Low-High	0.453	0.453	0.453	0.453
		Inter-High	0.067	0.067	0.067	0.067
	Mask_resegmented Minimum	Low-Inter	1.000	1.000	1.000	1.000
		Low-High	0.453	0.453	0.453	0.453
		Inter-High	0.067	0.067	0.067	0.067
First order	Minimum	Low-Inter	1.000	1.000	1.000	1.000
		Low-High	0.453	0.453	0.453	0.453
		Inter-High	0.067	0.067	0.067	0.067
GLCM	Imc2	Low-Inter		0.091	0.068	0.066
		Low-High		0.108	0.109	0.120
		Inter-High		1.000	1.000	0.820
	MCC	Low-Inter		0.129		
		Low-High		0.095		
		Inter-High		1.000		
GLDM	Dependence	Low-Inter		0.096		
	NonUniformity	Low-High		0.039		
	Normalized	Inter-High		1.000		
GLSZM	LowGrayLevel ZoneEmphasis	Low-Inter	1.000	1.000		
		Low-High	0.135	0.197		
		Inter-High	0.041	0.011		
	SmallArea LowGrayLevelEmphasis	Low-Inter	1.000			
		Low-High	0.149			
		Inter-High	0.038			

Table 24. P-values of post hoc analysis for each pair of classes for variables shown in Table 23. P-values below 0.05 are shown in bold. Gr: group of features.

For these three features, boxplots differing between low-, intermediate- and high-risk groups are shown in Figure 103. Figure 103a shows not only the difference between low- and high-risk groups, but also the similarity between intermediate- and high-risk ones. This feature represents the heterogeneity within dependences, being higher for low-risk patients. On the other hand, Figure 103bc represents similar information, since both acquire higher values when the image contains more zones of low grey level values. For both, high-risk group is distinguishable from the intermediate one.

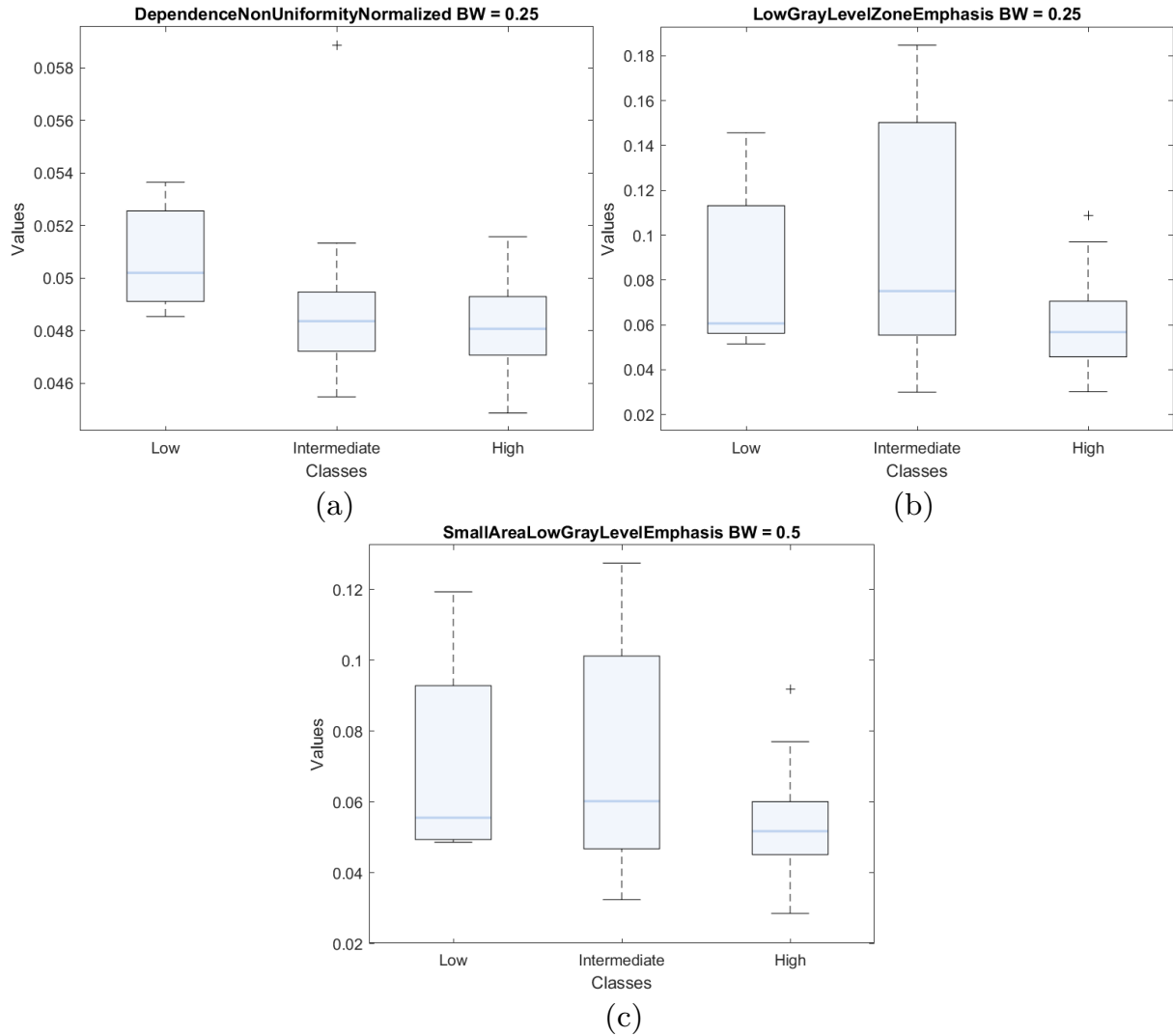


Figure 103. Boxplot for PETHEMA labelling of SMM patients, differing between low-, intermediate- and high-risk of progression to MM, for radiomics features: (a) DependenceNonUniformityNormalized, (b) LowGrayLevelZoneEmphasis and (c) SmallAreaLowGrayLevelEmphasis.

Although only variables with a p-value lower than 0.05 are included in this section, univariate analysis is performed for all of the extracted radiomics features. P-values of the rest of variables can be consulted in *Annex J. Univariate analysis results for risk assessment in SMM*.

5.2.3.2. Discussion

SMM is an asymptomatic condition, which usually progress to MM. However, time to progression varies from some months to several years, becoming therapeutic decisions a challenge for haematological experts. In this way, an attempt is made to separate patients into groups defined by the likelihood of progressing more rapidly. In the current section, univariate analyses of radiomics variables extracted from $[^{18}\text{F}]\text{FDG}$ PET imaging in BM region are developed to find potential prognostic

biomarkers of this rapid progression, differentiating patients into three groups: low, intermediate and high risk of progression to MM.

In spite of the limited database, with only 45 patients, some tendencies are observed. First, for IMWG labelling, LowGrayLevelZoneEmphasis obtains a p-value of 0.030 for all the classes, with significant differences between low- and high-risk groups (p-value = 0.029). Considering the low representation of low-risk class (5 out of 45), it must be noted the potential of this feature as biomarker, which represents a higher number of low-grey-level zones for low-risk group. Nonetheless, the lack of variables capable of distinguishing low- and intermediate-risk and intermediate- and high-risk groups should be mentioned, being the distinguishing of IMWG labelling classes impossible.

On the other hand, for PETHEMA labelling, a more balanced database, more features with significant differences between all groups are observed (8 out of 125). However, only three of them obtain significant differences between pairs of classes. Despite expecting more differences between low- and high- risk groups, only DependenceNonUniformityNormalized shows significant differences between these two classes, with a p-value of 0.039. Contrarily to expected, this feature, which is directly correlated to heterogeneity, is higher for low-risk group than for the rest. Meanwhile classes which are expected to be more similar, intermediate- and high-risk, are differentiable by two radiomics features, LowGrayLevelZoneEmphasis, the same one than for IMWG low- and high-risk differentiation, and SmallAreaLowGrayLevelEmphasis, with the best p-values of 0.011 and 0.038, respectively. Both represent a lower number of small low-grey-level zones for high-risk group than for the rest of classes. In clinical terms, the lower the risk, the more abundant are the low-grey-level zones. This result is reasonable, since for progression to MM, where the BM shows a patchy and heterogeneous pattern, the pattern of BM in SMM patients is expected to be more similar the higher the risk of progression. Thanks to these features or biomarkers, PETHEMA classification can be quantified for differentiating between classes, except for low- and intermediate-risk groups.

Summarizing all the results, independent on the labelling method, no features for low- and intermediate-risk differentiation are found. This is an expected result since there is a high discrepancy between labelling methods, especially for patients of intermediate-risk based on IMWG and low-risk based on PETHEMA. In addition, the most represented group in both labelling methods is the high-risk since it represents nearly 50% of cases. This way it can be concluded that, in spite of the limited database, some potential biomarkers are found, demonstrating the potential of radiomics features to predict risk of progression to MM. These results are in line with the study of Ripani *et al.* [118], which demonstrates the potential of radiomics features to predict the progression to MM, using the same number of samples.

However, their study only evaluates the risk to progression or not progression, without a differentiation between levels of risk. No other studies of this field are found.

As shown in the results, the main limitation of this quantification is the database. Although some potential biomarkers are found, the database is extremely limited, with unrepresentative number of samples of each class. Furthermore, a real GT based on patients where time to progression is measured would be the optimal way to define classes, while IMWG and PETHEMA labelling methods are based on clinical variables of the moment of the diagnosis of SMM.

Due to the limited database, no other quantification methods described in the present thesis are applied. However, the first future works are increasing the amount and representation of samples of each class to compose a database big enough to performed both quantification based on ML models and quantification based on CNN architectures. Following this improvement of the database, only patients with a [^{18}F]FDG PET image of the moment of SMM diagnosis and a diagnosed progression to MM can be used for conforming a more realistic database.

5.2.3.3. Conclusions

Since SMM condition tends to progress to MM, haematological experts try to define the risk of progression. This way, a stricter follow-up can be prescribed for a patient with a higher risk, improving timing for MM diagnosis and reducing unnecessary test for low-risk patients. Despite the limited database, the present thesis set the scene for a research line focused on finding quantitative image biomarkers to help clinicians to identify patients with higher risk, in order to personalise the follow-up and therapeutic decisions of the patient. There is still a long way to go to generate accurate decision support systems in this field, but the trend can be observed thanks to this proof of concept.

5.3. Bone marrow Non-Hodgkin's Lymphoma quantification in [^{18}F]FDG PET/CT

For NHL quantification, the cohort is composed of 273 [^{18}F]FDG PET/CT images of diagnosis with their corresponding visual assessment by experts in nuclear medicine. In addition, for 234 patients, a BM biopsy evaluation of BM status is also available. From these 273 NHL patients, 153 belong to the FL group, 116 to the DLBCL group and 4 are patients with transformed NHL lymphoma, i.e. patients whose lymphoma evolves from one subtype to another. As it can be seen in Table 25, there is a proportion of BMI-/BMI+/rBM based on PET visual assessment of 70.7%/22%/7.3%, while based on MFC it is of 56.8%/28.6%/14.5%. These same

proportions for the FL group are 66.6%/28.1%/5.2% and 51.8%/38.8%/9.3%, respectively, and 75%/14.6%/10.3% and 65.9%/13.2%/20.9% respectively for the DLBCL group. It must be mentioned that 37.6% (88 out of 234), 40.3% (56 out of 139) and 31.9% (29 out of 91) of NHL, FL and DLBCL patients with both, MFC and PET assessment, belong to different classes. Following the rules described in subsection 4.3.1. *Pseudo-ground truth labelling*, the pGT labelling results in: 146 BMI-, 91 BMI+ and 36 rBM, for the NHL group; 70 BMI-, 70 BMI+ and 13 rBM for the FL group; and 75 BMI-, 20 BMI+ and 21 rBM for the DLBCL group.

		PET assessment								
		NHL			FL			DLBCL		
	Class	BMI-	BMI+	rBM	BMI-	BMI+	rBM	BMI-	BMI+	rBM
Biopsy assessment (MFC)	BMI-	110	16	7	58	12	2	51	4	5
	BMI+	30	31	6	23	25	6	6	6	0
	rBM	24	5	5	11	2	0	11	3	5
	X	29	8	2	10	4	0	19	4	2

Table 25. Confusion matrix comparing PET and MFC labelling. X: no MFC test available.

For comparing quantification methods with visual assessment, taking MFC and pGT labelling as GT, the performance metrics are obtained for the PET visual assessment summarised in Table 26. As it can be seen, taking MFC as GT, no acceptable values (greater than 0.7) are obtained for any of the metrics with the exception of majoritarian F1_score. As pGT is obtained partially from PET labelling, these results are generally better than those obtained with MFC labelling. The superiority of F1_score in all its variants (F1_macro, F1_macro* and F1_weighted) for the more aggressive subtype group must be noted, i.e. DLBCL, compared to the indolent and the complete cohort, i.e. FL and NHL.

	MFC			pGT		
	NHL	FL	DLBCL	NHL	FL	DLBCL
Sensitivity	0.479	0.423	0.538	0.551	0.509	0.600
Precision	0.515	0.424	0.571	0.617	0.526	0.673
F1(BMI-)	0.741	0.806	0.797	0.783	0.763	0.722
F1(BMI+)	0.521	0.463	0.480	0.696	0.705	0.815
F1(rBM)	0.192	0.000	0.345	0.192	0.000	0.610
F1_m	0.485	0.423	0.541	0.557	0.489	0.619
F1_m*	0.623	0.597	0.656	0.687	0.654	0.716
F1_w	0.598	0.597	0.661	0.666	0.664	0.699

Table 26. Sensitivity, precision, F1(BMI-): F1_score for BMI- class; F1(BMI+): F1_score for BMI+ class; F1(rBM): F1_score for rBM class; F1_m: F1_score_macro, calculated as the mean F1_score of the three previous metrics; F1_m*: F1_score_macro*, calculated as the mean

F1_score considered as binary problem for any class; F1_w: F1_score_weighted; all of them using PET labelling as predicted and MFC or pGT labelling as GT.

Since differences between NHL subtypes are observed in the state of the art, due to their aggressiveness and their adherence to $[^{18}\text{F}]\text{FDG}$, quantification methodologies are individually applied to each subgroup, i.e. not only for the whole cohort of NHL, but also for the subcohorts of FL and DLBCL. However, this differentiation is only applied for ML models, due to the limited database for CNN quantification.

5.3.1. Quantification by radiomics extraction

For the first quantification methodology, Table 27, Table 28 and Table 29 show the performance metrics for each cohort: NHL, FL and DLBCL, respectively. This way, for each combination of labelling method (PET, MFC and pGT), group of features (ORIG, R, OS and OS_R) and cohort (NHL, FL and DLBCL) only the ML model and BW with the best performance, measured by both F1_macro* and F1_weighted, are shown, i.e. 1 out of 44. The performance metrics for all the models are included in *Annex K. BM quantification in NHL patients by ML models*. The lack of significant features should be noted for NHL and DLBCL cohorts in MFC labelling. In addition, significant features for the FL cohort in MFC labelling are only found for BWs of 0.0625SUV and 0.125SUV, and for the DLBCL cohort in pGT labelling only for 0.0625SUV.

As seen in Table 27, the NHL cohort obtains better performance for PET labelling than the rest of labelling methods, being unacceptable for MFC and pGT. Although the OS group obtains the highest F1_macro*, as mentioned above for MRD detection, the ORIG group is prioritised for avoiding overfitting. For ORIG group, the NHL cohort obtains a maximum F1_macro* of 0.803 and F1_weighted of 0.818 by selecting the first metric as priority. On the other hand, a better F1_weighted is obtained by prioritizing it, with a F1_macro* of 0.786 and F1_weighted of 0.834. Nonetheless, the performance for predicting the minoritarian class, i.e. rBM, is represented by a F1_score of 0.333, demonstrating the lack of generalization of the model for this class. For MFC labelling, F1_macro* maximum is 0.659 with a F1_weighted of 0.644, remarking the unacceptable generalization for BMI+ class with a F1_score of 0.444 for this class. Finally, pGT labelling improves slightly MFC performance, increasing BMI+ generalization by worsening the other two classes classification. From all the possible combinations, only the model of pGT labelling for F1_weighted maximization improves with the features reduction.

		S	Pr	F1 (BMI-)	F1 (BMI+)	F1 (rBM)	F1_m	F1_m*	F1_w
F1_macro*	PET	ORIG	0.681	0.833	0.900	0.778	0.500	0.726	0.818
		RED	0.672	0.629	0.852	0.842	0.222	0.639	0.792
		OS	0.692	0.917	0.918	0.824	0.500	0.747	0.843
		OS_R	0.758	0.686	0.873	0.700	0.571	0.715	0.803
	MFC	ORIG	0.533	0.222	0.727	0.444	0.500	0.557	0.644
		OS	0.633	0.565	0.744	0.167	0.667	0.526	0.637
	pGT	ORIG	0.551	0.545	0.682	0.615	0.333	0.543	0.672
		RED	0.551	0.545	0.682	0.615	0.333	0.543	0.672
		OS	0.548	X	0.846	0.720	0.000	0.522	0.686
		OS_R	0.489	0.454	0.698	0.710	0.000	0.469	0.682
F1_weighted	PET	ORIG	0.655	0.736	0.918	0.800	0.333	0.684	0.834
		RED	0.483	0.428	0.722	0.593	0.000	0.438	0.663
	MFC	ORIG	0.483	0.428	0.722	0.593	0.000	0.438	0.663
		OS	0.450	0.435	0.743	0.538	0.000	0.427	0.667
	pGT	ORIG	0.496	0.449	0.723	0.690	0.000	0.471	0.679
		RED	0.513	0.449	0.698	0.727	0.000	0.475	0.686

Table 27. Performance metrics for the ML model with the best F1_macro (upper) and F1_weighted (lower) for BM assessment in NHL patients. Green cells represent coincidence between the model with the best F1_macro* and the best F1_weighted, excluding the metrics from F1_w set to avoid redundant information. Outstanding values, i.e. above 0.85, are marked in bold. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted; X: indeterminate number resulting from dividing by 0.

For the FL cohort, as seen in Table 28, equally to NHL cohort, the best performances are obtained for PET labelling, achieving a maximum F1_macro* of 0.836, with a F1_weighted of 0.768, for OS group and a F1_macro* of 0.812, with a F1_weighted of 0.962, for RED group. In spite of the high metrics for RED group, the F1_score for rBM class is 0, i.e. none of the rBM test samples are correctly classified, being the other two perfectly classified (BMI-) and outstandingly classified (BMI+). For F1_weighted optimization, acceptable metrics but lower than for F1_macro* optimization are obtained, with a maximum of F1_macro* of 0.774 and F1_weighted of 0.909. None of the remaining labelling methods, i.e. MFC and pGT, for any of the F1_score metrics optimization obtains acceptable performances, apart from pGT labelling for OS_R group, with a F1_macro* of 0.740 and F1_weighted of 0.745, with more balanced F1_scores of each class 0.737, 0.720 and 0.5 for BMI-, BMI+ and rBM, respectively.

		S	Pr	F1 (BMI-)	F1 (BMI+)	F1 (rBM)	F1_m	F1_m*	F1_w	
F1_macro*	PET	ORIG	0.533	0.558	0.903	0.727	0.000	0.543	0.710	0.815
		RED	0.667	X	1.000	0.923	0.000	0.641	0.812	0.962
		OS	0.855	0.824	0.815	0.667	1.000	0.827	0.836	0.768
		OS_R	0.845	0.687	0.897	0.727	0.500	0.708	0.797	0.835
	MFC	ORIG	0.485	0.640	0.500	0.421	0.667	0.529	0.605	0.532
		RED	0.455	0.398	0.421	0.471	0.333	0.408	0.546	0.517
		OS	0.451	0.451	0.727	0.625	0.000	0.451	0.620	0.683
		OS_R	0.443	0.596	0.455	0.353	0.667	0.492	0.569	0.476
	pGT	ORIG	0.591	0.574	0.667	0.667	0.333	0.556	0.673	0.694
		RED	0.591	0.556	0.696	0.632	0.333	0.554	0.674	0.696
		OS	0.485	0.465	0.727	0.696	0.000	0.474	0.641	0.707
		OS_R	0.651	0.673	0.737	0.720	0.500	0.652	0.740	0.745
F1_weighted	PET	OS	0.533	0.533	0.933	0.667	0.000	0.533	0.714	0.842
		OS_R	0.611	X	0.968	0.833	0.000	0.600	0.774	0.909
	MFC	ORIG	0.439	X	0.720	0.533	0.000	0.418	0.589	0.631
		RED	0.443	0.394	0.500	0.375	0.333	0.403	0.543	0.521
		OS_R	0.348	X	0.545	0.444	0.000	0.330	0.504	0.515
	pGT	ORIG	0.515	X	0.727	0.750	0.000	0.492	0.659	0.727
		RED	0.485	0.464	0.750	0.667	0.000	0.472	0.640	0.706

Table 28. Performance metrics for the ML model with the best F1_macro (upper) and F1_weighted (lower) for BM assessment in FL patients. Green cells represent coincidence between the model with the best F1_macro* and the best F1_weighted, excluding the metrics from F1_w set to avoid redundant information. Outstanding values, i.e. above 0.85, are marked in bold. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted; X: indeterminate number resulting from dividing by 0.

Finally, for DLBCL patients, as shown in Table 29, all the labelling methods obtain acceptable F1_macro* values for some model. In the case of PET, the OS_R group obtains the best F1_macro* with a value of 0.759 and a F1_weighted of 0.807. However, its F1_score for BMI-, the majoritarian class, hides the poor classification of both minoritarian classes. For this reason, ORIG and RED groups, both with the same metrics, are considered as better performances, with a F1_macro* of 0.757, slightly lower than the OS_R group, and F1_weighted of 0.741. Similar hiding of poor minoritarian classification is observed for F1_weighted models. For MFC labelling, the ORIG group model must be remarked since it obtains an acceptable F1_macro* with a value of 0.705 for F1_macro* optimization, while the OS group for optimizing F1_weighted obtains a F1_macro* of 0.7 and F1_weighted of 0.728. Finally, for pGT and F1_macro* optimization, only the ORIG features group

obtains acceptable values of F1_macro* and F1_weighted, 0.744 and 0.732, respectively.

		S	Pr	F1 (BMI-)	F1 (BMI+)	F1 (rBM)	F1_m	F1_m*	F1_w
F1_macro*	PET	ORIG	0.611	0.938	0.897	0.500	0.667	0.688	0.741
		RED	0.611	0.938	0.897	0.500	0.667	0.688	0.741
		OS	0.585	0.600	0.857	0.500	0.500	0.619	0.690
		OS_R	0.611	0.789	0.929	0.500	0.500	0.643	0.807
	MFC	ORIG	0.611	0.730	0.750	0.667	0.444	0.620	0.699
		OS	0.574	0.711	0.842	0.667	0.333	0.614	0.724
	pGT	ORIG	0.636	0.812	0.833	0.500	0.667	0.667	0.744
		RED	0.576	0.709	0.727	0.500	0.500	0.576	0.668
		OS	0.576	0.709	0.727	0.500	0.500	0.576	0.668
		OS_R	0.656	0.736	0.700	0.500	0.600	0.600	0.666
F1_weighted	PET	RED	0.585	0.730	0.889	0.500	0.400	0.596	0.754
		OS	0.444	X	0.897	0.400	0.000	0.432	0.691
	MFC	ORIG	0.519	X	0.889	0.000	0.500	0.463	0.748
		OS	0.574	0.711	0.842	0.400	0.500	0.581	0.728
	pGT	RED	0.626	0.452	0.667	0.571	0.444	0.561	0.657
		OS	0.495	0.667	0.783	0.500	0.286	0.523	0.654

Table 29. Performance metrics for the ML model with the best F1_macro (upper) and F1_weighted (lower) for BM assessment in DLBCL patients. Green cells represent coincidence between the model with the best F1_macro* and the best F1_weighted, excluding the metrics from F1_w set to avoid redundant information. Outstanding values, i.e. above 0.85, are marked in bold. S: sensitivity; Pr: precision; F1_m: F1_macro; F1_w: F1_weighted; X: indeterminate number resulting from dividing by 0.

Due to the large imbalance in the data that generates a high masking of minority classes in the F1_weighted result, F1_macro* is used as the metric to optimise. In the same way, in order to avoid model overfitting, models without oversampling processes are considered suitable. Avoiding oversampling, the model is not trained with samples that might not exist in real life and the current recommendations are followed [232,233]. Consequently, these rules are followed to simplify comparisons between presented models. This way, when maximums are mentioned, they are related to not oversampled groups, i.e. only ORIG and RED features groups.

Comparing labelling methods, since the features studied are directly extracted from the [¹⁸F]FDG PET image, without any clinical aspect, the best F1_macro* results are obtained for the PET labelling for the three cohorts with a maximum of 0.803, 0.812 and 0.757 for NHL, FL and DLBCL, respectively. In line with this result, pGT labelling obtains better performance, 0.662, 0.674 and 0.744, respectively, than MFC,

with 0.652, 0.605 and 0.705, since it is partially defined by PET labelling. Comparing different cohorts of patients, the best performance of PET labelling corresponds to FL, i.e., the indolent subgroup, with a $F1\_macro^*$ of 0.812, followed by the NHL group with a $F1\_macro^*$ of 0.803, which includes both FL and DLBCL. However, for MFC and pGT labelling, the DLBCL cohort, i.e. the aggressive subtype of lymphoma, obtains better performances than FL or NHL, with a metric of 0.705 and 0.744, respectively.

In terms of each class performance, the best $F1\_score$ for the majoritarian class, i.e. BMI- class, is obtained in the three cohorts of patients for the PET labelling, with values of 0.9, 1 and 0.897, respectively. Similarly, the BMI+ class for NHL and FL cohorts obtains an optimal classification in PET labelling, with $F1\_scores$ of 0.842 and 0.923. On the other hand, the optimal classification of this BM status for the DLBCL cohort is obtained for MFC labelling, with a $F1\_score$ of 0.667. Finally, for the most minoritarian class, i.e. rBM, it must be remarked that for PET labelling ML models are incapable of classifying this class in the FL cohort, with a $F1\_score$ of 0 for the optimal $F1\_macro^*$ model, while a maximum of 0.667 for $F1\_score$ of this class is obtained for the DLBCL cohort. For MFC labelling, is just the opposite, with a $F1\_score$ for rBM class of 0.667 and 0.444 for FL and DLBCL cohorts, respectively.

Table 30 shows the model and BW corresponding to the models whose performance are shown in Table 27, Table 28 and Table 29. In terms of ML models for $F1\_macro^*$ optimization without oversampling methods, there is a notable difference between some models that result in the best performance in multiple cases, such as LR, RF, MLP1, MLP2 or GB (best models in 6, 4, 2, 2 and 2 cases, respectively), compared to others that obtain the best result in only one (DT, NB or k-NN) or zero (any SVM model) combinations. Consequently, more complex models generalise better for BM status classification, in both FL and DLBCL cohorts. Finally, the smallest BW obtains more often the best performance being the number of times: 8, 4, 3 and 4, for 0.0625, 0.125, 0.25 and 0.5, respectively. Nevertheless, not a clear tendency of improvement depending on the BW is observed.

		NHL		FL		DLBCL			
		BW	MOD	BW	MOD	BW	MOD		
F1_macro*	PET	ORIG	0.5	LR	0.125/0.0625	RF/LR	0.0625	GB	
		R	0.125	LR	0.5	RF	0.25/0.0625	GB/MLP2	
		OS	0.25	GB	0.25	RF	0.125	MLP1	
		OS_R	0.125	MLP1	0.125	RF	0.0625	MLP1	
	MFC	ORIG	0.0625	NB	0.0625	MLP1	0.25/0.125/0.0625	LR	
		R			0.0625	RF			
		OS	0.0625	SIG	0.125	MLP1	0.5/0.125	RF	
		OS_R			0.0625	RF			
	pGT	ORIG	0.25	MLP1	0.5	DT	0.5	RF	
		R	0.0625	MLP2	0.125	LR	0.0625	LR/k-NN	
		OS	0.5	LIN	0.0625	MLP1	0.125	DT	
		OS_R	0.5	GB	0.125	RF	0.0625	RF	
	F1_weighted	PET	ORIG	0.25	LR	0.125/0.0625	RF/LR	0.0625	GB
			R	0.125	LR	0.5	RF	0.0625	MLP1
			OS	0.25	GB	0.0625	DT	0.25	GB
			OS_R	0.125	MLP1	0.25	RF	0.0625	MLP1
MFC		ORIG	0.125	MLP2	0.5	MLP2	0.25	MLP1	
		R			0.0625	DT			
		OS	0.5	LR	0.125	MLP1	0.25	DT	
		OS_R			0.0625	LR			
pGT		ORIG	0.125	GB	0.25	MLP2	0.5	RF	
		R	0.25	GB	0.25	RF	0.0625	RBF	
		OS	0.5	LIN	0.0625	MLP1	0.0625	GB	
		OS_R	0.5	GB	0.125	RF	0.0625	RF	

Table 30. Bin Width (BW) and model type for the ML model with the best F1_macro (upper) and F1_weighted (lower) for BM assessment in NHL, FL and DLBCL. Grey cells represent the lack of those groups, i.e., no significant features after reduction. Green cells represent coincidence between the model with the best F1_macro* and the best F1_weighted.

5.3.2. Quantification using Deep or Machine Learning models

Due to disbalanced database and especially limited for minoritarian classes (with a minimum of 8 samples for rBM in the worst case), the second quantification methodology, i.e. the one based on CNN architectures, is applied only in the NHL cohort, not for FL and DLBCL. For this experiment Table 31 shows the performance metrics for the combination of parameters, i.e. architecture, F and class weights, with the best F1_macro* for each combination of labelling (PET, MFC and pGT), model input (3D, pRGB and MIP) and application of data augmentation (applied or not applied). This way, same as for MRD quantification, only 1 model is shown

for the 60 possibilities (4 values of F, 5 architectures and 3 class weights). The rest of metrics can be found in *Annex L. BM quantification in NHL patients by CNNs*.

As it can be seen in Table 31, there are not outstanding values for none F1_macro* or F1_weighted in any combination of input and labelling. In terms of PET labelling, the best F1_macro* is obtained for MIP without the data augmentation process, with a performance defined by a F1_macro* of 0.732 and F1_weighted of 0.731. On the other hand, the best F1_weighted is obtained for 3D without data augmentation input, with F1_macro* of 0.720 and F1_weighted of 0.790. However, the F1_score for the rBM class is lower than 0.5 in this last case (which means that fails more than randomized). Taking this metric also into account, the best model is the one defined previously for the best F1_macro*, since the F1_score of the rBM class increases at the expense of losing precision in BMI- and BMI+ classification. For MFC and pGT labelling, as for ML models performance, pGT obtains generally better results, since this labelling method is partially defined by the PET result. Although pGT is generally better, the maximum F1_macro* of MFC labelling is higher than the one for pGT, 0.681 against 0.666. However, none of their models obtain a rBM F1_score higher than 0.4, denoting the impossibility of detecting this class defined by MFC test by using images as the only input. As a result, no good models for generalizing this problem based on these two last labelling methods are obtained.

			S	Pr	F1 (BMI-)	F1 (BMI+)	F1 (rBM)	F1_m	F1_m*	F1_w	
F1_macro*	PET	3D	X	0.581	X	0.931	0.700	0.000	0.544	0.720	0.790
			✓	0.659	0.745	0.847	0.500	0.545	0.631	0.724	0.741
		pRGB	X	0.556	0.747	0.792	0.500	0.400	0.564	0.678	0.690
			✓	0.599	0.726	0.900	0.500	0.400	0.600	0.727	0.764
		MIP	X	0.626	0.764	0.807	0.526	0.667	0.667	0.732	0.731
			✓	0.519	X	0.889	0.667	0.000	0.519	0.677	0.753
	MFC	3D	X	0.476	0.512	0.700	0.381	0.364	0.482	0.617	0.537
			✓	0.494	0.542	0.714	0.421	0.364	0.500	0.630	0.557
		pRGB	X	0.447	0.458	0.714	0.421	0.182	0.439	0.594	0.521
			✓	0.561	0.766	0.732	0.667	0.286	0.561	0.681	0.638
		MIP	X	0.465	0.480	0.634	0.571	0.200	0.469	0.603	0.542
			✓	0.547	0.547	0.667	0.571	0.400	0.546	0.662	0.586
	pGT	3D	X	0.526	X	0.809	0.667	0.000	0.492	0.661	0.662
			✓	0.532	X	0.792	0.750	0.000	0.514	0.666	0.681
		pRGB	X	0.555	0.553	0.636	0.545	0.375	0.519	0.636	0.573
			✓	0.500	0.503	0.727	0.593	0.182	0.501	0.644	0.614
		MIP	X	0.571	0.590	0.571	0.632	0.444	0.549	0.657	0.576
			✓	0.445	0.454	0.636	0.538	0.167	0.447	0.594	0.545
F1_weighted	PET	3D	✓	0.519	X	0.889	0.667	0.000	0.519	0.677	0.753
		pRGB	X	0.556	0.747	0.792	0.500	0.400	0.564	0.654	0.707
		MIP	X	0.578	0.608	0.897	0.588	0.286	0.590	0.725	0.770
	MFC	3D	X	0.478	X	0.711	0.600	0.000	0.437	0.595	0.539
			✓	0.475	0.629	0.708	0.500	0.250	0.486	0.607	0.569
	pGT	pRGB	X	0.481	X	0.744	0.588	0.000	0.444	0.616	0.600
		MIP	X	0.506	0.593	0.720	0.560	0.286	0.522	0.645	0.612
			✓	0.452	X	0.772	0.500	0.000	0.424	0.588	0.585

Table 31. Performance metrics for CNNs architectures with the best F1_macro* (upper) and F1_weighted (lower) in BM status classification. DA: Data Augmentation; S: sensitivity; Pr: precision; F1_m: F1_macro; F1_m*: F1_macro* F1_w: F1_weighted; X: indeterminate number resulting from dividing by 0. Outstanding values, i.e. above 0.85, are marked in bold. Green cells represent coincidence between the model with the best F1_m* and the best F1_w.

From the results shown in Table 31, some observations are made. First, no input shows a superiority over the rest, since similar results are obtained for 3D, pRGB and MIP. Consequently, pRGB or MIP can be good solutions for computational resources limitations. Second, a tendency to improve the performance of the CNNs is observed by using data augmentation (7 out of 9), especially for 3D and pRGB databases. Third, only two models obtain an F1_score for rBM higher than 0.5, both

for PET labelling. Consequently, rBM class defined by MFC or pGT labelling methods is impossible to be detected. Due to the limited number of samples from this class, this is an expected result. By contrast, the majoritarian class obtains F1_scores acceptable or even outstanding in most of the models, due to the abundance of samples. The BMI+ class represents an intermediate performance, with F1_scores near to an acceptable value. Finally, as for ML model quantification, the disbalance between classes generates a high masking of minority classes in the F1_weighted, being a useless metric for measuring performance.

Table 32 shows the combination of parameters for the performance metrics shown in Table 31. In terms of F parameters, a balanced repetition of all values is observed, so no trend is concluded. Related to the CNN architecture, there is a clear tendency to the medium complexity models, since AlexNet and VGG16 are the most repeated, six times each. Meanwhile, 3D_RPET_NET, the simplest, and ResNet, a bit more complex are only three times repeated, and DenseNet is not even in the table of best performance models. Taking only models with an acceptable F1_macro*, VGG16 outstands over the rest (4 out of 5), suggesting that the BM status detection in NHL patients represents a medium complexity problem and no better performances are obtained due to the limited database. Finally, in terms of class weights, A seems to be the most adequate, since it appears 10 times, while B and C appear only 4. This demonstrates the uselessness of these class weights formula for imbalanced databases.

		DA	Architecture	F	Class weights	
F1_macro*	PET	3D	X	VGG16	4	A
			✓	AlexNet	8	B
		pRGB	X	VGG16	8	B
			✓	VGG16	1	C
		MIP	X	VGG16	4	B
			✓	VGG16	1	A
	MFC	3D	X	AlexNet	1	A
			✓	AlexNet	1	A
		pRGB	X	VGG16	4	A
			✓	RPET_NET	2	C
		MIP	X	ResNet	4	A
			✓	AlexNet	2	B
	pGT	3D	X	ResNet	2	A
			✓	AlexNet	4	A
		pRGB	X	AlexNet	8	C
			✓	RPET_NET	2	A
		MIP	X	ResNet	2	A
			✓	RPET_NET	8	C
F1_weighted	PET	3D	✓	AlexNet	4	B
		pRGB	X	AlexNet	2	A
		MIP	X	VGG16	4	A
	MFC	3D	X	ResNet	8	C
			✓	DenseNet	4	B
	pGT	pRGB	X	AlexNet	4	A
		MIP	X	AlexNet	2	A
			✓	AlexNet	8	A

Table 32. Architecture, F value and class weights for the CNN with the best F1_macro in NHL patients for BM status classification. DA: data augmentation. DA column is marked in green for combinations which obtains an acceptable value of F1_macro.

5.3.3. Discussion

Between 20% and 40% of NHL patients suffer from BMI, a condition which directly affects NHL staging at its diagnosis. An adequate detection can be essential in prognosis and therapeutic decisions [46,47]. As for MRD detection, the combination of BM biopsy evaluated by MFC and [^{18}F]FDG PET visual assessment is the most common technique to detect BMI [47]. However, the heterogeneous BM pattern caused by BMI is similar to the one observed in patients which suffer from

inflammatory or infectious states, due to the overproduction of immune cells. For this reason, an adequate identification of BM status, distinguishing between BMI-, BMI+ and rBM, can be useful for helping experts in clinical decision making. In this section, a cohort of 273 ^{18}F FDG PET/CT images of NHL diagnosis is used to performed two different quantification methodologies, one based on ML models and the other based on CNN architectures. This way, both radiomics features extracted from ^{18}F FDG PET imaging and the image itself are used as input to differentiate the three possible BM status, defined by three different labelling methods: ^{18}F FDG PET visual assessment, MFC evaluation of the BM biopsy and pGT defined from the two previous ones.

The first quantitative BM methodology based on extracting radiomics features from ^{18}F FDG PET imaging denotes the potential of these features and ML models for distinguishing BM status in NHL patients. Excluding models for features groups obtained with oversampling methods, the optimal model for PET labelling classification achieves F1_macro* values of 0.803, 0.812 and 0.757 for NHL, FL and DLBCL, respectively. From these results the potential to detect different status in both subtypes of NHL through radiomics features is demonstrated. Furthermore, due to the superior performance of the FL model over the DLBCL model, it could be deduced that it is easier to differentiate these states through PET for the indolent subtype than for the aggressive subtype. However, considering that the DLBCL database is more limited, this cannot be stated with certainty. By contrast, MFC labelling seems to be easier to predict for DLBCL than for FL subtype with a medium performance for NHL, with F1_macro* of 0.705, 0.605 and 0.659, respectively, suggesting that due to the higher aggressiveness of the subtype, MFC assessment yields a lower number of false negatives, i.e. biopsy samples are more likely to be representative of the entire BM. As another demonstration of this result, F1_score of the majoritarian class, i.e. BMI-, in MFC labelling is acceptable for DLBCL patients, while not even near to acceptable for FL. Finally, pGT obtains generally better performances than MFC following the same tendency, with F1_macro* of 0.668, 0.674 and 0.744 for NHL, FL and DLBCL, respectively.

On the other hand, the second quantification methodology, i.e. the one based on CNNs, results in acceptable metrics only for PET labelling with a F1_macro* of 0.732. MFC and pGT maximum F1_macro* are 0.681 and 0.666. In addition, the maximum F1_score for the rBM class is 0.6, being most of them lower than 0.5. Comparing architectures, a clear superiority of VGG16 over the rest of them is demonstrated, implying a medium level of complexity of the classification problem. All these results indicate that the database used is not large enough, or at least the number of samples of each minoritarian class, to be able to create complex models capable of differentiating between these classes. However, the potential of this

quantification method is demonstrated, since although the limited database and no subtypes cohorts are used, results are near to acceptable. Comparing both quantification methodologies demonstrates that ML models based on radiomics features are the optimal solution to distinguish these classes, at least with the current database. Nonetheless, it should be noted the superiority of MFC labelling classification with CNNs, since it obtains better results than ML models. Finally, comparing them with visual assessment summarised in Table 26, both obtain a slightly better performance with F1_macro* for the NHL cohort and MFC labelling of 0.659 and 0.681 for ML models and CNNs approach, respectively, against the 0.623 of visual assessment.

Taking into account the studies included in Table 6, the majority of them are limited to SUV metrics and ratios between them and no classification methodologies. Only two of them [136,140] use radiomics features to detect BMI+ and differentiate it from BMI-. However, their methodology is also limited to a VOI of a maximum SUV, preventing the whole BM evaluation. In addition, no studies which include the rBM status are identified. On the other hand, in line with what was observed in the studies included in the reviews of Pakos *et al.* [130], Chen *et al.* [131] and, Almainani *et al.* [132], in aggressive subtypes, BMI+ is easier to visualise, without any type of quantification (Table 26), as well as rBM. Nonetheless, in the ML models, contrary to the studies without quantification, the indolent subgroup, i.e. FL, obtains better results than the aggressive one, i.e. DLBCL, with F1_macro* of 0.812 versus 0.757. It is interesting to note that BMI+ status is better identified in patients with a more aggressive type than in those with an indolent type with MFC labelling, with F1_scores for BMI+ class of 0.667 versus 0.471, while for PET labelling the behaviour is the opposite, with F1_scores of 0.923 for the indolent type versus 0.5 for the aggressive one.

Limitations of this study must be separated into both quantification methodologies. However, there is a common limitation for both: the database. As for MRD, labelling method represents the main limitation. Both, MFC and PET assessments are labelling approaches used in the clinical practice to diagnose the patients. However, considering the discrepancies between both labels for the three groups of patients: 37.6%, 40.29% and 31.86% for NHL, FL and DLBCL, respectively, a real GT based on the follow-up of the patients, especially for rBM differentiation, would be a better and more realistic GT for this experiment. Furthermore, the number of patients of any cohort, i.e. NHL, FL and DLBCL, only allows to develop proof-of-concept ML models, since it is not enough cases to state their validity for being included in clinical practice, being even worse for CNNs approach. Moreover, all of the cohorts are clearly disbalanced between number of cases for any class, especially for rBM, since there are 20, 34 and 36 classified as rBM for PET, MFC, and pGT labelling,

respectively, for the 273 patients in the NHL group; 8, 13, and 13, out of 153 FL patients; and 12, 19 and 21 out of 116 DLBCL patients. Due to this imbalance, models are incapable of generalising adequately. As seen in both quantification methodologies, the models encounter many difficulties in classifying the minoritarian class, i.e. rBM, with almost all models having a F1_score value for this class under 0.5. Although attempts to control this problem through the use of oversampling are made, it remains a very small number of example patients for the creation of new data and is insufficient to obtain a generalised model. In addition, as mentioned above, oversampling methods are expected to improve the model by overfitting it, and not by a better generalization. Finally, and same as for segmentation, all images are acquired from the same scanner, acquisition protocols and centre, limiting the generalizability of the developed models. In terms of quantification methodologies, limitations are exactly the same as for MRD detection since the quantification methods are equally applied. For ML models quantification, the limitations are number of extracted features, features selection method, and fixed separation between classes thresholds. Meanwhile, for CNNs quantification, limitations are focused on computational resources, such as the 3D inputs resolution, limitations of the F parameter, irreproducibility of trainings due to data augmentation on-the-fly, and the fixed selection of CNNs parameters along with only using one MIP as input, instead of MIPs from the three planes.

From these limitations, multiple future works can be considered, similar to those proposed for MRD detection due to the similarities in database and applied methodology. For database limitation, the creation of a wider and more balanced database based on follow-up GT instead of PET and MFC assessment, in order to generalise better the ML models and CNNs to the problem. Ideally, increasing both NHL subtypes cohorts, i.e. FL and DLBCL, enough to train CNNs architectures focused on each subtype, since the potential of dividing models per each subtype is demonstrated through ML models. Additionally, and considering the great variety of subtypes which conform NHL disease, obtaining more cohorts for more subtypes of NHL, for developing new methods adapted to these new subtypes. In terms of methodology, future works are exactly the same as for MRD detection. For feature obtention, looking for features which suit better the problem by extracting a larger number of radiomics features, and by using dimensionality reduction techniques such as PCA for avoiding information based on features combination. Then, for ML models, in order to compensate the imbalanced database, optimising the threshold for class division based on the training set. For CNNs inputs, apart from computational limitations, using the original resolution of 3D images and using a MIP image for each plane, i.e. axial, coronal and sagittal. Finally, a parameter optimization for CNNs by varying learning rate, type of optimiser, loss function, etc.

Finally, using radiomics features as inputs for CNNs architectures, in order to find a more suitable model to this classification problem.

In addition to future works based on limitations, other research lines, considering the premise that for rBM bladder uptake of $[^{18}\text{F}]\text{FDG}$ increases due to immune cells overproduction being even higher than the reference organ, i.e. liver, explained in the subsection 1.2.2.4. *Reactive bone marrow*, are proposed. Using segmentation of both organs, i.e. bladder and liver, along with BM, a quantification methodology based on more clinical factors for BM status in NHL patients can be developed. This way, not only BM uptake or functionality is considered for defining BM status, but also functionality of bladder and liver. For this future work, a conference article [239] where both liver and bladder are segmented through U-Net architectures, and the uptake of $[^{18}\text{F}]\text{FDG}$ of three organs, BM, liver and bladder, is used to train AlexNet models by using a MIP image. However, a more adequate investigation is proposed, by extracting radiomics individually for each organ for ML models quantification and by using a greater variety of architectures and 3D inputs for CNN quantification. Moreover, considering radiomics features individually extracted for each organ, training CNN architectures with these several features as input.

As described, and despite database limitations, the quantification methodologies implemented in the present thesis, especially ML models based on radiomics features, not only demonstrate their potential for defining BM status in NHL patients, but are also the basis for multiple future works related to this field. In this way, by improving the database in terms of extension and imbalance, it will be possible to get a better differentiation between classes, especially the least obvious, i.e. rBM. Once these future works are developed in order to stop being only a proof of concept, quantification methodologies could be included as part of a clinical decision support system for helping nuclear medicine and haematology experts managing BM status for NHL diagnosis.

5.3.4. Conclusions

For NHL, visual interpretation of $[^{18}\text{F}]\text{FDG}$ PET/CT imaging is performed in clinical practice, not only for main lesions assessment, but also for BMI evaluation and its differentiation from rBM status, due to its importance in the diagnosis of NHL patients and the therapeutic consequences of this BM status. In this thesis, two quantification methodologies for BM status differentiation between three classes, BMI-, BMI+ and rBM, are proposed.

From the results of the ML models quantification methodology, the possible usefulness and potential of this method is demonstrated with a maximum F1_macro^* of 0.803, 0.812 and 0.757 for the NHL, FL and DLBCL cohorts,

respectively, for PET labelling. This potential is also observed for MFC labelling, with poorer results of 0.659, 0.605 and 0.705, respectively. Nevertheless, as this problem is more complex since radiomics features are extracted from PET image, the potential for detection MFC results should be remarked. In this way, the models can be used for helping clinicians in this diagnosis, because the performance of the models exceeds the visual assessment. On the other hand, for CNNs approach a tendency to be potentially adapted to the classification model is observed. However, due to the limited database, actual trainings are only a proof of concept, being limited to using the whole NHL cohort, not dividing into subtypes. The continuation of this research line towards more robust models by using a higher representativeness of the different NHL subtypes would allow for better staging by improving the identification of BM status at diagnosis.

6. Conclusions and future works

In this chapter, the rejection or confirmation of the research hypotheses proposed in section 2.1. *Research hypotheses* of this doctoral thesis is discussed based on the results presented in chapter 5. *Results*. Finally, the main contributions of this thesis and possible future research directions are described.

6.1. Discussion of research hypotheses

H1. Quantitative analysis of bone marrow, using [^{18}F]FDG PET image processing methodologies, allows to quantify the differences in uptake patterns for the study of bone marrow involvement or affectation. The integration of these methodologies in a platform supports the clinical decision-making process in the evaluation of bone marrow.

Hypothesis partially confirmed.

The assessment of BM affectation for various conditions, such as evaluating tumour burden in MM, assessing residual tumour burden in MRD, or evaluating BM infiltration or inflammatory/infectious status in NHL, is performed using a combination of two techniques in clinical practice: haematological evaluation of a BM biopsy and visual assessment of a [^{18}F]FDG PET/CT image. Nevertheless, both techniques identify different information.

On the one hand, the biopsy identifies tumour burden in a small sample of the BM. Despite its high sensitivity, the significant likelihood that the sample is not representative of the entire BM limits the information it can provide, making its combination with [^{18}F]FDG PET/CT imaging necessary for diagnostic confirmation. On the other hand, [^{18}F]FDG PET/CT imaging, while valuable as a complementary test, is subjectively evaluated by a nuclear medicine expert, making it an observer-dependent technique. For this reason, despite the advantages of combining both techniques, there are many cases where the two tests provide discrepant information.

Although medical image processing techniques are increasingly being developed to create objective quantification methodologies that support clinicians in decision-making, particularly in oncology, there are only few examples related to the clinical contexts addressed in this thesis. Moreover, most of the examples found are based on simple methodologies or a partial assessment of the BM, returning to the limitation of lacking an adequate quantification to represent the entire BM.

The quantification methodologies developed for this thesis enable the application of artificial intelligence, a rising tool, for the objective evaluation of [^{18}F]FDG PET

imaging, aiming to provide a diagnostic support tool for nuclear medicine experts, confirming the first part of the hypothesis. In this way, it is expected to contribute, especially in the most challenging cases, to the identification of distinctive patterns of BM pathological conditions, thereby reducing not only false negatives but also the discrepancy between the two diagnostic techniques.

Although the presented results demonstrate the potential of these methodologies, especially considering the limitations of the databases in the different clinical contexts in terms of both extension and representativeness of all possible BM states, integrating these models into a platform that can be directly used in current clinical practice poses a bioethical conflict due to the level of generalisation of the resulting models.

For this reason, studies with more representative and extensive datasets are needed to achieve better model generalisation, enabling the demonstrated potential of this thesis to be directed towards its direct application as a decision-making support tool. With this in mind, the quantification methodologies, despite their ease of integration into the platform, are not developed, thus making the second part of the hypothesis impossible to confirm.

H2. Automatic segmentation models from CT images enable the generation of bone marrow masks to perform a subsequent quantification and/or evaluation of its condition, providing a less time-consuming alternative compared to manual segmentation.

Hypothesis confirmed.

Despite the several object segmentation techniques available and their widespread application in medical imaging, the automatic or semi-automatic segmentation of the BM is typically overlooked due to its complexity. Only a few studies based on unsupervised methodologies are identified attempting this segmentation, without any quantitative evaluation of their performance. Thus, the segmentation of this organ is limited to exclusively manual segmentation, which, due to its extensive presence in the body and its challenging differentiation from compact bone in certain regions, makes this process extremely tedious and time-consuming. For the manual review of the BM masks obtained for this study, an average of 6 to 12 hours is required to segment a case of approximately 300 slices.

This thesis develops two automatic BM segmentation techniques and, as an intermediate step, bone segmentation. The presented results demonstrate the precision of these approaches, particularly highlighting the segmentation based on DL over traditional image processing algorithms, being this integrated into the implemented segmentation platform.

Despite the significant time required to train the multiple combinations of models implemented for this thesis, once the model is integrated into the platform, the total time for complete processing, from the conversion of DICOM files from the CT to the generation of a volumetric mask for both regions, bone and BM, (including stretcher removal, PET volumetrics creation from DICOMs, etc.), is approximately between 2 and 5 minutes, depending on the computational power of the computer used. Furthermore, it is worth noting that an efficient bone segmentation is also achieved, which, although not the primary focus of this thesis, can be used in multiple clinical contexts, such as the evaluation of weakness or fractures in conditions like osteoporosis, or in the assessment of bone metastases.

Finally, the implementation of this segmentation in a platform that also allows for manual review of the masks, including multiple functions to facilitate this tedious task, along with the visualization of the results, provides an effective tool to lay the basis for other potential studies.

H3. [^{18}F]FDG PET images used in the assessment of treatment response in patients with Multiple Myeloma can be quantified to detect Measurable Residual Disease through analysis of bone marrow.

Hypothesis partially confirmed.

As mentioned earlier, the assessment of the BM status in current clinical practice is performed through the combination of two different techniques: BM biopsy and visual evaluation of the [^{18}F]FDG PET image. For this reason, the evaluation is limited by the potential discrepancies between both tests. Regarding MM, it is a disease characterised by the uncontrolled growth of monoclonal plasma cells in the BM, leading to clear affectation of this organ. Thanks to medical advances in recent decades, an increasing number of patients achieve a CR after treatment. However, those who do not reach this response maintain a residual tumour burden in the BM that is patchy and heterogeneous, known as MRD, presenting a threat to the recovery of the patient.

Despite presenting a higher uptake pattern than that observed in patients with a CR, in some cases, due to the high sensitivity of biopsy evaluation, it can be confirmed that there are doubtful cases where nuclear medicine experts are unable to distinguish this involvement. For this reason, this thesis implements BM quantification methodologies based on artificial intelligence, aiming to develop tools capable of distinguishing patterns or potential image-based biomarkers related to the BM status as an aid to diagnosis.

Despite the innovation of applying these techniques in the present clinical context and their demonstrated potential, the proposed hypothesis is considered partially

confirmed. That is, although the possibility of quantifying the heterogeneous patterns of this type of involvement is confirmed, even in doubtful cases (i.e. patients which are visually MRD- through ^{18}F FDG PET assessment), the algorithms are trained with a limited database, meaning that the implemented models cannot be considered generalisable unless better results are achieved. This is particularly important as assuming these models to be fully reliable could have consequences on clinical decisions and, consequently, on the lives of the patients.

H4. Progression to Measurable Residual Disease can be predicted by quantifying ^{18}F FDG uptake in bone marrow on Multiple Myeloma diagnostic imaging.

Hypothesis partially confirmed.

Similarly, the application of these methodologies to other clinical context of MM also demonstrates high potential despite the limited databases. In the case of MRD prediction from ^{18}F FDG PET/CT diagnostic images, it is important to note the difficulty of this prediction, as it does not involve, as in the previous case, a classification based on current information. Instead, it aims to predict the highest likelihood of experiencing MRD based on diagnostic images, without further patient information, either haematological or regarding treatment. For this reason, a perfect classification of the cases is not expected from this quantification, yet the results obtained clearly demonstrate potential. Due to the expected complexity of this prediction, which yields acceptable, but not sufficiently satisfactory results, the hypothesis is considered partially confirmed.

H5. Potential biomarkers of risk for progression to Multiple Myeloma can be identified by quantitative analysis of diagnostic ^{18}F FDG PET images of Smouldering Multiple Myeloma.

Hypothesis partially confirmed.

In the case of assessing the risk of progression to MM in patients with SMM, it is important to mention the inability to implement multivariate quantification methodologies, as the database does not meet the necessary criteria to produce a result that could be considered valid. However, the univariate study of radiomics features demonstrates their potential as possible biomarkers of progression risk. Despite having found potential biomarkers to differentiate low- and high- risk, and intermediate- and high- risk groups, no biomarkers are found to differentiate low- and intermediate- risk groups. For this reason, the hypothesis is considered partially confirmed. Nevertheless, among these markers, it is worth highlighting their great

utility for current clinical practice, as they can help avoid overtreatment or prescribe more intensive monitoring of patients, thus achieving more personalised medicine.

In this way, it can be confirmed that this thesis, due to its novelty in this field, lays the foundation for future clinical decision support systems based on images that are already routinely taken in clinical practice. Furthermore, thanks to the three implemented subcontexts in MM field, it can guide the physician throughout the entire patient follow-up, from an asymptomatic MM, through its symptomatic state, to a possible impairment caused by residual tumour load.

H6. [^{18}F]FDG PET images used in the Non-Hodgkin's Lymphoma diagnosis can be quantified to detect bone marrow infiltration or reaction to inflammation or infection.

Hypothesis partially confirmed.

NHL is a lymphoid neoplasm that encompasses multiple subtypes, with the most common being FL and DLBCL. These subtypes differ primarily in their aggressiveness, with FL being indolent and DLBCL being one of the most aggressive subtypes. However, both share BMI in 20-40% of cases. Although not the primary lesion, this infiltration has significant therapeutic and prognostic consequences.

For its assessment, the combination of techniques previously mentioned is used. However, the heterogeneous uptake observed through [^{18}F]FDG PET/CT diagnostic images in patients with BMI is similar to that seen in patients undergoing an inflammatory or infectious process, where the BM is reactive, as it is responsible for the overproduction of immune cells to control the process affecting the patient.

The quantification methodologies implemented in this thesis demonstrate the possibility of identifying patterns capable of objectively and quantitatively distinguishing these BM states in patients with NHL, thus confirming the hypothesis. In fact, the individual evaluation by NHL subtypes, due to differences in uptake based on aggressiveness, shows even greater potential. However, due to the limitations of the database, the generalisation of the implemented models is constrained, and for bioethical reasons, their immediate clinical application as a decision support system is not intended. For this reason, the hypothesis is partially confirmed.

Nevertheless, thanks to the novelty and potential demonstrated, this thesis lays the foundation for the use of the developed methodologies, so that with a more comprehensive database, a future clinical decision support system can be developed. In this way, the aim is to evolve towards ensuring more personalised treatment and improving the quality of life for patients.

6.2. Contributions of the thesis

Based on the main hypothesis and the obtained results, a combination of contributions can be detailed. These are divided into main contributions and scientific publications.

6.2.1. Main contributions

The contributions included in this section are the main contributions that can be defined on the basis of the hypotheses, their confirmation or rejection, and the results presented.

MC1. Definition of an automatic bone and BM segmentation methodology from CT images based on traditional image processing algorithms.

Due to the need for an accurate segmentation of BM for its quantification, the development of an automatic segmentation emerges as a secondary objective of this thesis. Based on the reviewed state of the art, traditional image processing algorithms are applied in limited cases to BM segmentation, with bone segmentation often being an essential preliminary step.

As an initial approach for segmenting this organ and following the steps outlined in the state of the art, a methodology for automatic bone segmentation is developed using traditional image processing algorithms, such as thresholding and morphological operations. Subsequently, the compact bone is removed to retain only the BM. Despite being a preliminary approach and involving unsupervised segmentation, acceptable results are achieved for this organ. Furthermore, the essential steps that are repeated in subsequent image segmentation processes are established: creation of the CT volume from DICOM files, “co-registration” of the PET with the CT image, and creation of the SUV map for subsequent quantification. This methodology, which appears as an initial segmentation approach, is published as part of a JCR-indexed journal article:

E. Milara, A. Gómez-Grande, S. Tomás-Soler, A.P. Seiffert, R. Alonso, E.J. Gómez, J. Martínez-López, P. Sánchez-González, “Bone marrow segmentation and radiomics analysis of [¹⁸F]FDG PET/CT images for measurable residual disease assessment in multiple myeloma”, Comput Methods Programs Biomed 225 (2022) 107083. <https://doi.org/10.1016/J.CMPB.2022.107083>.

MC2. Design and implementation of a platform for the automatic segmentation of bone and BM, through the integration of automatic models, and their subsequent manual edition.

As previously mentioned, a precise segmentation is required for adequate quantification. Therefore, using this initial approach of BM segmentation and addressing its limitations, a manual editing platform for the masks obtained automatically is implemented. First, the platform integrates automatic segmentation to ease its application. Subsequently, to enhance usability, multiple functions are incorporated into the platform, including basic mask editing, visualisation adjustments, volume rendering, and more complex editing functions based on thresholding, dilation, etc. In this way, a platform is implemented for automatic segmentation with optional manual editing of bone and BM, which can be generalised to any other segmentation based on CT if desired by the user. In addition, the platform can be used as a mean for [^{18}F]FDG PET imaging visualization over CT imaging, allowing the user to evaluate the functional nature of the [^{18}F]FDG PET imaging.

Despite the complexity introduced by the extensive number of added functions, the essential ones are located in the main window to ensure a quicker and more user-friendly experience. Additionally, the user has access to multiple information buttons to provide clear guidance throughout the entire segmentation process.

Finally, it establishes an initial foundation for implementing all the steps developed in this thesis into clinical practice. Thus, by advancing its implementation and improving the generalisation of the quantification models, it would be possible to achieve not only segmentation but also quantification of BM in various clinical contexts, creating a comprehensive clinical decision support system.

MC3. Definition of an automatic bone and BM segmentation methodology from CT images based on a deep learning model.

Using BM masks automatically segmented through the methodology based on traditional image processing algorithms, followed by manual review of each case using the implemented platform, accurate segmentations of both organs are obtained for training supervised segmentation models. This way, independent models for bone and BM segmentation are created.

Regarding bone, as observed in the state of the art, a U-Net architecture is employed due to its potential in this clinical context, and its robustness in handling databases with a limited number of samples. The results of applying this DL technique for bone segmentation are presented in the following JCR-indexed journal article:

E. Milara, A. Gómez-Grande, P. Sarandeses, A.P. Seiffert, E.J. Gómez, P. Sánchez-González, “Automatic Skeleton Segmentation in CT Images Based on U-Net”, *Journal of Imaging Informatics in Medicine* (2024).
<https://doi.org/10.1007/s10278-024-01127-5>.

Based on this study, the segmentation of BM, the organ of interest for this thesis, is designed. Using the same database as for bone segmentation, pGT of BM masks are generated from bone masks to significantly reduce the time required for manual segmentations. Continuing to benefit from this study, the bone pre-processing methodology is adapted to focus on BM as the area of interest. Finally, using transfer learning due to the limited real GT in the database, the bone segmentation model is retrained to achieve optimal BM segmentation. Moreover, despite relying on the initial bone segmentation model, its pre-processing, and its database for training the BM segmentation model, the final model is entirely independent of bone segmentation. Once trained, the final model uses only the CT as input, generating the BM mask as output without requiring the intermediate step of obtaining the bone mask. This represents an improvement not only over the previous segmentation methodology but also over most studies found in the state of the art. The results of this segmentation are published as part of the following JCR-indexed journal article:

E. Milara, P. Sarandeses, A. Jiménez-Ubieto, A. Saviatto, A.P. Seiffert, F.J. Gárate, D. Moreno-Blanco, M. Poza, E.J. Gómez, A. Gómez-Grande, P. Sánchez-González, “Machine Learning Models Based on $[^{18}\text{F}]$ FDG PET Radiomics for Bone Marrow Assessment in Non-Hodgkin Lymphoma”, Applied Sciences 14 (2024) 10291. <https://doi.org/10.3390/app142210291>.

Finally, both segmentation models are included in the previously detailed platform, in order to obtain the optimal segmentation for both, bone and BM. This way, an automatic and optimal segmentation of bone and BM is integrated into a platform for clinical or research aims. As observed in most BM quantification studies in the state of the art, the proposed algorithms are often based on ROIs of maximum SUV or specific manually selected locations. Thanks to the precise segmentation achieved, BM quantification can be performed comprehensively, assessing not only the points of maximum uptake but also the entire organ. This approach enables the evaluation of characteristics related to its heterogeneous uptake pattern in pathological conditions.

MC4. Definition and implementation of methodologies for BM quantification in $[^{18}\text{F}]$ FDG PET/CT imaging.

Despite the growing importance of visually assessing $[^{18}\text{F}]$ FDG PET/CT imaging to evaluate BM status in various clinical contexts where it may be affected, particularly as a complementary test to biopsy, this assessment remains purely visual, making it observer-dependent. On the other hand, although image quantification techniques aimed at objectifying evaluations are gaining prominence over the years, BM evaluation remains an underexplored field. In the state of the art, some studies also

implement radiomics features, and even classifiers based on artificial intelligence, for assessing BM status. However, none of these studies focus on whole-body BM quantification, limiting their scope to analysing bone or BM in specific regions such as vertebrae, femurs, or pelvis.

In this thesis, a robust quantification methodology based on whole-body BM is developed, divided into two distinct approaches, aiming to provide the optimal solution for each problem. On the one hand, a quantification methodology is implemented, focusing on training ML models based on radiomics features extracted from [^{18}F]FDG PET imaging in BM region, aiming to identify textural features that provide information about its functional activity. Before training the models, a univariate analysis is conducted to identify key biomarkers that enable better differentiation of the different BM states. On the other hand, quantification is performed using CNN architectures, taking the original image as input rather than extracted features. In this way, the potential of radiomics features in classifying BM status is assessed, while also exploring the feasibility of identifying more complex architectures, which find deep radiomics, capable of performing this distinction.

Although the limitations of the available databases prevent performing both quantifications for all clinical contexts, the quantification methodology is adapted to each case, proposing a specific solution for each available clinical context. Moreover, the implemented methodology is not limited to BM quantification; with proper segmentation of any other organ and an analysis of its functionality through [^{18}F]FDG PET imaging, it could be adapted to quantify any organ.

MC5. Identification of classifiers for the differentiation of MRD+ and MRD- in patients diagnosed with MM from [^{18}F]FDG PET/CT treatment response evaluation images.

Due to the discrepancies observed between the two techniques used for MRD diagnosis, i.e. biopsy and visual assessment of [^{18}F]FDG PET/CT imaging, and the subjectivity inherent in this visual evaluation, there is a need to develop quantification methodologies to assist clinicians in making a more objective assessment. For this reason, as identified in the state of the art, image quantification is gaining importance to assist clinicians in the field of MM. However, no studies, apart from those encompassed within this thesis, directly address MRD.

Both the precise segmentation methodology and the robust quantification methodology are applied to detect MRD in patients diagnosed with MM and subsequently treated. The quantification results highlight the potential of both quantification methodologies not only for distinguishing MRD+ and MRD- patients assessed through both [^{18}F]FDG PET and biopsy but also for identifying uptake

patterns not visible to the human eye, such as patients visually evaluated as negative-MRD. The results of this study are presented in a conference proceeding:

E. Milara, A.P. Seiffert, F.J. Gárate, D. Moreno-Blanco, E.J. Gómez, A. Gómez-Grande, R. Alonso, P. Sánchez-González, “Segmentación y cuantificación de la médula ósea en imágenes $[^{18}\text{F}]$ FDG PET/CT para la detección de la Enfermedad Residual Medible en pacientes con Mieloma Múltiple”, in: *XLII Congreso Anual de La Sociedad Española de Ingeniería Biomédica, Sevilla, 2024: pp. 497–500.*

Regarding the presentation of these results on MRD detection using $[^{18}\text{F}]$ FDG PET/CT imaging, an initial proof of concept was conducted with a total of 39 patients (8 MRD+/31 MRD-) and the extraction of only 30 radiomics features. In this case, quantification is also limited to the application of ML models without hyperparameter optimisation and the segmentation methodology based on traditional algorithms. These preliminary results are published in a JCR-indexed journal article:

E. Milara, A. Gómez-Grande, S. Tomás-Soler, A.P. Seiffert, R. Alonso, E.J. Gómez, J. Martínez-López, P. Sánchez-González, “Bone marrow segmentation and radiomics analysis of $[^{18}\text{F}]$ FDG PET/CT images for measurable residual disease assessment in multiple myeloma”, *Comput Methods Programs Biomed* 225 (2022) 107083. <https://doi.org/10.1016/J.CMPB.2022.107083>.

Continuing with this study, the same methodology is applied but exclusively evaluating the regions corresponding to the biopsy area, i.e. breastbone, iliac left crest or iliac right crest. The aim is to assess the representativeness of these flat bones, commonly used for biopsy obtention, in evaluating BM affection in MRD. To achieve this, two tests are performed: (1) a correlation study of the radiomics features extracted from the whole-body BM and those extracted from only the biopsy regions; and (2) the development of ML models based on the radiomics features extracted solely from the biopsy regions for MRD detection. All these results are included in a study published in a JCR-indexed journal:

E. Milara, R. Alonso, L. Maseing, A.P. Seiffert, A. Gómez-Grande, E.J. Gómez, J. Martínez-López, P. Sánchez-González, “Radiomics analysis of bone marrow biopsy locations in $[^{18}\text{F}]$ FDG PET/CT images for measurable residual disease assessment in multiple myeloma”, *Phys Eng Sci Med* 46 (2023) 903–913. <https://doi.org/10.1007/s13246-023-01265-0>.

MC6. Identification of classifiers for the prediction of MRD+ or MRD- in patients diagnosed with MM from $[^{18}\text{F}]$ FDG PET/CT diagnostic images.

Due to the prognostic importance of MRD detection, the possibility not only to assist in its detection but also to help predict it, or at least to identify patients with a higher likelihood of developing it, is evaluated. Just as studies for MRD detection are scarce, studies for its prediction are non-existent. No studies are found that attempt to predict the higher or lower probability of developing MRD based on the images used for diagnosis. The quantification methodology implemented in this thesis is limited to the first type of quantification, based on ML models and radiomics, for the available MM database. However, despite the limitations and the complexity of the study, as it aims to predict the subsequent condition of MM without considering any treatment-related factors, the results obtained show a certain trend, meaning that the state of the BM at the time of diagnosis may affect a higher probability of developing MRD. Optimizing this study would improve patient care with greater treatment personalization. A proof of concept for this study, with a smaller database and more limited feature extraction, is presented in a conference article:

E. Milara, R. Cenarro, A. Gómez-Grande, A.P. Seiffert, R. Alonso, E.J. Gómez, P. Sánchez-González, “Análisis textural de imágenes $[^{18}\text{F}]$ FDG PET/CT para la predicción de enfermedad mínima residual en pacientes con sintomatología de mieloma múltiple”, in: *XL Congreso Anual de La Sociedad Española de Ingeniería Biomédica, Valladolid, 2022: pp. 22–25.*

MC7. Identification of potential biomarkers for risk of progression to MM prediction in patients diagnosed with SMM from $[^{18}\text{F}]$ FDG PET/CT diagnostic images.

SMM is a condition characterised by being an asymptomatic form of MM. In most cases, patients diagnosed with SMM progress to MM, with the progression time being uncertain. For this reason, experts find it difficult to define a review protocol. To facilitate this, they differentiate SMM patients into three risk groups for progression (low, intermediate, and high) based on clinical characteristics, as this state is practically undetectable through $[^{18}\text{F}]$ FDG PET/CT imaging.

In the present study, due to the limited data available, univariate analysis, which is part of the first quantification methodology, is applied to identify potential biomarkers capable of distinguishing the different levels of progression risk. In this way, some characteristics are identified that reveal distinct uptake patterns to differentiate certain pairs of progression risk labels, demonstrating the potential of these biomarkers to assist in the diagnosis of SMM through imaging evaluation.

MC8. Identification of classifiers for the differentiation of BMI-, BMI+ and rBM in patients diagnosed with NHL from [^{18}F]FDG PET/CT diagnostic images.

As mentioned earlier, BMI is a common condition in patients with NHL, and its identification is crucial for staging patients due to its therapeutic and prognostic implications. This field is marked not only by the usual discrepancies between BM evaluation tests but also by the simplicity of the studies found in the state of the art, where the methodology and/or the ROI limit their applicability to current clinical practice. Furthermore, no studies that differentiate between healthy or BMI+ states and the heterogeneity observed in the BM of patients undergoing an infectious or inflammatory process, which leads to overproduction of immune cells with increased activity, at the time of diagnostic imaging acquisition for NHL, are found.

In the present thesis, the aforementioned quantification methodologies are applied to differentiate between the different states of the BM: BMI-, BMI+, and rBM. Furthermore, in the case of the first quantification, differentiation is also performed by NHL subtypes, with the aim of observing the possibility of identifying these states differently based on the aggressiveness of each type. The results of this application demonstrate the potential of the quantification methodologies for identifying these states, as well as laying the groundwork for future studies and clinical decision support systems. These results are published in a JCR-indexed journal article:

E. Milara, P. Sarandeses, A. Jiménez-Ubieto, A. Saviatto, A.P. Seiffert, F.J. Gárate, D. Moreno-Blanco, M. Poza, E.J. Gómez, A. Gómez-Grande, P. Sánchez-González, “Machine Learning Models Based on [^{18}F]FDG PET Radiomics for Bone Marrow Assessment in Non-Hodgkin Lymphoma”, *Applied Sciences* 14 (2024) 10291. <https://doi.org/10.3390/app142210291>.

Similarly to the MRD, an initial proof of concept was developed, using a more limited database and radiomic feature extraction, aiming to differentiate only between BMI- and BMI+ patients. This initial proof is presented in a conference article:

J.I. Velasco, **E. Milara**, P. Sarandeses, A.P. Seiffert, A. Gómez-Grande, E.J. Gómez, P. Sánchez-González, “Metodologías de cuantificación de imágenes [^{18}F]FDG PET/CT para la detección de infiltración medular en pacientes con linfoma folicular”, in: *XL Congreso Anual de La Sociedad Española de Ingeniería Biomédica, Valladolid, 2022: pp. 6–9*.

Finally, a future line based on this quantification, which evaluates not only BM uptake but also the spleen (due to its role as an accessory organ of the immune system) and the liver (as the reference organ) is also developed. Despite its proof-of-concept nature, it is also presented in a conference article:

L. Rodríguez-Luque, **E. Milara**, A. Jiménez-Ubieto, D. Moreno-Blanco, A. Saviatto, M. Poza, E.J. Gómez, A. Gómez-Grande, P. Sarandeses, P. Sánchez-González, “Modelo de clasificación de médula ósea en pacientes diagnosticados con linfoma no Hodgkin mediante la cuantificación de médula, bazo e hígado en imágenes [^{18}F]FDG PET/CT”, in: *XLII Congreso Anual de La Sociedad Española de Ingeniería Biomédica, Sevilla, 2024: pp. 325–328*.

6.2.2. Scientific publications

This section includes all the publications made throughout the scientific projection of this doctoral thesis, including articles in JCR indexed journals with impact factor and national congresses.

6.2.2.1. Journal articles

A.P. Seiffert, A. Gómez-Grande, **E. Milara**, S. Llamas-Velasco, A. Villarejo-Galende, E.J. Gómez, P. Sánchez-González, “Texture-Based Analysis of ^{18}F -Labeled Amyloid PET Brain Images”, *Applied Sciences* 11 (2021) 1991. <https://doi.org/10.3390/app11051991>.

E. Milara, A. Gómez-Grande, S. Tomás-Soler, A.P. Seiffert, R. Alonso, E.J. Gómez, J. Martínez-López, P. Sánchez-González, “Bone marrow segmentation and radiomics analysis of [^{18}F]FDG PET/CT images for measurable residual disease assessment in multiple myeloma”, *Comput Methods Programs Biomed* 225 (2022) 107083. <https://doi.org/10.1016/J.CMPB.2022.107083>.

E. Milara, R. Alonso, L. Maseing, A.P. Seiffert, A. Gómez-Grande, E.J. Gómez, J. Martínez-López, P. Sánchez-González, “Radiomics analysis of bone marrow biopsy locations in [^{18}F]FDG PET/CT images for measurable residual disease assessment in multiple myeloma”, *Phys Eng Sci Med* 46 (2023) 903–913. <https://doi.org/10.1007/s13246-023-01265-0>.

D. Palomino-Fernández[†], **E. Milara**[†], Á. Galiana, M. Sánchez-Ortiz, A.P. Seiffert, J. Jiménez-Almonacid, A. Gómez-Grande, S. Ruiz-Solís, A. Ruiz-Alonso, E.J. Gómez, M.J. Tabuenca, P. Sánchez-González, “Textural and Conventional Pretherapeutic [^{18}F]FDG PET/CT Parameters for Survival Outcome Prediction in Stage III and IV Oropharyngeal Cancer Patients”, *Applied Sciences* 14 (2024) 1454. <https://doi.org/10.3390/app14041454>. (†: shared first authorship)

E. Milara, A. Gómez-Grande, P. Sarandeses, A.P. Seiffert, E.J. Gómez, P. Sánchez-González, “Automatic Skeleton Segmentation in CT Images Based on U-Net”, *Journal of Imaging Informatics in Medicine* (2024). <https://doi.org/10.1007/s10278-024-01127-5>.

E. Milara, P. Sarandeses, A. Jiménez-Ubieto, A. Saviatto, A.P. Seiffert, F.J. Gárate, D. Moreno-Blanco, M. Poza, E.J. Gómez, A. Gómez-Grande, P. Sánchez-González, “Machine Learning Models Based on $[^{18}\text{F}]$ FDG PET Radiomics for Bone Marrow Assessment in Non-Hodgkin Lymphoma”, *Applied Sciences* 14 (2024) 10291. <https://doi.org/10.3390/app142210291>.

6.2.2.2. National congresses

E. Milara, A.P. Seiffert, A. Gómez-Grande, A. Villarejo-Galende, H. Bueno-Zamora, E.J. Gómez-Aguilera, P. Sánchez-González, “Texture analysis of ^{18}F -florbetapir PET brain images for the diagnosis of Alzheimer’s disease”, in: *XXXVII Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Cantabria, 2019: pp. 17–20.

E. Milara, R. Cenarro, A. Gómez-Grande, A.P. Seiffert, R. Alonso, E.J. Gómez, P. Sánchez-González, “Análisis textural de imágenes $[^{18}\text{F}]$ FDG PET/CT para la predicción de enfermedad mínima residual en pacientes con sintomatología de mieloma múltiple”, in: *XL Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Valladolid, 2022: pp. 22–25.

J.I. Velasco, **E. Milara**, P. Sarandeses, A.P. Seiffert, A. Gómez-Grande, E.J. Gómez, P. Sánchez-González, “Metodologías de cuantificación de imágenes $[^{18}\text{F}]$ FDG PET/CT para la detección de infiltración medular en pacientes con linfoma folicular”, in: *XL Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Valladolid, 2022: pp. 6–9.

J. De Benavides, **E. Milara**, R. Alonso, A.P. Seiffert, A. Gómez-Grande, E.J. Gómez, P. Sánchez-González, “Algoritmo y herramienta para la segmentación y volumetría esplénica sobre imágenes tomográficas”, in: *XL Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Valladolid, 2022: pp. 490–493.

E. Milara, P. Sarandeses, A. Gómez-Grande, E.J. Gómez, P. Sánchez-González, “Propuesta de modelo basado en redes neuronales para la clasificación de regiones del esqueleto a partir de imágenes TC”, in: *XLI Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Cartagena, 2023: pp. 298–301.

M. Calvo de Mora, **E. Milara**, E.J. Gómez, M. Chiva de Agustín, P. Sánchez-González, “Herramienta para la detección de los márgenes de resección en cirugía conservadora de cáncer de mama”, in: *XLI Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Cartagena, 2023: p. 302.

M. Calvo de Mora-Román, M. Chiva-de Agustín, **E. Milara**, P. Sánchez-González, E.J. Gómez, “Algoritmo y herramienta de apoyo en la evaluación de radiografía de

la pieza quirúrgica en cirugía conservadora del cáncer de mama”, in: *37 Congreso Nacional de La Sociedad Española de Radiología Médica*, Barcelona, 2024: p. 234.

M. de Diego-Peña, S. Vegas-Viedma, M. del Olmo-Reillo, P. Quintana-Zapata, E. Mendoza-Bielsa, N. Brox-Torrecilla, M.I. García-Gómez Muriel, **E. Milara**, P. Sánchez-González, P. Valderrábano, “Trabajo de fin de grado: algoritmo de aprendizaje profundo para la segmentación automática de múltiples tejidos en imágenes de ecografía tiroidea”, in: *65º Congreso de La Sociedad Española de Endocrinología y Nutrición*, Oviedo, 2024: p. 7.

E. Milara, A.P. Seiffert, F.J. Gárate, D. Moreno-Blanco, E.J. Gómez, A. Gómez-Grande, R. Alonso, P. Sánchez-González, “Segmentación y cuantificación de la médula ósea en imágenes $[^{18}\text{F}]\text{FDG}$ PET/CT para la detección de la Enfermedad Residual Medible en pacientes con Mieloma Múltiple”, in: *XLII Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Sevilla, 2024: pp. 497–500.

L. Rodríguez-Luque, **E. Milara**, A. Jiménez-Ubieto, D. Moreno-Blanco, A. Saviatto, M. Poza, E.J. Gómez, A. Gómez-Grande, P. Sarandeses, P. Sánchez-González, “Modelo de clasificación de médula ósea en pacientes diagnosticados con linfoma no Hodgkin mediante la cuantificación de médula, bazo e hígado en imágenes $[^{18}\text{F}]\text{FDG}$ PET/CT”, in: *XLII Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Sevilla, 2024: pp. 325–328.

M. de Diego-Peña, **E. Milara**, E.J. Gómez, S. Vegas-Viedma, M. del Olmo-Reillo, P. Quintana-Zapata, N. Brox-Torrecilla, M.I. García-Gómez Muriel, E. Mendoza-Bielsa, H.G. Pián-Arias, I. Ruz-Caracuel, P. Luengo-Pierrard, P. Valderrábano-Herrero, P. Sánchez-González, “Herramienta para la segmentación automática de videos ecográficos tiroideos basada en modelos de aprendizaje profundo”, in: *XLII Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Sevilla, 2024: pp. 169–172.

M. Bonilla-Iñigo, **E. Milara**, S. Vegas-Viedma, M. del Olmo-Reillo, P. Quintana-Zapata, E.J. Gómez, T. Presa-Abos, M. Chiva-de Agustín, B. Pérez-Mies, J. Palacios-Calvo, P. Luengo-Pierrard, P. Valderrábano-Herrero, P. Sánchez-González, “Sistema de ayuda a la decisión clínica para la predicción de malignidad de nódulos mamarios a partir de imágenes ecográficas”, in: *XLII Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Sevilla, 2024: pp. 17–20.

A. Alonso-Luna, **E. Milara**, Á. Tejerizo-García, E.J. Gómez, M.R. Oliver-Pérez, P. Sánchez-González, “Modelos para la predicción de la invasión del espacio linfovascular en pacientes con cáncer de endometrio”, in: *XLII Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Sevilla, 2024: pp. 265–268.

6.3. Future works

This section presents the proposed future works for continuing with the research developed in the present thesis. All of them are individually mentioned in discussion sections. Some of them, shown in Figure 104, are the ones corresponding to the observed limitations. The rest appears as completely new research lines, based on issues observed during the thesis development.

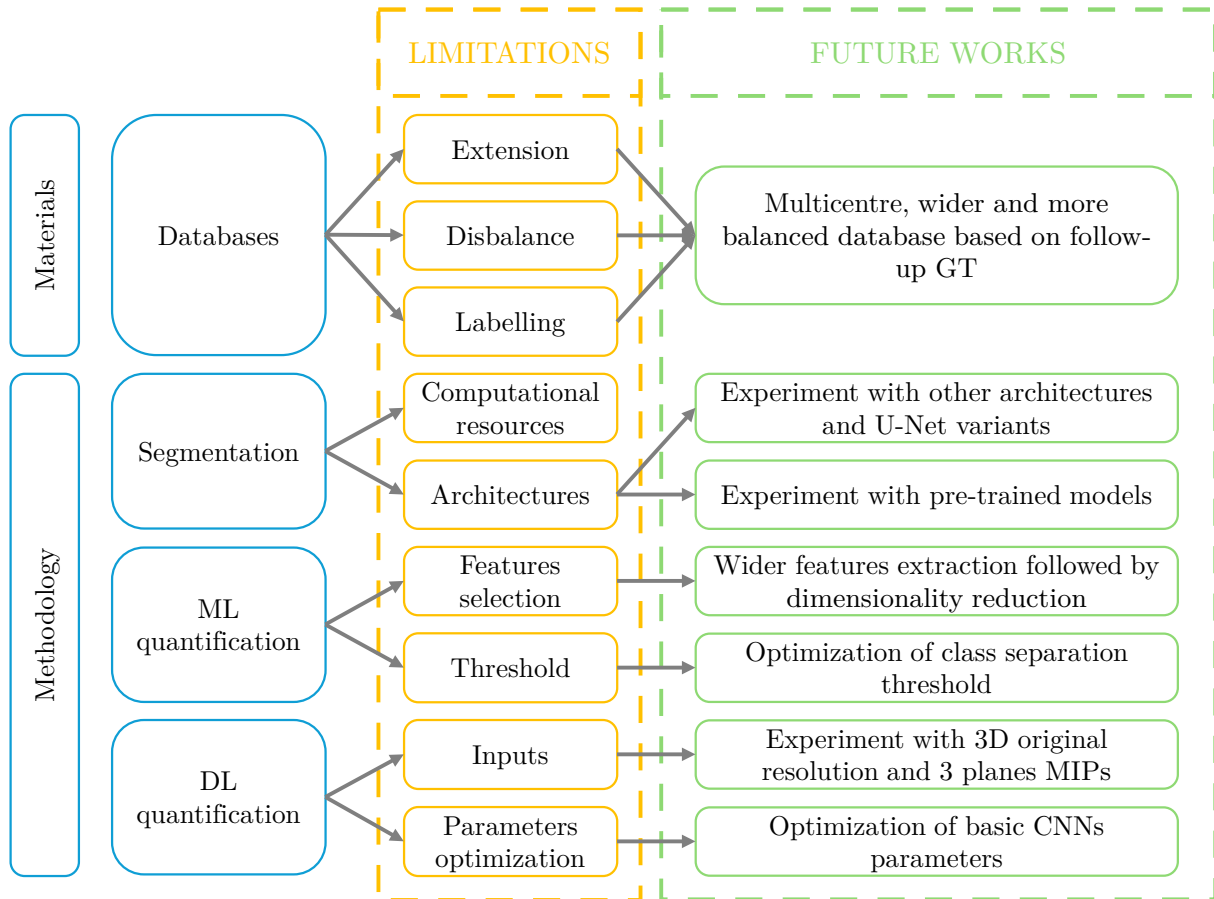


Figure 104. Outline of possible future lines of action based on the limitations encountered.

6.3.1. Future works based on limitations

Multicentre, wider and more balanced database based on follow-up GT

As mentioned in all of the clinical contexts, the main limitation is the used databases. In terms of segmentation, although the developed models are trained for 2D segmentation, ensuring that the number of samples is adequate, the origin of these samples is the same, 77 CT scans. In addition, all of them are acquired from the same centre, scanner and acquisition protocols. In terms of quantification, since samples are the volumetric data, the number of samples is even more critical. Furthermore, the clear disbalance between classes in all of the clinical contexts, and

the discrepancy between biopsy and ^{18}F FDG PET/CT visual assessment, makes the representativeness of the database even worse.

For this reason, the main research line is a proposal of a multicentre project where different hospitals provide ^{18}F FDG PET/CT images for creating a wider and more balanced database. In terms of segmentation, a robust data augmentation process is also proposed, in order to obtain a better performance. For example, in the case of the stretcher removal model, this data augmentation process would allow the removal of this structure even if it is rotated in a different direction. In terms of quantification, a real GT based on the follow-up of the patients, instead of biopsy and ^{18}F FDG PET/CT evaluations, should be taken as the label to predict. This way, the optimal information for creating models which can generalize the presented problems, i.e. models which can be used as clinical decision support, would be used. The difficulty of this last part should be considered, since only patients which are not being treated by clinical decision based on current practical routine can be used, because the treatment can change the follow-up while the ^{18}F FDG PET/CT shows actually another BM state.

Experiment with other architectures and U-Net variants

Although the U-Net architecture is being used in the present thesis due to its usefulness for bone segmentation demonstrated in previous studies [78,79], and its capability of handling small datasets, there are several architectures which can be used with the same aim, such as DeepLab [226], Fully Convolutional Network [227] or variants of the U-Net [228–231]. This limitation is extended to BM segmentation, since due to the limited number of GT, the bone segmentation model is re-trained to segment BM, thanks to transfer learning capability. Observing the benefit of the BM segmentation results due to bone model use, the implementation of other architectures for bone segmentation could lead to other pre-trained models for BM segmentation re-training, being a possibility of obtaining a better performance for BM segmentation.

Experiment with pre-trained models

In addition to the possibility of using other architectures, there is increasing interest in using models that have been previously trained for different purposes, especially in cases with insufficient number of samples. This way, a possible future work is the use of previously trained models for other purposes, which are published online free to be used, to achieve, by means of fine-tuning methods, a possible better segmentation, compensating databases limitations.

Wider features extraction followed by dimensionality reduction

For the first quantification methodology, the radiomics features selection is limited to the extraction of 125 features and a univariate analysis limited to study individually each feature, without considering possible information obtained from combination of some of them. For this reason, a new future work for a better features selection is proposed. First, reducing the extraction limitation by using wavelets and other filters for obtaining a larger number of radiomics features. This way, finding features that suit the problem better is desired, rather than using even more features for training. Secondly, substituting univariate analysis by dimensional reduction techniques such as PCA, trying to retain essential information due to features combination.

Optimization of class separation threshold

Also for the first quantification methodology, oversampling techniques are used for compensating database disbalance. Nevertheless, as mentioned, the use of this technique is being unrecommended, due to the possibility of unrepresentative synthetic samples creation. For this reason, a future work focused on the optimization of separation between classes threshold is proposed. This way, imbalance between classes could be compensated without using false or synthetic samples.

Experiment with 3D original resolution and 3 planes MIPs

For the second quantification methodology, based on CNNs, the training process is limited to the inputs used. In terms of 3D input, the main limitation is the computational resources, which prevent from using the whole 3D resolution. For this reason, if a less restrictive computational resources are available, a future work based on using the original resolution 3D input is proposed. On the other hand, for MIP input, which is also demonstrated to be a useful option when resources are limited, the use, not only the MIP of coronal plane, but also the MIP of the rest of main planes (sagittal and axial) to increase the potential information used for training CNNs, is proposed.

Optimization of basic CNNs parameters

Also for the second quantification methodology, fixed values for the parameters are used. This way, the optimization of the quantification performance is limited by the potential of these values for the presented problems. For this reason, a future work based on the optimisation of these parameters, with the aim of improving quantification results, is proposed. This way, different values for learning rate, optimiser and loss function would be assessed.

6.3.2. New research lines

Post-processing for limbs and head removal

Although bone and BM segmentation can be completely automatic, skipping the manual selection of the first and last slices, since a method for their automatic location is implemented, can lead to errors in the location of these slices in cases with unusually high soft tissue uptake. For this reason, implementing a post-processing methodology for limbs and head removal after the automatic segmentation appears as a possible future work. This way, depending on its performance, a better automatic performance could be obtained.

Python language platform integration

The platform developed is based on MATLAB. Despite the usefulness of MATLAB for medical image processing, it is computationally limited. Consequently, the segmentation models and the entire quantification methodology are implemented in Python. For this reason, implementing the platform in Python is proposed as a future development.

Validation of the platform by nuclear medicine experts

Although the platform is designed with the intention of becoming a useful tool for bone and BM segmentation by nuclear medicine experts, no validation is performed. The initial positive assessment obtained does not appreciate the possible complexity provided by the diversity of functions implemented. A possible clinical evaluation with a panel of experts would improve the user experience and allow the implementation of functions that, from the point of view of the experts, would allow easier and/or more accurate segmentation.

Integration of quantification methodologies into the platform

As mentioned on multiple occasions throughout this thesis, there are currently no quantification models sufficiently generalised to the presented problems to avoid ethical conflicts in developing a diagnostic support tool. However, after obtaining more representative databases and higher-performing models (as a consequence of other proposed future works), the integration of optimal quantification methodologies into the developed platform, as a new tab, is proposed. This would enable the creation of a complete decision support system, encompassing automatic segmentation of the VOI and prediction of the BM state, capable to handle all the clinical contexts discussed. This way, a tool for helping clinicians in order to personalise the prescribed treatment or follow-up could be generated.

Individualised assessment of the different regions of the skeleton

Based on the studies previously developed in the framework of the present thesis, one focused on evaluating the representativeness of biopsy sites [235], and the other focused on the proposal of a DL model for differentiating 8 different anatomical structures [238], a new future work is proposed. This future line is based on the defined quantification methodologies, substituting the whole-body BM evaluation for the features of each region. This way, potential regions with a higher informative value can be identified, serving as a help for focusing visual assessment or even for limiting the follow-up acquisition images to a concrete region.

Inclusion of clinical features into databases

Quantification methodologies are exclusively based on [^{18}F]FDG PET imaging, complicating the MFC results classification. However, the majority of BM affectations have also clinical sings. For this reason, a future work based on combining both clinical features and features extracted from [^{18}F]FDG PET imaging is proposed. This work can be especially important for MRD prediction or risk to progression to MM in SMM, since currently only imaging information is used for MRD prediction, without considering treatment information for quantification, and as mentioned, [^{18}F]FDG PET imaging is doubtful for visual assessment of risk evaluation in SMM patients. For both, clinical features based on the diagnosis parameters would be included: percentage of plasma cell in BM biopsy, level of paraprotein in serum or urine, haematological conditions such as hypercalcemia or anaemia, number of focal lesions, etc. In the case of SMM, this clinical information would be ideal if the GT were based on the follow-up of the patients, as this clinical information is currently used for the IMWG and PETHEMA labelling methods.

CNNs based on radiomics features

Although the use of images as input for CNNs architectures allows the obtention of deep radiomics, normal radiomics can be used as the input of CNNs architectures, trying to find more complex models based on these possible biomarkers which suit better the defined problems. This way, the previous future work based on including clinical features could be also included in CNNs architectures, and not only in the ML models quantification. Furthermore, the future work based on extracting radiomics features for each of the 8 anatomical regions, incrementing the number of extracted features to evaluate by 8, can be efficiently applied. For these reasons, the use of CNNs architectures based on features inputs instead of images, is proposed.

NHL subtypes quantification models

As observed in results, the division of both available NHL subtypes improves the performance of quantification methodologies, demonstrating the differences between

subtypes due to its aggressiveness. For this observation, a new proposed future work based on obtaining a satisfactory database for each NHL subtype is proposed. This way, quantification methodologies can be applied to any of these NHL subtypes databases, personalizing the model to its proper characteristics due to the subtype.

Spleen and liver quantification for BM status differentiation

As explained in subsection 1.2.2.4. *Reactive bone marrow*, bladder uptake of $[^{18}\text{F}]\text{FDG}$ should increase in rBM patients due to immune cells overproduction in their function as accessory immune organ, being even higher than the liver, i.e. reference organ. Considering this fact, a future work based on the quantification of not only BM, but also spleen and liver is proposed. For this future work, a proof of concept is presented in a conference article [239] where both liver and bladder are segmented using U-Net. Then, the uptake of $[^{18}\text{F}]\text{FDG}$ of three organs, i.e. BM, liver and bladder, is used to train an AlexNet architecture. Due to computational limitations, the input of the model is individually (1) the coronal MIP image of the three organs, and (2) three coronal MIP images, one per organ. For this future work, a more extensive investigation line is proposed, by extracting radiomics individually for each organ for ML models quantification and by using the variety of architectures and 3D inputs for CNN quantification detailed in the present thesis.

References

- [1] B.J. Bain, D.M. Clark, I.A. Lampert, B.S. Wilkins, Bone marrow pathology, Blackwell Sciences, 2001.
- [2] J. van der Walt, A. Orazi, D.A. Arber, Diagnostic Bone Marrow Haematopathology, Cambridge University Press, 2020. <https://doi.org/10.1017/9781316535042>.
- [3] T.J. Molina, Hematopathology, Springer, 2020.
- [4] National Cancer Institute, Bone Marrow Definition, Terese Winslow (2014). <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/bone-marrow> (accessed December 10, 2024).
- [5] D.A.G. Barisas, K. Choi, Extramedullary hematopoiesis in cancer, Exp Mol Med (2024). <https://doi.org/10.1038/s12276-024-01192-4>.
- [6] L. Sompayrac, How The Immune System works, 5th ed., Wiley Blackwell, 2016.
- [7] D. Shevyrev, V. Tereshchenko, T.N. Berezina, S. Rybtsov, Hematopoietic Stem Cells and the Immune System in Development and Aging, Int J Mol Sci 24 (2023) 5862. <https://doi.org/10.3390/ijms24065862>.
- [8] J. Wong, R. Jackson, L. Chen, J. Song, R. Pillai, M. Afkhami, O. Danilova, P. Aoun, K.K. Gaal, Y. Kim, Determination of age-dependent bone marrow normocellularity, Am J Clin Pathol 161 (2024) 170–176. <https://doi.org/10.1093/ajcp/aqad129>.
- [9] E. Matutes, B.J. Bain, A. Wotherspoon, Lymphoid Malignancies: An Atlas of Investigation and Diagnosis, 1st ed., Oxford, 2007.
- [10] L. Rao, X. Wang, Z. Zong, Z. Chen, X. Shi, C. Yi, X. Zhang, Z. Yang, PET-CT for Evaluation of Spleen and Liver 18 F-FDG Diffuse Uptake Without Lymph Node Enlargement in Lymphoma, Medicine (United States) 95 (2016). <https://doi.org/10.1097/MD.0000000000003750>.
- [11] C. Nanni, E. Zamagni, A. Versari, S. Chauvie, A. Bianchi, M. Rensi, M. Bellò, I. Rambaldi, A. Gallamini, F. Patriarca, F. Gay, B. Gamberi, M. Cavo, S. Fanti, Image interpretation criteria for FDG PET/CT in multiple myeloma: a new proposal from an Italian expert panel. IMPeTUs (Italian Myeloma criteria for PET USe), Eur J Nucl Med Mol Imaging 43 (2016) 414–421. <https://doi.org/10.1007/s00259-015-3200-9>.

-
- [12] F. Fend, A. Dogan, J.R. Cook, Plasma cell neoplasms and related entities—evolution in diagnosis and classification, *Virchows Archiv* 482 (2023) 163–177. <https://doi.org/10.1007/s00428-022-03431-3>.
 - [13] S.V. Rajkumar, M.A. Dimopoulos, A. Palumbo, J. Blade, G. Merlini, M.-V. Mateos, S. Kumar, J. Hillengass, E. Kastritis, P. Richardson, O. Landgren, B. Paiva, A. Dispenzieri, B. Weiss, X. LeLeu, S. Zweegman, S. Lonial, L. Rosinol, E. Zamagni, S. Jagannath, O. Sezer, S.Y. Kristinsson, J. Caers, S.Z. Usmani, J.J. Lahuerta, H.E. Johnsen, M. Beksac, M. Cavo, H. Goldschmidt, E. Terpos, R.A. Kyle, K.C. Anderson, B.G.M. Durie, J.F.S. Miguel, International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma, *Lancet Oncol* 15 (2014) e538–e548. [https://doi.org/10.1016/S1470-2045\(14\)70442-5](https://doi.org/10.1016/S1470-2045(14)70442-5).
 - [14] K. Brigle, B. Rogers, Pathobiology and Diagnosis of Multiple Myeloma, *Semin Oncol Nurs* 33 (2017) 225–236. <https://doi.org/10.1016/j.soncn.2017.05.012>.
 - [15] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D.M. Parkin, M. Piñeros, A. Znaor, F. Bray, Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods, *Int J Cancer* 144 (2019) 1941–1953. <https://doi.org/10.1002/ijc.31937>.
 - [16] P. Lin, Multiple Myeloma and Other Plasma Cell Neoplasms, in: *Diagnosis of Blood and Bone Marrow Disorders*, Springer International Publishing, Cham, 2018: pp. 383–403. https://doi.org/10.1007/978-3-319-20279-2_13.
 - [17] R.A. Kyle, T.M. Therneau, S.V. Rajkumar, J.R. Offord, D.R. Larson, M.F. Plevak, L.J. Melton, A Long-Term Study of Prognosis in Monoclonal Gammopathy of Undetermined Significance, *New England Journal of Medicine* 346 (2002) 564–569. <https://doi.org/10.1056/NEJMoa01133202>.
 - [18] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, *CA Cancer J Clin* 71 (2021) 209–249. <https://doi.org/10.3322/caac.21660>.
 - [19] J. Ferlay, M. Colombet, I. Soerjomataram, D.M. Parkin, M. Piñeros, A. Znaor, F. Bray, Cancer statistics for the year 2020: An overview, *Int J Cancer* 149 (2021) 778–789. <https://doi.org/10.1002/ijc.33588>.
 - [20] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, F. Bray, Multiple Myeloma Fact Sheet. Global Cancer Observatory: Cancer Today., Lyon, France: International Agency for Research on Cancer (2024).

- <https://gco.iarc.who.int/media/globocan/factsheets/cancers/35-multiple-myeloma-fact-sheet.pdf> (accessed December 2, 2024).
- [21] B.G.M. Durie, S.E. Salmon, A clinical staging system for multiple myeloma correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival, *Cancer* 36 (1975) 842–854. [https://doi.org/10.1002/1097-0142\(197509\)36:3<842::AID-CNCR2820360303>3.0.CO;2-U](https://doi.org/10.1002/1097-0142(197509)36:3<842::AID-CNCR2820360303>3.0.CO;2-U).
 - [22] R.A. Kyle, B.G.M. Durie, S. V Rajkumar, O. Landgren, J. Blade, G. Merlini, N. Kröger, H. Einsele, D.H. Vesole, M. Dimopoulos, J. San Miguel, H. Avet-Loiseau, R. Hajek, W.M. Chen, K.C. Anderson, H. Ludwig, P. Sonneveld, S. Pavlovsky, A. Palumbo, P.G. Richardson, B. Barlogie, P. Greipp, R. Vescio, I. Turesson, J. Westin, M. Boccadoro, Monoclonal gammopathy of undetermined significance (MGUS) and smoldering (asymptomatic) multiple myeloma: IMWG consensus perspectives risk factors for progression and guidelines for monitoring and management, *Leukemia* 24 (2010) 1121–1127. <https://doi.org/10.1038/leu.2010.60>.
 - [23] A. Palumbo, H. Avet-Loiseau, S. Oliva, H.M. Lokhorst, H. Goldschmidt, L. Rosinol, P. Richardson, S. Caltagirone, J.J. Lahuerta, T. Facon, S. Brinchen, F. Gay, M. Attal, R. Passera, A. Spencer, M. Offidani, S. Kumar, P. Musto, S. Lonial, M.T. Petrucci, R.Z. Orlowski, E. Zamagni, G. Morgan, M.A. Dimopoulos, B.G.M. Durie, K.C. Anderson, P. Sonneveld, J. San Miguel, M. Cavo, S.V. Rajkumar, P. Moreau, Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group, *Journal of Clinical Oncology* 33 (2015) 2863–2869. <https://doi.org/10.1200/JCO.2015.61.2267>.
 - [24] M.-V. Mateos, J.-F. San Miguel, Smoldering Multiple Myeloma: When to Observe and When to Treat?, *American Society of Clinical Oncology Educational Book* (2015) e484–e492. https://doi.org/10.14694/EdBook_AM.2015.35.e484.
 - [25] E. Pérez-Persona, M.-B. Vidriales, G. Mateo, R. García-Sanz, M.-V. Mateos, A.G. de Coca, J. Galende, G. Martín-Nuñez, J.M. Alonso, N. de las Heras, J.M. Hernández, A. Martín, C. López-Berges, A. Orfao, J.F. San Miguel, New criteria to identify risk of progression in monoclonal gammopathy of uncertain significance and smoldering multiple myeloma based on multiparameter flow cytometry analysis of bone marrow plasma cells, *Blood* 110 (2007) 2586–2592. <https://doi.org/10.1182/blood-2007-05-088443>.
 - [26] M.-V. Mateos, M.-T. Hernández, P. Giraldo, J. de la Rubia, F. de Arriba, L.L. Corral, L. Rosiñol, B. Paiva, L. Palomera, J. Bargay, A. Oriol, F. Prosper, J.

- López, E. Olavarria, N. Quintana, J.-L. García, J. Bladé, J.-J. Lahuerta, J.-F. San Miguel, Lenalidomide plus Dexamethasone for High-Risk Smoldering Multiple Myeloma, *New England Journal of Medicine* 369 (2013) 438–447. <https://doi.org/10.1056/NEJMoa1300439>.
- [27] S.V. Rajkumar, D. Larson, R.A. Kyle, Diagnosis of Smoldering Multiple Myeloma, *New England Journal of Medicine* 365 (2011) 474–475. <https://doi.org/10.1056/NEJMc1106428>.
- [28] E. Kastiris, E. Terpos, L. Moulopoulos, M. Spyropoulou-Vlachou, N. Kanellias, E. Eleftherakis-Papaiakovou, M. Gkotsamanidou, M. Migkou, M. Gavriatopoulou, M. Roussou, A. Tasidou, M.A. Dimopoulos, Extensive bone marrow infiltration and abnormal free light chain ratio identifies patients with asymptomatic myeloma at high risk for progression to symptomatic disease, *Leukemia* 27 (2013) 947–953. <https://doi.org/10.1038/leu.2012.309>.
- [29] B.M. Cherry, N. Korde, M. Kwok, E.E. Manasanch, M. Bhutani, M. Mulquin, D. Zuchlinski, M.A. Yancey, I. Maric, K.R. Calvo, R. Braylan, M. Stetler-Stevenson, C. Yuan, P. Tembhare, A. Zingone, R. Costello, M.J. Roschewski, O. Landgren, Modeling progression risk for smoldering multiple myeloma: results from a prospective clinical study, *Leuk Lymphoma* 54 (2013) 2215–2218. <https://doi.org/10.3109/10428194.2013.764419>.
- [30] F. Kraeber-Bodere, B. Jamet, D. Bezzi, E. Zamagni, P. Moreau, C. Nanni, New Developments in Myeloma Treatment and Response Assessment, *Journal of Nuclear Medicine* 64 (2023) 1331–1343. <https://doi.org/10.2967/jnumed.122.264972>.
- [31] B. Paiva, J.J.M. van Dongen, A. Orfao, New criteria for response assessment: role of minimal residual disease in multiple myeloma, *Blood* 125 (2015) 3059–3068. <https://doi.org/10.1182/blood-2014-11-568907>.
- [32] L. Burgos, N. Puig, M.-T. Cedená, M.-V. Mateos, J.J. Lahuerta, B. Paiva, J.F. San-Miguel, Measurable residual disease in multiple myeloma: ready for clinical practice?, *J Hematol Oncol* 13 (2020) 82. <https://doi.org/10.1186/s13045-020-00911-4>.
- [33] K.T. Soh, N. Came, G.E. Otteson, D. Jevremovic, M. Shi, H. Olteanu, A. Natoni, A. Lagoo, E. Theakston, J.P. Óskarsson, M. Gorniak, G. Grigoriadis, M. Arroz, M. Fletcher, P. Lin, P. Ludwig, P. Tembhare, R. Matuzeviciene, M. Radzevicius, S. Kay, W. Chen, C. Cabrita, P.K. Wallace, Evaluation of multiple myeloma measurable residual disease by high sensitivity flow cytometry: An international harmonized approach for data analysis, *Cytometry B Clin Cytom* 102 (2022) 88–106. <https://doi.org/10.1002/cyto.b.22053>.

- [34] M. Fulciniti, N.C. Munshi, J. Martinez-Lopez, F. Di Raimondo, Deep Response in Multiple Myeloma: A Critical Review, *Biomed Res Int* 2015 (2015) 1–7. <https://doi.org/10.1155/2015/832049>.
- [35] S. Kumar, B. Paiva, K.C. Anderson, B. Durie, O. Landgren, P. Moreau, N. Munshi, S. Lonial, J. Bladé, M.-V. Mateos, M. Dimopoulos, E. Kastritis, M. Boccadoro, R. Orłowski, H. Goldschmidt, A. Spencer, J. Hou, W.J. Chng, S.Z. Usmani, E. Zamagni, K. Shimizu, S. Jagannath, H.E. Johnsen, E. Terpos, A. Reiman, R.A. Kyle, P. Sonneveld, P.G. Richardson, P. McCarthy, H. Ludwig, W. Chen, M. Cavo, J.-L. Harousseau, S. Lentzsch, J. Hillengass, A. Palumbo, A. Orfao, S.V. Rajkumar, J.S. Miguel, H. Avet-Loiseau, International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma, *Lancet Oncol* 17 (2016) e328–e346. [https://doi.org/10.1016/S1470-2045\(16\)30206-6](https://doi.org/10.1016/S1470-2045(16)30206-6).
- [36] N.C. Munshi, H. Avet-Loiseau, A.C. Rawstron, R.G. Owen, J.A. Child, A. Thakurta, P. Sherrington, M.K. Samur, A. Georgieva, K.C. Anderson, W.M. Gregory, Association of Minimal Residual Disease With Superior Survival Outcomes in Patients With Multiple Myeloma, *JAMA Oncol* 3 (2017) 28. <https://doi.org/10.1001/jamaoncol.2016.3160>.
- [37] K.R. Shankland, J.O. Armitage, B.W. Hancock, Non-Hodgkin lymphoma, *The Lancet* 380 (2012) 848–857. [https://doi.org/10.1016/S0140-6736\(12\)60605-9](https://doi.org/10.1016/S0140-6736(12)60605-9).
- [38] R. Alaggio, C. Amador, I. Anagnostopoulos, A.D. Attygalle, I.B. de O. Araujo, E. Berti, G. Bhagat, A.M. Borges, D. Boyer, M. Calaminici, A. Chadburn, J.K.C. Chan, W. Cheuk, W.-J. Chng, J.K. Choi, S.-S. Chuang, S.E. Coupland, M. Czader, S.S. Dave, D. de Jong, M.-Q. Du, K.S. Elenitoba-Johnson, J. Ferry, J. Geyer, D. Gratzinger, J. Guitart, S. Gujral, M. Harris, C.J. Harrison, S. Hartmann, A. Hochhaus, P.M. Jansen, K. Karube, W. Kempf, J. Khoury, H. Kimura, W. Klapper, A.E. Kovach, S. Kumar, A.J. Lazar, S. Lazzi, L. Leoncini, N. Leung, V. Leventaki, X.-Q. Li, M.S. Lim, W.-P. Liu, A. Louissaint, A. Marcogliese, L.J. Medeiros, M. Michal, R.N. Miranda, C. Mitteldorf, S. Montes-Moreno, W. Morice, V. Nardi, K.N. Naresh, Y. Natkunam, S.-B. Ng, I. Oschlies, G. Ott, M. Parrens, M. Pulitzer, S.V. Rajkumar, A.C. Rawstron, K. Rech, A. Rosenwald, J. Said, C. Sarkozy, S. Sayed, C. Saygin, A. Schuh, W. Sewell, R. Siebert, A.R. Sohani, R. Tooze, A. Traverse-Glehen, F. Vega, B. Vergier, A.D. Wechalekar, B. Wood, L. Xerri, W. Xiao, The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms, *Leukemia* 36 (2022) 1720–1748. <https://doi.org/10.1038/s41375-022-01620-2>.

- [39] C. Laurent, M. Baron, N. Amara, C. Haioun, M. Dandoit, M. Maynadié, M. Parrens, B. Vergier, C. Copie-Bergman, B. Fabiani, A. Traverse-Glehen, N. Brousse, M.-C. Copin, P. Tas, T. Petrella, M.-C. Rousselet, J. Brière, F. Charlotte, C. Chassagne-Clement, T. Rousset, L. Xerri, A. Moreau, A. Martin, D. Damotte, P. Dartigues, I. Soubeyran, M. Peoch, P. Dechelotte, J.-F. Michiels, A. de Mascarel, F. Berger, C. Bossard, F. Arbion, I. Quintin-Roué, J.-M. Picquenot, M. Patey, B. Fabre, H. Sevestre, C. Le Naoures, M.-P. Chenard-Neu, C. Bastien, S. Thiebault, L. Martin, M. Delage, T. Filleron, G. Salles, T.J. Molina, G. Delsol, P. Brousset, P. Gaulard, Impact of Expert Pathologic Review of Lymphoma Diagnosis: Study of Patients From the French Lymphopath Network, *Journal of Clinical Oncology* 35 (2017) 2008–2017. <https://doi.org/10.1200/JCO.2016.71.2083>.
- [40] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, F. Bray, Hodgkin Lymphoma Fact Sheet. Global Cancer Observatory: Cancer Today., Lyon, France: International Agency for Research on Cancer (2024). <https://gco.iarc.who.int/media/globocan/factsheets/cancers/33-hodgkin-lymphoma-fact-sheet.pdf> (accessed December 2, 2024).
- [41] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, F. Bray, Non-Hodgkin Lymphoma Fact Sheet. Global Cancer Observatory: Cancer Today., Lyon, France: International Agency for Research on Cancer (2024). <https://gco.iarc.who.int/media/globocan/factsheets/cancers/34-non-hodgkin-lymphoma-fact-sheet.pdf> (accessed December 2, 2024).
- [42] K.S. Kurz, S. Kalmbach, M. Ott, A.M. Staiger, G. Ott, H. Horn, Follicular Lymphoma in the 5th Edition of the WHO-Classification of Haematolymphoid Neoplasms—Updated Classification and New Biological Data, *Cancers (Basel)* 15 (2023) 785. <https://doi.org/10.3390/cancers15030785>.
- [43] M. Martelli, A.J.M. Ferreri, C. Agostinelli, A. Di Rocco, M. Pfreundschuh, S.A. Pileri, Diffuse large B-cell lymphoma, *Crit Rev Oncol Hematol* 87 (2013) 146–171. <https://doi.org/10.1016/j.critrevonc.2012.12.009>.
- [44] J.K.C. Chan, P.M. Banks, M.L. Cleary, G. Delsol, C. De Wolf-Peeters, B. Falini, K.C. Gatter, T.M. Grogan, N.L. Harris, P.G. Isaacson, E.S. Jaffe, D.M. Knowles, D.Y. Mason, H.-K. Muller-Hermelink, S.A. Pileri, M.A. Piris, E. Ralfkiaer, H. Stein, R.A. Warnke, A Revised European-American Classification of Lymphoid Neoplasms Proposed by the International Lymphoma Study Group: *A Summary Version*, *Am J Clin Pathol* 103 (1995) 543–560. <https://doi.org/10.1093/ajcp/103.5.543>.

- [45] S. Li, K.H. Young, L.J. Medeiros, Diffuse large B-cell lymphoma, *Pathology* 50 (2018) 74–87. <https://doi.org/10.1016/j.pathol.2017.09.006>.
- [46] L.-M. Wu, F.-Y. Chen, X.-X. Jiang, H.-Y. Gu, Y. Yin, J.-R. Xu, 18F-FDG PET, combined FDG-PET/CT and MRI for evaluation of bone marrow infiltration in staging of lymphoma: A systematic review and meta-analysis, *Eur J Radiol* 81 (2012) 303–311. <https://doi.org/10.1016/j.ejrad.2010.11.020>.
- [47] R. Nakajima, A.J. Moskowitz, L. Michaud, A. Mauguen, C.L. Batlevi, A. Dogan, H. Schöder, Baseline FDG-PET/CT detects bone marrow involvement in follicular lymphoma and provides relevant prognostic information, *Blood Adv* 4 (2020) 1812–1823. <https://doi.org/10.1182/bloodadvances.2020001579>.
- [48] C.B. Johnson, J. Zhang, D. Lucas, The Role of the Bone Marrow Microenvironment in the Response to Infection, *Front Immunol* 11 (2020). <https://doi.org/10.3389/fimmu.2020.585402>.
- [49] S.-H. Lee, W.N. Erber, A. Porwit, M. Tomonaga, L.C. Peterson, ICSH guidelines for the standardization of bone marrow specimens and reports, *Int J Lab Hematol* 30 (2008) 349–364. <https://doi.org/10.1111/j.1751-553X.2008.01100.x>.
- [50] T. Bennike, H. Gormsen, B. Moller, Comparative Studies of Bone Marrow Punctures of the Sternum, the Iliac Crest, and the Spinous Process, *Acta Med Scand* 155 (1956) 377–396. <https://doi.org/10.1111/j.0954-6820.1956.tb14386.x>.
- [51] P. Pasqualetti, R. Casale, A. Collacciani, B.P. Abruzzo, D. Colantonio, Multiple myeloma: Relationship between survival and cellular morphology, *Am J Hematol* 33 (1990) 145–147. <https://doi.org/10.1002/ajh.2830330214>.
- [52] J. Flores-Montero, L. Sanoja-Flores, B. Paiva, N. Puig, O. García-Sánchez, S. Böttcher, V.H.J. van der Velden, J.-J. Pérez-Morán, M.-B. Vidriales, R. García-Sanz, C. Jimenez, M. González, J. Martínez-López, A. Corral-Mateos, G.-E. Grigore, R. Fluxá, R. Pontes, J. Caetano, L. Sedek, M.-C. del Cañizo, J. Bladé, J.-J. Lahuerta, C. Aguilar, A. Báez, A. García-Mateo, J. Labrador, P. Leoz, C. Aguilera-Sanz, J. San-Miguel, M.-V. Mateos, B. Durie, J.J.M. van Dongen, A. Orfao, Next Generation Flow for highly sensitive and standardized detection of minimal residual disease in multiple myeloma, *Leukemia* 31 (2017) 2094–2103. <https://doi.org/10.1038/leu.2017.29>.
- [53] R. Alonso, M.T. Cedená, A. Gómez-Grande, R. Ríos, J.M. Moraleda, V. Cabañas, M.J. Moreno, J. López-Jiménez, F. Martín, A. Sanz, A. Valeri, A. Jiménez, R. Sánchez, J.J. Lahuerta, J. Martínez-López, Imaging and bone marrow assessments improve minimal residual disease prediction in multiple

- myeloma, *Am J Hematol* 94 (2019) 853–861. <https://doi.org/10.1002/ajh.25507>.
- [54] B. Jamet, C. Bailly, T. Carlier, C. Touzeau, C. Nanni, E. Zamagni, L. Barré, A.-V. Michaud, M. Chérel, P. Moreau, C. Bodet-Milin, F. Kraeber-Bodéré, Interest of Pet Imaging in Multiple Myeloma, *Front Med (Lausanne)* 6 (2019). <https://doi.org/10.3389/fmed.2019.00069>.
- [55] E. Zamagni, C. Nanni, L. Dozza, T. Carlier, C. Bailly, P. Tacchetti, A. Versari, S. Chauvie, A. Gallamini, B. Gamberi, D. Caillot, F. Patriarca, M. Macro, M. Boccadoro, L. Garderet, S. Barbato, S. Fanti, A. Perrot, F. Gay, P. Sonneveld, L. Karlin, M. Cavo, C. Bodet-Milin, P. Moreau, F. Kraeber-Bodéré, Standardization of ^{18} F-FDG–PET/CT According to Deauville Criteria for Metabolic Complete Response Definition in Newly Diagnosed Multiple Myeloma, *Journal of Clinical Oncology* 39 (2021) 116–125. <https://doi.org/10.1200/JCO.20.00386>.
- [56] International Myeloma Working Group, Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group., *Br J Haematol* 121 (2003) 749–57. <https://doi.org/10.1046/j.1365-2141.2003.04355.x>.
- [57] L. Rasche, D. Alapat, M. Kumar, G. Gershner, J. McDonald, C.P. Wardell, R. Samant, R. Van Hemert, J. Epstein, A.F. Williams, S. Thanendrarajan, C. Schinke, M. Bauer, C. Ashby, R.G. Tytarenko, F. van Rhee, B.A. Walker, M. Zangari, B. Barlogie, F.E. Davies, G.J. Morgan, N. Weinhold, Monitoring of Residual Disease in Myeloma, *Leukemia* 33 (2019) 1713–1722. <https://doi.org/10.1038/s41375-018-0329-0>.
- [58] P. Moreau, M. Attal, D. Caillot, M. Macro, L. Karlin, L. Garderet, T. Facon, L. Benboubker, M. Escoffre-Barbe, A.M. Stoppa, K. Laribi, C. Hulin, A. Perrot, G. Marit, J.R. Eveillard, F. Caillon, C. Bodet-Milin, B. Pegourie, V. Dorvaux, C. Chateaux, K. Anderson, P. Richardson, N.C. Munshi, H. Avet-Loiseau, A. Gaultier, J.M. Nguyen, B. Dupas, E. Frampas, F. Kraeber-Bodere, Prospective evaluation of magnetic resonance imaging and [18F]fluorodeoxyglucose positron emission tomography-computed tomography at diagnosis and before maintenance therapy in symptomatic patients with multiple myeloma included in the IFM/DFCI 2009 trial, *Journal of Clinical Oncology* 35 (2017) 2911–2918. <https://doi.org/10.1200/JCO.2017.72.2975>.
- [59] C. Sachpekidis, H. Goldschmidt, A. Dimitrakopoulou-Strauss, Positron Emission Tomography (PET) Radiopharmaceuticals in Multiple Myeloma, *Molecules* 25 (2019) 134. <https://doi.org/10.3390/molecules25010134>.

- [60] C. Nanni, E. Zamagni, M. Farsad, P. Castellucci, P. Tosi, D. Cangini, E. Salizzoni, R. Canini, M. Cavo, S. Fanti, Role of 18F-FDG PET/CT in the assessment of bone involvement in newly diagnosed multiple myeloma: preliminary results, *Eur J Nucl Med Mol Imaging* 33 (2006) 525–531. <https://doi.org/10.1007/s00259-005-0004-3>.
- [61] E. Zamagni, C. Nanni, F. Patriarca, E. Englaro, P. Castellucci, O. Geatti, P. Tosi, P. Tacchetti, D. Cangini, G. Perrone, M. Ceccolini, A. Brioli, S. Buttignol, R. Fanin, E. Salizzoni, M. Baccarani, S. Fanti, M. Cavo, A prospective comparison of 18F-fluorodeoxyglucose positron emission tomography-computed tomography, magnetic resonance imaging and whole-body planar radiographs in the assessment of bone disease in newly diagnosed multiple myeloma, *Haematologica* 92 (2007) 50–55. <https://doi.org/10.3324/haematol.10554>.
- [62] E. Zamagni, F. Patriarca, C. Nanni, B. Zannetti, E. Englaro, A. Pezzi, P. Tacchetti, S. Buttignol, G. Perrone, A. Brioli, L. Pantani, C. Terragna, F. Carobolante, M. Baccarani, R. Fanin, S. Fanti, M. Cavo, Prognostic relevance of 18-F FDG PET/CT in newly diagnosed multiple myeloma patients treated with up-front autologous transplantation, *Blood* 118 (2011) 5989–5995. <https://doi.org/10.1182/blood-2011-06-361386>.
- [63] C. Mesguich, C. Hulin, V. Latrabe, A. Lascaux, L. Bordenave, E. Hindié, G. Marit, Prospective comparison of 18-FDG PET/CT and whole-body diffusion-weighted MRI in the assessment of multiple myeloma, *Ann Hematol* (2020) 2869–2880. <https://doi.org/10.1007/s00277-020-04265-2>.
- [64] R.C. Walker, T.L. Brown, L.B. Jones-Jackson, L. De Blanche, T. Bartel, Imaging of multiple myeloma and related plasma cell dyscrasias, *Journal of Nuclear Medicine* 53 (2012) 1091–1101. <https://doi.org/10.2967/jnumed.111.098830>.
- [65] E. Zamagni, P. Tacchetti, M. Cavo, Imaging in multiple myeloma: How? When?, *Blood* 133 (2019) 644–651. <https://doi.org/10.1182/blood-2018-08-825356>.
- [66] C. Mesguich, C. Hulin, V. Latrabe, A. Lascaux, L. Bordenave, E. Hindié, 18F-FDG PET/CT and MRI in the Management of Multiple Myeloma: A Comparative Review, *Frontiers in Nuclear Medicine* 1 (2022). <https://doi.org/10.3389/fnume.2021.808627>.
- [67] H. Schirrmeister, M. Bommer, A. Buck, S. Müller, P. Messer, D. Bunjes, H. Döhner, L. Bergmann, S. Reske, Initial results in the assessment of multiple myeloma using 18F-FDG PET, *Eur J Nucl Med* 29 (2002) 361–366. <https://doi.org/10.1007/s00259-001-0711-3>.

-
- [68] J. Hillengass, L.A. Moulopoulos, S. Delorme, V. Koutoulidis, J. Mosebach, T. Hielscher, M. Drake, S. V. Rajkumar, B. Oestergaard, N. Abildgaard, M. Hinge, T. Plesner, Y. Suehara, K. Matsue, N. Withofs, J. Caers, A. Waage, H. Goldschmidt, M.A. Dimopoulos, S. Lentzsch, B. Durie, E. Terpos, Whole-body computed tomography versus conventional skeletal survey in patients with multiple myeloma: A study of the International Myeloma Working Group, *Blood Cancer J* 7 (2017). <https://doi.org/10.1038/bcj.2017.78>.
- [69] R. Zukovs, C. Antke, E. Mamlins, L.M. Sawicki, A. Mohring, D. Lopez y Nienhoff, A. Boquoi, M. Kondakci, G. Antoch, H.-W. Müller, R. Fenk, R. Haas, 18F-FDG-PET/CT in relapsed multiple myeloma: Are prognostic thresholds different from first-line therapy?, *BMC Med Imaging* 22 (2022) 63. <https://doi.org/10.1186/s12880-022-00788-4>.
- [70] C. Nanni, E. Zamagni, A. Versari, S. Chauvie, A. Bianchi, M. Rensi, M. Bellò, I. Rambaldi, A. Gallamini, F. Patriarca, F. Gay, B. Gamberi, M. Cavo, S. Fanti, Image interpretation criteria for FDG PET/CT in multiple myeloma: a new proposal from an Italian expert panel. IMPeTUs (Italian Myeloma criteria for PET USe), *Eur J Nucl Med Mol Imaging* 43 (2016) 414–421. <https://doi.org/10.1007/s00259-015-3200-9>.
- [71] Y. Wang, E. Chiu, J. Rosenberg, S.S. Gambhir, Standardized uptake value atlas: Characterization of physiological 2-Deoxy-2-[18F]fluoro-d-glucose uptake in normal tissues, *Mol Imaging Biol* 9 (2007) 83–90. <https://doi.org/10.1007/s11307-006-0075-y>.
- [72] E. Zamagni, C. Nanni, L. Dozza, T. Carlier, C. Bailly, P. Tacchetti, A. Versari, S. Chauvie, A. Gallamini, B. Gamberi, D. Caillot, F. Patriarca, M. Macro, M. Boccadoro, L. Garderet, S. Barbato, S. Fanti, A. Perrot, F. Gay, P. Sonneveld, L. Karlin, M. Cavo, C. Bodet-Milin, P. Moreau, F. Kraeber-Bodéré, Standardization of 18 F-FDG-PET/CT According to Deauville Criteria for Metabolic Complete Response Definition in Newly Diagnosed Multiple Myeloma, *Journal of Clinical Oncology* 39 (2021) 116–125. <https://doi.org/10.1200/JCO.20.00386>.
- [73] E. Pelosi, D. Penna, D. Deandreis, A. Chiappella, A. Skanjeti, U. Vitolo, G. Bisi, FDG-PET in the detection of bone marrow disease in Hodgkin's disease and aggressive non-Hodgkin's lymphoma and its impact on clinical management., *The Quarterly Journal of Nuclear Medicine and Molecular Imaging* 52 (2008) 9–16.
- [74] G. Çetin, M.A. Çıkırcıoğlu, T. Özkan, C. Karatoprak, M.C. Ar, A.E. Eşkazan, M. Ayer, A. Cerit, K. Gözübenli, B. Borkü Uysal, S. Erdem, N. Ergül, G. Tatar, T.F. Çermik, Can Positron Emission Tomography and Computed

- Tomography Be a Substitute for Bone Marrow Biopsy in Detection of Bone Marrow Involvement in Patients with Hodgkin's or Non-Hodgkin's Lymphoma?, *Turkish Journal of Hematology* 32 (2015). <https://doi.org/10.4274/tjh.2013.0336>.
- [75] S.F. Barrington, N.G. Mikhaeel, L. Kostakoglu, M. Meignan, M. Hutchings, S.P. Müller, L.H. Schwartz, E. Zucca, R.I. Fisher, J. Trotman, O.S. Hoekstra, R.J. Hicks, M.J. O'Doherty, R. Hustinx, A. Biggi, B.D. Cheson, Role of Imaging in the Staging and Response Assessment of Lymphoma: Consensus of the International Conference on Malignant Lymphomas Imaging Working Group, *Journal of Clinical Oncology* 32 (2014) 3048–3058. <https://doi.org/10.1200/JCO.2013.53.5229>.
- [76] S.F. Barrington, N.G. Mikhaeel, L. Kostakoglu, M. Meignan, M. Hutchings, S.P. Müller, L.H. Schwartz, E. Zucca, R.I. Fisher, J. Trotman, O.S. Hoekstra, R.J. Hicks, M.J. O'Doherty, R. Hustinx, A. Biggi, B.D. Cheson, Role of imaging in the staging and response assessment of lymphoma: Consensus of the international conference on malignant lymphomas imaging working group, *Journal of Clinical Oncology* 32 (2014) 3048–3058. <https://doi.org/10.1200/JCO.2013.53.5229>.
- [77] J.-A. Pérez-Carrasco, B. Acha, C. Suárez-Mejías, J.-L. López-Guerra, C. Serrano, Joint segmentation of bones and muscles using an intensity and histogram-based energy minimization approach, *Comput Methods Programs Biomed* 156 (2018) 85–95. <https://doi.org/10.1016/j.cmpb.2017.12.027>.
- [78] A. Klein, J. Warszawski, J. Hillengaß, K.H. Maier-Hein, Automatic bone segmentation in whole-body CT images, *Int J Comput Assist Radiol Surg* 14 (2019) 21–29. <https://doi.org/10.1007/s11548-018-1883-7>.
- [79] S. Noguchi, M. Nishio, M. Yakami, K. Nakagomi, K. Togashi, Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques, *Comput Biol Med* 121 (2020) 103767. <https://doi.org/10.1016/j.compbimed.2020.103767>.
- [80] C. Sachpekidis, O. Enqvist, J. Ulén, A. Kopp-Schneider, L. Pan, A. Jauch, M. Hajiyaanni, L. John, N. Weinhold, S. Sauer, H. Goldschmidt, L. Edenbrandt, A. Dimitrakopoulou-Strauss, Application of an artificial intelligence-based tool in [18F]FDG PET/CT for the assessment of bone marrow involvement in multiple myeloma, *Eur J Nucl Med Mol Imaging* 50 (2023) 3697–3708. <https://doi.org/10.1007/s00259-023-06339-5>.
- [81] C. Fiandra, S. Rosati, F. Arcadipane, N. Dinapoli, M. Fato, P. Franco, E. Gallio, D. Scaffidi Gennarino, P. Silvetti, S. Zara, U. Ricardi, G. Balestra, Active bone marrow segmentation based on computed tomography imaging in

- anal cancer patients: A machine-learning-based proof of concept, *Physica Medica* 113 (2023) 102657. <https://doi.org/10.1016/j.ejmp.2023.102657>.
- [82] A. Perasso, C. Campi, A.M. Massone, M.C. Beltrametti, Spinal canal and spinal marrow segmentation by means of the hough transform of special classes of curves, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2015: pp. 590–600. https://doi.org/10.1007/978-3-319-23231-7_53.
- [83] G. Sambuceti, M. Brignone, C. Marini, M. Massollo, F. Fiz, S. Morbelli, A. Buschiazzo, C. Campi, R. Piva, A.M. Massone, M. Piana, F. Frassoni, Estimating the whole bone-marrow asset in humans by a computational approach to integrated PET/CT imaging, *Eur J Nucl Med Mol Imaging* 39 (2012) 1326–1338. <https://doi.org/10.1007/s00259-012-2141-9>.
- [84] F. Martínez-Martínez, J. Kybic, L. Lambert, Z. Mecková, Fully automated classification of bone marrow infiltration in low-dose CT of patients with multiple myeloma based on probabilistic density model and supervised learning, *Comput Biol Med* 71 (2016) 57–66. <https://doi.org/10.1016/j.compbimed.2016.02.001>.
- [85] Y. Nishida, S. Kimura, H. Mizobe, J. Yamamichi, K. Kojima, A. Kawaguchi, M. Fujisawa, K. Matsue, Automatic digital quantification of bone marrow myeloma volume in appendicular skeletons - Clinical implications and prognostic significance, *Sci Rep* 7 (2017). <https://doi.org/10.1038/s41598-017-13255-w>.
- [86] H.R. Buie, G.M. Campbell, R.J. Klinck, J.A. MacNeil, S.K. Boyd, Automatic segmentation of cortical and trabecular compartments based on a dual threshold technique for in vivo micro-CT bone analysis, *Bone* 41 (2007) 505–515. <https://doi.org/10.1016/j.bone.2007.07.007>.
- [87] M. Fujino, K. Suzuki, M. Nishio, N. Nishiyama, Y. Osaka, Strategy of radiation therapy for bone metastases and MSCC in breast cancer patients, *Breast Cancer* 18 (2011) 238–243. <https://doi.org/10.1007/s12282-011-0288-z>.
- [88] T. Bäuerle, W. Semmler, Imaging response to systemic therapy for bone metastases, *Eur Radiol* 19 (2009) 2495–2507. <https://doi.org/10.1007/s00330-009-1443-1>.
- [89] D.E. Oprea-Lager, M.C.F. Cysouw, R. Boellaard, C.M. Deroose, L.-F. de Geus-Oei, E. Lopci, L. Bidaut, K. Herrmann, L.S. Fournier, T. Bäuerle, N.M. deSouza, F.E. Lecouvet, Bone Metastases Are Measurable: The Role of Whole-Body MRI and Positron Emission Tomography, *Front Oncol* 11 (2021). <https://doi.org/10.3389/fonc.2021.772530>.

- [90] B. Grubmüller, D. Senn, G. Kramer, P. Baltzer, D. D'Andrea, K.H. Grubmüller, M. Mitterhauser, H. Eidherr, A.R. Haug, W. Wadsak, S. Pfaff, S.F. Shariat, M. Hacker, M. Hartenbach, Response assessment using ⁶⁸Ga-PSMA ligand PET in patients undergoing ¹⁷⁷Lu-PSMA radioligand therapy for metastatic castration-resistant prostate cancer, *Eur J Nucl Med Mol Imaging* 46 (2019) 1063–1072. <https://doi.org/10.1007/s00259-018-4236-4>.
- [91] M.M. Heck, M. Retz, C. D'Alessandria, I. Rauscher, K. Scheidhauer, T. Maurer, E. Storz, F. Janssen, M. Schottelius, H.-J. Wester, J.E. Gschwend, M. Schwaiger, R. Tauber, M. Eiber, Systemic Radioligand Therapy with [¹⁷⁷] Lu Labeled Prostate Specific Membrane Antigen Ligand for Imaging and Therapy in Patients with Metastatic Castration Resistant Prostate Cancer, *Journal of Urology* 196 (2016) 382–391. <https://doi.org/10.1016/j.juro.2016.02.2969>.
- [92] H.K. Genant, K. Engelke, S. Prevrhal, Advanced CT bone imaging in osteoporosis, *Rheumatology* 47 (2008) iv9–iv16. <https://doi.org/10.1093/rheumatology/ken180>.
- [93] Y. Kang, K. Engelke, W.A. Kalender, A new accurate and precise 3-D segmentation method for skeletal structures in volumetric CT data, *IEEE Trans Med Imaging* 22 (2003) 586–598. <https://doi.org/10.1109/TMI.2003.812265>.
- [94] S. Banik, R.M. Rangayyan, G.S. Boag, Automatic Segmentation of the Ribs, the Vertebral Column, and the Spinal Canal in Pediatric Computed Tomographic Images, *J Digit Imaging* 23 (2010) 301–322. <https://doi.org/10.1007/s10278-009-9176-x>.
- [95] J. Kratky, J. Kybic, Three-dimensional segmentation of bones from CT and MRI using fast level sets, in: *Medical Imaging 2008: Image Processing*, SPIE, 2008: p. 691447. <https://doi.org/10.1117/12.770954>.
- [96] Y. Zhang, B.J. Matuszewski, L.K. Shark, C.J. Moore, Medical image segmentation using new hybrid level-set method, in: *Proceedings - 5th International Conference BioMedical Visualization, Information Visualization in Medical and Biomedical Informatics, MediVis 2008*, 2008: pp. 71–76. <https://doi.org/10.1109/MediVis.2008.12>.
- [97] H. Li, Contour extraction of hand-wrist skeletal based on active contour model, in: *2009 WRI World Congress on Software Engineering, WCSE 2009*, 2009: pp. 283–286. <https://doi.org/10.1109/WCSE.2009.101>.
- [98] H. Yousefi, M. Fatehi, M. Amian, R.A. Zoroofi, A fully automated segmentation of radius bone based on active contour in wrist MRI data set, in:

- 20th Iranian Conference on Biomedical Engineering, Redhawk Publications, Tehran, 2013.
- [99] S.Y. Ababneh, J.W. Prescott, M.N. Gurcan, Automatic graph-cut based segmentation of bones from knee magnetic resonance images for osteoarthritis research, *Med Image Anal* 15 (2011) 438–448. <https://doi.org/10.1016/j.media.2011.01.007>.
 - [100] L. Ben Younes, Y. Nakajima, T. Saito, Fully automatic segmentation of the Femur from 3D-CT images using primitive shape recognition and statistical shape models, *Int J Comput Assist Radiol Surg* 9 (2014) 189–196. <https://doi.org/10.1007/s11548-013-0950-3>.
 - [101] R. Korez, B. Likar, F. Pernuš, T. Vrtovec, Model-Based Segmentation of Vertebral Bodies from MR Images with 3D CNNs, in: 2016: pp. 433–441. https://doi.org/10.1007/978-3-319-46723-8_50.
 - [102] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. MICCAI 2015., Springer, Cham, 2015: pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
 - [103] N. Archip, P.J. Erard, M. Egmont-Petersen, J.M. Haeffliger, J.F. Germond, A knowledge-based approach to automatic detection of the spinal cord in CT images, *IEEE Trans Med Imaging* 21 (2002) 1504–1516. <https://doi.org/10.1109/TMI.2002.806578>.
 - [104] L.G. Nyúl, J. Kanyó, E. Máté, G. Makay, E. Balogh, M. Fidrich, A. Kuba, Method for Automatically Segmenting the Spinal Cord and Canal from 3D CT Images, in: *Computer Analysis of Images and Patterns, CAIP 2005*, Springer-Verlag, 2005: pp. 456–463.
 - [105] R.M. Rangayyan, H.J. Deglint, G.S. Boag, Method for the automatic detection and segmentation of the spinal canal in computed tomographic images, *J Electron Imaging* 15 (2006) 033007. <https://doi.org/10.1117/1.2234770>.
 - [106] B. De Leener, S. Kadoury, J. Cohen-Adad, Robust, accurate and fast automatic segmentation of the spinal cord, *Neuroimage* 98 (2014) 528–536. <https://doi.org/10.1016/j.neuroimage.2014.04.051>.
 - [107] N. Li, S.S. Noticewala, C.W. Williamson, H. Shen, I. Sirak, R. Tarnawski, U. Mahantshetty, C.K. Hoh, K.L. Moore, L.K. Mell, Feasibility of atlas-based active bone marrow sparing intensity modulated radiation therapy for cervical cancer, *Radiotherapy and Oncology* 123 (2017) 325–330. <https://doi.org/10.1016/j.radonc.2017.02.017>.

- [108] C. Fiandra, S. Rosati, F. Arcadipane, N. Dinapoli, M. Fato, P. Franco, E. Gallio, D. Scaffidi Gennarino, P. Silvetti, S. Zara, U. Ricardi, G. Balestra, Active bone marrow segmentation based on computed tomography imaging in anal cancer patients: A machine-learning-based proof of concept, *Physica Medica* 113 (2023) 102657. <https://doi.org/10.1016/j.ejmp.2023.102657>.
- [109] F. Fiz, C. Marini, R. Piva, M. Miglino, M. Massollo, F. Bongioanni, S. Morbelli, G. Bottoni, C. Campi, A. Bacigalupo, P. Bruzzi, F. Frassoni, M. Piana, G. Sambuceti, Adult advanced chronic lymphocytic leukemia: Computational analysis of whole-body CT documents a bone structure alteration, *Radiology* 271 (2014) 805–813. <https://doi.org/10.1148/radiol.14131944>.
- [110] F. Fiz, H. Dittman, C. Campi, S. Morbelli, C. Marini, M. Brignone, M. Bauckneht, R. Piva, A.M. Massone, M. Piana, G. Sambuceti, C. La Fougère, Assessment of Skeletal Tumor Load in Metastasized Castration-Resistant Prostate Cancer Patients: A Review of Available Methods and an Overview on Future Perspectives, *Bioengineering* 5 (2018) 58. <https://doi.org/10.3390/bioengineering5030058>.
- [111] D. Schenone, A. Dominietto, C. Campi, F. Frassoni, M. Cea, S. Aquino, E. Angelucci, F. Rossi, L. Torri, B. Bignotti, A.S. Tagliafico, M. Piana, Radiomics and artificial intelligence for outcome prediction in multiple myeloma patients undergoing autologous transplantation: A feasibility study with ct data, *Diagnostics* 11 (2021). <https://doi.org/10.3390/diagnostics11101759>.
- [112] M. Wennmann, J. Chmelik, F. Bauer, A. Klein, C. Uhlenbrock, J. Lochner, M. Grözinger, L. Rotkopf, S. Sauer, D. Bonekamp, J. Kleesiek, T. Weber, J. Hillengass, H. Goldschmidt, H.-P. Schlemmer, R. Floca, N. Weinhold, K. Maier-Hein, S. Delorme, Automatic bone marrow segmentation in whole-body magnetic resonance imaging: towards comprehensive, objective MRI-phenotypic bone marrow characterization in multiple myeloma, in: 18th International Myeloma Workshop, 2021: pp. 45–46.
- [113] M. Wennmann, A. Klein, F. Bauer, J. Chmelik, C. Uhlenbrock, M. Grözinger, L. Rotkopf, S. Sauer, H. Thierjung, M. Götz, D. Bonekamp, J. Kleesiek, T. Weber, J. Hillengass, H. Goldschmidt, H.-P. Schlemmer, R. Floca, N. Weinhold, K. Maier-Hein, S. Delorme, V.H. Jimenez-Zepeda, D.E. Reece, R. Rigo, P. Gogna, S. Kong, X.Y. Hu, P. Chapani, W. Cheung, D. Brenner, R.K. Plante, K. Shi, A. Husain, D. Tankala, D. Boyne, Automatic analysis of magnetic resonance imaging in multiple myeloma patients: deep-learning based pelvic bone marrow segmentation and radiomics analysis for prediction of plasma cell infiltration, in: 18th International Myeloma Workshop, 2021: p. 49.

-
- [114] M.A. Lodge, Repeatability of SUV in oncologic 18F-FDG PET, *Journal of Nuclear Medicine* 58 (2017) 523–532. <https://doi.org/10.2967/jnumed.116.186353>.
 - [115] R. Bar-Shalom, A.Y. Valdivia, M.D. Blafox, PET imaging in oncology, *Semin Nucl Med* 30 (2000) 150–185. <https://doi.org/10.1053/snuc.2000.7439>.
 - [116] E. Alipour, A. Pooyan, F. Shomal Zadeh, A.D. Darbandi, P.A. Bonaffini, M. Chalian, Current Status and Future of Artificial Intelligence in MM Imaging: A Systematic Review, *Diagnostics* 13 (2023) 3372. <https://doi.org/10.3390/diagnostics13213372>.
 - [117] L. Manco, D. Albano, L. Urso, M. Arnaboldi, M. Castellani, L. Florimonte, G. Guidi, A. Turra, A. Castello, S. Panareo, Positron Emission Tomography-Derived Radiomics and Artificial Intelligence in Multiple Myeloma: State-of-the-Art, *J Clin Med* 12 (2023). <https://doi.org/10.3390/jcm12247669>.
 - [118] D. Ripani, C. Caldarella, T. Za, E. Rossi, V. De Stefano, A. Giordano, Progression to Symptomatic Multiple Myeloma Predicted by Texture Analysis-Derived Parameters in Patients Without Focal Disease at 18F-FDG PET/CT, *Clin Lymphoma Myeloma Leuk* 21 (2021) 536–544. <https://doi.org/10.1016/j.clml.2021.03.014>.
 - [119] X. Xiong, J. Wang, S. Hu, Y. Dai, Y. Zhang, C. Hu, Differentiating Between Multiple Myeloma and Metastasis Subtypes of Lumbar Vertebra Lesions Using Machine Learning-Based Radiomics, *Front Oncol* 11 (2021). <https://doi.org/10.3389/fonc.2021.601699>.
 - [120] Z. Jin, Y. Wang, Y. Wang, Y. Mao, F. Zhang, J. Yu, Application of 18F-FDG PET-CT Images Based Radiomics in Identifying Vertebral Multiple Myeloma and Bone Metastases, *Front Med (Lausanne)* 9 (2022). <https://doi.org/10.3389/fmed.2022.874847>.
 - [121] P. Mannam, A. Murali, P. Gokulakrishnan, E. Venkatachalapathy, P. Venkata Sai, Radiomic analysis of positron-emission tomography and computed tomography images to differentiate between multiple myeloma and skeletal metastases, *Indian Journal of Nuclear Medicine* 37 (2022) 217–226. https://doi.org/10.4103/ijnm.ijnm_111_21.
 - [122] Y. Li, Y. Liu, P. Yin, C. Hao, C. Sun, L. Chen, S. Wang, N. Hong, MRI-Based Bone Marrow Radiomics Nomogram for Prediction of Overall Survival in Patients With Multiple Myeloma, *Front Oncol* 11 (2021). <https://doi.org/10.3389/fonc.2021.709813>.
 - [123] B. Jamet, L. Morvan, C. Nanni, A.-V. Michaud, C. Bailly, S. Chauvie, P. Moreau, C. Touzeau, E. Zamagni, C. Bodet-Milin, F. Kraeber-Bodéré, D.

- Mateus, T. Carlier, Random survival forest to predict transplant-eligible newly diagnosed multiple myeloma outcome including FDG-PET radiomics: a combined analysis of two independent prospective European trials, *Eur J Nucl Med Mol Imaging* 48 (2021) 1005–1015. <https://doi.org/10.1007/s00259-020-05049-6>.
- [124] B. Ni, G. Huang, H. Huang, T. Wang, X. Han, L. Shen, Y. Chen, J. Hou, Machine Learning Model Based on Optimized Radiomics Feature from 18F-FDG-PET/CT and Clinical Characteristics Predicts Prognosis of Multiple Myeloma: A Preliminary Study, *J Clin Med* 12 (2023). <https://doi.org/10.3390/jcm12062280>.
- [125] H. Zhong, D. Huang, J. Wu, X. Chen, Y. Chen, C. Huang, 18F-FDG PET/CT based radiomics features improve prediction of prognosis: multiple machine learning algorithms and multimodality applications for multiple myeloma, *BMC Med Imaging* 23 (2023). <https://doi.org/10.1186/s12880-023-01033-2>.
- [126] M. Wennmann, W. Ming, F. Bauer, J. Chmelik, A. Klein, C. Uhlenbrock, M. Grözinger, K.-C. Kahl, T. Nonnenmacher, M. Debic, T. Hielscher, H. Thierjung, L.T. Rotkopf, N. Stanczyk, S. Sauer, A. Jauch, M. Götz, F.T. Kurz, K. Schlamp, M. Horger, S. Afat, B. Besemer, M. Hoffmann, J. Hoffend, D. Kraemer, U. Graeven, A. Ringelstein, D. Bonekamp, J. Kleesiek, R.O. Floca, J. Hillengass, E.K. Mai, N. Weinhold, T.F. Weber, H. Goldschmidt, H.-P. Schlemmer, K. Maier-Hein, S. Delorme, P. Neher, Prediction of Bone Marrow Biopsy Results From MRI in Multiple Myeloma Patients Using Deep Learning and Radiomics, *Invest Radiol* 58 (2023) 754–765. <https://doi.org/10.1097/RLI.0000000000000986>.
- [127] C. Mesguich, E. Hindie, B.D. De Senneville, G. Tlili, J.B. Pinaquy, G. Marit, O. Saut, Improved 18-FDG PET/CT diagnosis of multiple myeloma diffuse disease by radiomics analysis, *Nucl Med Commun* 42 (2021) 1135–1143. <https://doi.org/10.1097/MNM.0000000000001437>.
- [128] M.E.S. Takahashi, C. Mosci, E.M. Souza, S.Q. Brunetto, E. Etchebehere, A.O. Santos, M.R. Camacho, E. Miranda, M.C.L. Lima, B.J. Amorim, C. de Souza, F. V. Pericole, I. Lorand-Metze, C.D. Ramos, Proposal for a Quantitative 18F-FDG PET/CT Metabolic Parameter to Assess the Intensity of Bone Involvement in Multiple Myeloma, *Sci Rep* 9 (2019) 1–8. <https://doi.org/10.1038/s41598-019-52740-2>.
- [129] M.E.S. Takahashi, C. Mosci, G.O. Duarte, F. V. Pericole, K. Metze, I.G.H. Lorand-Metze, C.D. Ramos, Intensity of bone involvement: a quantitative 18F-FDG PET/CT evaluation for monitoring outcome of multiple myeloma, *Nucl*

- Med Commun 42 (2021) 1375–1381.
<https://doi.org/10.1097/MNM.0000000000001470>.
- [130] E.E. Pakos, A.D. Fotopoulos, J.P.A. Ioannidis, 18F-FDG PET for evaluation of bone marrow infiltration in staging of lymphoma: a meta-analysis., *Journal of Nuclear Medicine* 46 (2005) 958–63.
- [131] Y.-K. Chen, C.-L. Yeh, C.-C. Tsui, J.-A. Liang, J.-H. Chen, C.-H. Kao, F-18 FDG PET for Evaluation of Bone Marrow Involvement in Non-Hodgkin Lymphoma, *Clin Nucl Med* 36 (2011) 553–559.
<https://doi.org/10.1097/RLU.0b013e318217aeff>.
- [132] J. Almaini, C. Tsoumpas, R. Feltbower, I. Polycarpou, FDG PET/CT versus Bone Marrow Biopsy for Diagnosis of Bone Marrow Involvement in Non-Hodgkin Lymphoma: A Systematic Review, *Applied Sciences* 12 (2022) 540. <https://doi.org/10.3390/app12020540>.
- [133] R. Elstrom, L. Guan, G. Baker, K. Nakhoda, J.A. Vergilio, H. Zhuang, S. Pitsilos, A. Bagg, L. Downs, A. Mehrotra, S. Kim, A. Alavi, S.J. Schuster, Utility of FDG-PET scanning in lymphoma by WHO classification, *Blood* 101 (2003) 3875–3876. <https://doi.org/10.1182/blood-2002-09-2778>.
- [134] H.J.A. Adams, T.C. Kwee, R. Fijnheer, S. V. Dubois, R.A.J. Nievelstein, J.M.H. de Klerk, Bone marrow 18F-fluoro-2-deoxy-d-glucose positron emission tomography/computed tomography cannot replace bone marrow biopsy in diffuse large B-cell lymphoma, *Am J Hematol* 89 (2014) 726–731.
<https://doi.org/10.1002/ajh.23730>.
- [135] H.J.A. Adams, T.C. Kwee, R. Fijnheer, S. V. Dubois, P.E. Blase, R.A.J. Nievelstein, J.M.H. De Klerk, Utility of quantitative FDG-PET/CT for the detection of bone marrow involvement in follicular lymphoma: A histopathological correlation study, *Skeletal Radiol* 43 (2014) 1231–1236.
<https://doi.org/10.1007/s00256-014-1919-1>.
- [136] N. Aide, M. Talbot, C. Fruchart, G. Damaj, C. Lasnon, Diagnostic and prognostic value of baseline FDG PET/CT skeletal textural features in diffuse large B cell lymphoma, *Eur J Nucl Med Mol Imaging* 45 (2018) 699–711.
<https://doi.org/10.1007/s00259-017-3899-6>.
- [137] I. El-Najjar, S. Montoto, A. McDowell, J. Matthews, J. Gribben, T.A. Szyszko, The value of semiquantitative analysis in identifying diffuse bone marrow involvement in follicular lymphoma, *Nucl Med Commun* 35 (2014) 311–315.
<https://doi.org/10.1097/MNM.0000000000000059>.
- [138] C. Perry, H. Lerman, E. Joffe, N. Sarid, O. Amit, I. Avivi, M. Kesler, J. Ben-Ezra, E. Even-Sapir, Y. Herishanu, The value of PET/CT in detecting bone

- marrow involvement in patients with follicular lymphoma, *Medicine (United States)* 95 (2016). <https://doi.org/10.1097/MD.0000000000002910>.
- [139] U. Asenbaum, R. Nolz, G. Karanikas, J. Furtner, R. Woitek, I. Simonitsch-Klupp, M. Raderer, M.E. Mayerhoefer, Bone Marrow Involvement in Malignant Lymphoma, *Acad Radiol* 25 (2018) 453–460. <https://doi.org/10.1016/j.acra.2017.10.024>.
- [140] J. Faudemer, N. Aide, A.C. Gac, G. Damaj, J.P. Vilque, C. Lasnon, Diagnostic value of baseline 18FDG PET/CT skeletal textural features in follicular lymphoma, *Sci Rep* 11 (2021). <https://doi.org/10.1038/s41598-021-03278-9>.
- [141] E.F. Hollinger, H. Alibazoglu, A. Ali, A. Green, G. Lamonica, Hematopoietic Cytokine-Mediated FDG Uptake Simulates the Appearance of Diffuse Metastatic Disease on Whole-Body PET Imaging, *Clin Nucl Med* 23 (1998) 93–98. <https://doi.org/10.1097/00003072-199802000-00007>.
- [142] S.J. Rosenbaum, T. Lind, G. Antoch, A. Bockisch, False-positive FDG PET uptake - The role of PET/CT, *Eur Radiol* 16 (2006) 1054–1065. <https://doi.org/10.1007/s00330-005-0088-y>.
- [143] P.Y. Salaun, T. Gastinne, C. Bodet-Milin, L. Campion, P. Cambefort, A. Moreau, S. Le Gouill, C. Berthou, P. Moreau, F. Kraeber-Bodéré, Analysis of 18F-FDG PET diffuse bone marrow uptake and splenic uptake in staging of Hodgkin's lymphoma: a reflection of disease infiltration or just inflammation?, *Eur J Nucl Med Mol Imaging* 36 (2009) 1813–1821. <https://doi.org/10.1007/s00259-009-1183-0>.
- [144] M. Chen, J. Rong, J. Zhao, Y. Teng, C. Jiang, J. Chen, J. Xu, PET-based radiomic feature based on the cross-combination method for predicting the mid-term efficacy and prognosis in high-risk diffuse large B-cell lymphoma patients, *Front Oncol* 14 (2024). <https://doi.org/10.3389/fonc.2024.1394450>.
- [145] I. Häggström, D. Leithner, J. Alvé, G. Campanella, M. Abusamra, H. Zhang, S. Chhabra, L. Beer, A. Haug, G. Salles, M. Raderer, P.B. Staber, A. Becker, H. Hricak, T.J. Fuchs, H. Schöder, M.E. Mayerhoefer, Deep learning for [18F]fluorodeoxyglucose-PET-CT classification in patients with lymphoma: a dual-centre retrospective analysis, *Lancet Digit Health* 6 (2024) e114–e125. [https://doi.org/10.1016/S2589-7500\(23\)00203-0](https://doi.org/10.1016/S2589-7500(23)00203-0).
- [146] R. Boellaard, R. Delgado-Bolton, W.J.G. Oyen, F. Giammarile, K. Tatsch, W. Eschner, F.J. Verzijlbergen, S.F. Barrington, L.C. Pike, W.A. Weber, S. Stroobants, D. Delbeke, K.J. Donohoe, S. Holbrook, M.M. Graham, G. Testanera, O.S. Hoekstra, J. Zijlstra, E. Visser, C.J. Hoekstra, J. Pruim, A. Willemsen, B. Arends, J. Kotzerke, A. Bockisch, T. Beyer, A. Chiti, B.J.

- Krause, European Association of Nuclear Medicine (EANM), FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0., *Eur J Nucl Med Mol Imaging* 42 (2015) 328–54. <https://doi.org/10.1007/s00259-014-2961-x>.
- [147] Grupo Español de Mieloma, Guía de Mieloma Múltiple Grupo Español de Mieloma, Madrid, 2021. https://www.fundacionpethema.es/sites/default/files/documentos/guides/GUI%CC%81A_MIELOMA%20MU%CC%81LTIPLE_PETHEMA_21ABRIL2021.pdf (accessed December 18, 2024).
- [148] The MathWorks, Inc., “MATLAB, version 9.15.0 (R2023b),” Natick, Massachusetts, 2023., (n.d.). <https://www.mathworks.com> (accessed October 19, 2024).
- [149] The Wellcome Centre for Human Neuroimaging, “Statistical Parametric Mapping,” (n.d.). <https://www.fil.ion.ucl.ac.uk/spm/software/spm12/> (accessed October 19, 2024).
- [150] Python Software Foundation, Python 3.6, (2016). <https://www.python.org/downloads/release/python-368> (accessed November 1, 2024).
- [151] Python Software Foundation, Python 3.9, (2020). <https://www.python.org/downloads/release/python-3913/> (accessed November 1, 2024).
- [152] P. Jaccard, THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE., *New Phytologist* 11 (1912) 37–50. <https://doi.org/10.1111/j.1469-8137.1912.tb05611.x>.
- [153] L.R. Dice, Measures of the Amount of Ecologic Association Between Species, *Ecology* 26 (1945) 297–302. <https://doi.org/10.2307/1932409>.
- [154] M.W. Vannier, T.K. Pilgram, C.M. Speidel, L.R. Neumann, D.L. Rickman, L.D. Schertz, Validation of Magnetic Resonance Imaging (MRI) multispectral tissue classification, *Computerized Medical Imaging and Graphics* 15 (1991) 217–223. [https://doi.org/10.1016/0895-6111\(91\)90079-B](https://doi.org/10.1016/0895-6111(91)90079-B).
- [155] A.A. Sharief, A. Badea, A.M. Dale, G.A. Johnson, Automated segmentation of the actively stained mouse brain using multi-spectral MR microscopy, *Neuroimage* 39 (2008) 136–145. <https://doi.org/10.1016/j.neuroimage.2007.08.028>.
- [156] F. Rheault, A. De Benedictis, A. Daducci, C. Maffei, C.M.W. Tax, D. Romascano, E. Caverzasi, F.C. Morency, F. Corrivetti, F. Pestilli, G. Girard, G. Theaud, I. Zemmoura, J. Hau, K. Glavin, K.M. Jordan, K. Pomiecko, M.

- Chamberland, M. Barakovic, N. Goyette, P. Poulin, Q. Chenot, S.S. Panesar, S. Sarubbo, L. Petit, M. Descoteaux, Tractostorm: The what, why, and how of tractography dissection reproducibility, *Hum Brain Mapp* 41 (2020) 1859–1874. <https://doi.org/10.1002/hbm.24917>.
- [157] H.-H. Chang, A.H. Zhuang, D.J. Valentino, W.-C. Chu, Performance measure characterization for evaluating neuroimage segmentation algorithms, *Neuroimage* 47 (2009) 122–135. <https://doi.org/10.1016/j.neuroimage.2009.03.068>.
- [158] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, *Med Image Anal* 42 (2017) 60–88. <https://doi.org/10.1016/j.media.2017.07.005>.
- [159] D.L. G Hill, P.G. Batchelor, M. Holden, D.J. Hawkes, Medical image registration, 2001. https://helios2.mi.parisdescartes.fr/~lomn/Cours/CV/BME/Registration/registration6_MedicalImageRegistration.pdf. (accessed November 27, 2024).
- [160] E. Henriksson, E. Kjellen, P. Wahlberg, T. Ohlsson, J. Wennerberg, E. Brun, 2-Deoxy-2-[18F]Fluoro-D-Glucose Uptake and Correlation to Intratumoral Heterogeneity, *Anticancer Res* 27 (2007) 21552160.
- [161] R. Boellaard, Standards for PET image acquisition and quantitative data analysis, *Journal of Nuclear Medicine* 50 (2009) 11S–20S. <https://doi.org/10.2967/jnumed.108.057182>.
- [162] J.J.M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H.J.W.L. Aerts, Computational Radiomics System to Decode the Radiographic Phenotype, *Cancer Res* 77 (2017) e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>.
- [163] S.S.F. Yip, H.J.W.L. Aerts, Applications and limitations of radiomics, *Phys Med Biol* 61 (2016) R150–R166. <https://doi.org/10.1088/0031-9155/61/13/R150>.
- [164] S. Chicklore, V. Goh, M. Siddique, A. Roy, P.K. Marsden, G.J.R. Cook, Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis, *Eur J Nucl Med Mol Imaging* 40 (2013) 133–140. <https://doi.org/10.1007/s00259-012-2247-0>.
- [165] J.J. Foy, K.R. Robinson, H. Li, M.L. Giger, H. Al-Hallaq, S.G. Armato, Variation in algorithm implementation across radiomics software, *Journal of Medical Imaging* 5 (2018) 1. <https://doi.org/10.1117/1.jmi.5.4.044505>.

- [166] M. Hatt, F. Tixier, L. Pierce, P.E. Kinahan, C.C. Le Rest, D. Visvikis, Characterization of PET/CT images using texture analysis: the past, the present... any future?, *Eur J Nucl Med Mol Imaging* 44 (2017) 151–165. <https://doi.org/10.1007/s00259-016-3427-0>.
- [167] A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J.W.L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G.J.R. Cook, C. Davatzikos, A. Depeursinge, M.C. Desseroit, N. Dinapoli, C.V. Dinh, S. Echegaray, I. El Naqa, A.Y. Fedorov, R. Gatta, R.J. Gillies, V. Goh, M. Götz, M. Guckenberger, S.M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R.T.H. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K.H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E.A.G. Pfaehler, A. Rahmim, A.U.K. Rao, J. Scherer, M.M. Siddique, N.M. Sijtsema, J. Socarras Fernandez, E. Spezi, R.J.H.M. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E.G.C. Troost, T. Upadhaya, V. Valentini, L. V. van Dijk, J. van Griethuysen, F.H.P. van Velden, P. Whybra, C. Richter, S. Löck, The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping, *Radiology* 295 (2020) 328–338. <https://doi.org/10.1148/radiol.2020191145>.
- [168] J.J.M. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.C. Fillion-Robin, S. Pieper, H.J.W.L. Aerts, Computational radiomics system to decode the radiographic phenotype, *Cancer Res* 77 (2017) e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>.
- [169] J. Yan, J.L. Chu-Shern, H.Y. Loi, L.K. Khor, A.K. Sinha, S.T. Quek, I.W.K. Tham, D. Townsend, Impact of Image Reconstruction Settings on Texture Features in ^{18}F -FDG PET, *Journal of Nuclear Medicine* 56 (2015) 1667–1673. <https://doi.org/10.2967/jnumed.115.156927>.
- [170] Image Biomarker Standardisation Initiative, Image Processing, (n.d.). https://ibsi.readthedocs.io/en/latest/02_Image_processing.html (accessed October 28, 2024).
- [171] R.T.H. Leijenaar, G. Nalbantov, S. Carvalho, W.J.C. Van Elmpt, E.G.C. Troost, R. Boellaard, H.J.W.L. Aerts, R.J. Gillies, P. Lambin, The effect of SUV discretization in quantitative FDG-PET Radiomics: The need for standardized methodology in tumor texture analysis, *Sci Rep* 5 (2015). <https://doi.org/10.1038/srep11075>.

- [172] Radiomics, Features — PyRadiomics documentation., (n.d.). <https://pyradiomics.readthedocs.io/en/latest/features.html> (accessed October 28, 2024).
- [173] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, Image biomarker standardisation initiative: Reference manual, (2016). https://ibsi.readthedocs.io/en/latest/02_Image_processing.html. (accessed November 27, 2024).
- [174] I. Haralick, R.M., Shanmugam, Dinstein, Textural features for Image Classification, IEEE Trans Syst Man Cybern SMC-3 (1973) 610–621. <https://doi.org/10.1109/TSMC.1973.4309314>.
- [175] C. Sun, W.G. Wee, Neighboring gray level dependence matrix for texture classification, Comput Vis Graph Image Process 23 (1983) 341–352. [https://doi.org/10.1016/0734-189X\(83\)90032-4](https://doi.org/10.1016/0734-189X(83)90032-4).
- [176] M.M. Galloway, Texture analysis using gray level run lengths, Computer Graphics and Image Processing 4 (1975) 172–179. [https://doi.org/10.1016/S0146-664X\(75\)80008-6](https://doi.org/10.1016/S0146-664X(75)80008-6).
- [177] G. Thibault, J. Angulo, F. Meyer, Advanced Statistical Matrices for Texture Characterization: Application to Cell Classification, IEEE Trans Biomed Eng 61 (2014) 630–637. <https://doi.org/10.1109/TBME.2013.2284600>.
- [178] M. Amadasun, R. King, Textural features corresponding to textural properties, IEEE Trans Syst Man Cybern 19 (1989) 1264–1274. <https://doi.org/10.1109/21.44046>.
- [179] D.L. Campbell, H. Kang, S. Shokouhi, Application of Hharalick texture features in brain [18F]-florbetapir positron emission tomography without reference region normalization, Clin Interv Aging 12 (2017) 2077–2086. <https://doi.org/10.2147/CIA.S143307>.
- [180] S.S. Shapiro, M.B. Wilk, An Analysis of Variance Test for Normality (Complete Samples), Biometrika 52 (1965) 591. <https://doi.org/10.2307/2333709>.
- [181] J.P. Royston, An Extension of Shapiro and Wilk’s W Test for Normality to Large Samples, Appl Stat 31 (1982) 115. <https://doi.org/10.2307/2347973>.
- [182] H.B. Mann, D.R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, The Annals of Mathematical Statistics 18 (1947) 50–60. <https://doi.org/10.1214/aoms/1177730491>.
- [183] B. Efron, T. Hastie, Computer Age Statistical Inference: Algorithms, Evidence, and Data Science., Cambridge University Press, Cambridge, UK, 2016.

-
- [184] W.H. Kruskal, W.A. Wallis, Use of Ranks in One-Criterion Variance Analysis, *J Am Stat Assoc* 47 (1952) 583. <https://doi.org/10.2307/2280779>.
- [185] G.H. McDonald, One-way ANOVA, in: Sparky House Publishing (Ed.), *Handbook of Biological Statistics*, 2014. <http://www.biostathandbook.com/onewayanova.html> (accessed October 28, 2024).
- [186] S.Y. Chen, Z. Feng, X. Yi, A general introduction to adjustment for multiple comparisons, *J Thorac Dis* 9 (2017) 1725–1729. <https://doi.org/10.21037/jtd.2017.05.34>.
- [187] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *J R Stat Soc Series B Stat Methodol* 57 (1995) 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- [188] K. Korthauer, P.K. Kimes, C. Duvallet, A. Reyes, A. Subramanian, M. Teng, C. Shukla, E.J. Alm, S.C. Hicks, A practical guide to methods controlling false discoveries in computational biology, *Genome Biol* 20 (2019). <https://doi.org/10.1186/s13059-019-1716-1>.
- [189] K. Pearson, Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia, *Philos Trans R Soc Lond A* 186 (1895) 343–414. <https://doi.org/10.1098/rsta.1895.0010>.
- [190] T. Van Der Ploeg, P.C. Austin, E.W. Steyerberg, Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints, *BMC Med Res Methodol* 14 (2014). <https://doi.org/10.1186/1471-2288-14-137>.
- [191] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem., in: T. Theeramunkong, B. Kijsirikul, N. Cercone, T. Ho (Eds.), *Advances in Knowledge Discovery and Data Mining. PAKDD 2009.*, Springer, Berlin, Heidelberg, 2009: pp. 475–482. https://doi.org/10.1007/978-3-642-01307-2_43.
- [192] N. V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research* 16 (2002) 321–357. <https://doi.org/10.1613/jair.953>.
- [193] E. Horvitz, D. Mulligan, Data, privacy, and the greater good, *Science* (1979) 349 (2015) 253–255. <https://doi.org/10.1126/science.aac4520>.
- [194] T.M.. Mitchell, *Machine learning*, McGraw-Hill, New York, 1997.

- [195] B. Farnham, S. Tokyo, B. Boston, F. Sebastopol, T. Beijing, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION, 2019.
- [196] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, NY, 2006.
- [197] Scikit-learn Developers, User Guide — scikit-learn 1.0.2 documentation, (2024). https://scikit-learn.org/stable/user_guide.html (accessed October 29, 2024).
- [198] F. Pedregosa, G. Varoquaux, A. Gramfort, B. Michel V. and Thirion, O. Grisel, M. Blondel, R. Prettenhofer P. and Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. <https://scikit-learn.org/stable/> (accessed November 27, 2025).
- [199] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984.
- [200] C. Cortes, V. Vapnik, L. Saitta, Support-Vector Networks Editor, Kluwer Academic Publishers, 1995.
- [201] L. Breiman, Random Forests, Mach Learn 45 (2001) 5–32. <https://doi.org/10.1023/A:1010933404324>.
- [202] J.H. Friedman, Greedy function approximation: A gradient boosting machine., The Annals of Statistics 29 (2001). <https://doi.org/10.1214/aos/1013203451>.
- [203] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer New York, NY, Nueva York, NY, EE. UU., 2009. <https://doi.org/10.1007/978-0-387-84858-7>.
- [204] J. Berkson, Application to the Logistic Function to Bio-Assay, J Am Stat Assoc 39 (1944) 357. <https://doi.org/10.2307/2280041>.
- [205] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, IEEE Trans Pattern Anal Mach Intell 5 (1973) 4–4. <https://doi.org/10.2307/1573081>.
- [206] A. McCallum, K. Nigam, A Comparison of Event Models for Naive Bayes Text Classification, in: Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, AAAI Press, 1998: pp. 41–48. <https://www.semanticscholar.org/paper/A-comparison-of-event-models-for-naive-bayes-text-McCallum-Nigam/04ce064505b1635583fa0d9cc07cac7e9ea993cc> (accessed November 27, 2024).

-
- [207] C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, 2008.
 - [208] T. Cover, P. Hart, Nearest neighbor pattern classification, *IEEE Trans Inf Theory* 13 (1967) 21–27. <https://doi.org/10.1109/TIT.1967.1053964>.
 - [209] G.E. Hinton, Connectionist Learning Procedures, *Artif Intell* 40 (1989) 185–234. [https://doi.org/10.1016/0004-3702\(89\)90049-0](https://doi.org/10.1016/0004-3702(89)90049-0).
 - [210] C. El Morr, M. Jammal, H. Ali-Hassan, W. El-Hallak, *Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics*, Springer, Cham, Suiza, 2022.
 - [211] Cross-validation: evaluating estimator performance — scikit-learn 1.6.0 documentation, (n.d.). https://scikit-learn.org/stable/modules/cross_validation (accessed December 13, 2024).
 - [212] A. Kebaili, J. Lapuyade-Lahorgue, S. Ruan, Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review, *J Imaging* 9 (2023). <https://doi.org/10.3390/jimaging9040081>.
 - [213] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, B.C. Van Esesn, A.A.S. Awwal, V.K. Asari, The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches, *ArXiv Preprint ArXiv:1803.01164* (2018). <https://doi.org/10.48550/arXiv.1803.01164>.
 - [214] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (2015) 436–444. <https://doi.org/10.1038/nature14539>.
 - [215] K. Fukushima, Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition, *IEEE Trans Syst Man Cybern* 18 (1988) 1–6. [https://doi.org/10.1016/0893-6080\(88\)90014-7](https://doi.org/10.1016/0893-6080(88)90014-7).
 - [216] A. Sarraf, M. Azhdari, S. Sarraf, A Comprehensive Review of Deep Learning Architectures for Computer Vision Applications, *American Scientific Research Journal for Engineering, Technology, and Sciences* 77 (2021) 1–29.
 - [217] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: F. Pereira, C.J. Burges, L. Bottou, K.Q. Weinberger (Eds.), *Adv Neural Inf Process Syst*, Curran Associates, Inc., 2012: pp. 1097–1105. <https://doi.org/10.1145/3065386>.
 - [218] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, *ArXiv Preprint ArXiv:1409.1556* (2014). <https://doi.org/10.48550/arXiv.1409.1556>.

- [219] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: pp. 770–778. <http://image-net.org/challenges/LSVRC/2015/>.
- [220] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, Proceedings of the AAAI Conference on Artificial Intelligence 31 (2017). <https://doi.org/10.1609/aaai.v31i1.11231>.
- [221] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: pp. 4700–4708.
- [222] A. Amyar, S. Ruan, I. Gardin, C. Chatelain, P. Decazes, R. Modzelewski, 3-D RPET-NET: Development of a 3-D PET Imaging Convolutional Neural Network for Radiomics Analysis and Outcome Prediction, IEEE Trans Radiat Plasma Med Sci 3 (2019) 225–231. <https://doi.org/10.1109/TRPMS.2019.2896399>.
- [223] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, ArXiv Preprint ArXiv:1412.6980 (2014). <https://doi.org/10.48550/arXiv.1412.6980>.
- [224] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (2015). <https://doi.org/10.1371/journal.pone.0118432>.
- [225] J. Davis, M. Goadrich, The Relationship Between Precision-Recall and ROC Curves, in: International Conference on Machine Learning, Pittsburgh, PA, 2006. <https://doi.org/10.1145/1143844.1143874>.
- [226] H. Kim, Y.D. Jeon, K.B. Park, H. Cha, M.-S. Kim, J. You, S.-W. Lee, S.-H. Shin, Y.-G. Chung, S. Bin Kang, W.S. Jang, D.-K. Yoon, Automatic segmentation of inconstant fractured fragments for tibia/fibula from CT images using deep learning, Sci Rep 13 (2023) 20431. <https://doi.org/10.1038/s41598-023-47706-4>.
- [227] B. Li, C. Liu, S. Wu, G. Li, Verte-Box: A Novel Convolutional Neural Network for Fully Automatic Segmentation of Vertebrae in CT Image, Tomography 8 (2022) 45–58. <https://doi.org/10.3390/tomography8010005>.
- [228] H.A. Khan, M.A. Haider, H.A. Ansari, H. Ishaq, A. Kiyani, K. Sohail, M. Muhammad, S.A. Khurram, Automated feature detection in dental periapical radiographs by using deep learning, Oral Surg Oral Med Oral Pathol Oral Radiol 131 (2021) 711–720. <https://doi.org/10.1016/j.oooo.2020.08.024>.

-
- [229] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O.R. Zaiane, M. Jagersand, U2-Net: Going deeper with nested U-structure for salient object detection, *Pattern Recognit* 106 (2020) 107404. <https://doi.org/10.1016/j.patcog.2020.107404>.
 - [230] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation, in: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020: pp. 1055–1059. <https://doi.org/10.1109/ICASSP40776.2020.9053405>.
 - [231] M.Z. Alom, M. Hasan, C. Yakopcic, T.M. Taha, V.K. Asari, Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation, *ArXiv Preprint ArXiv:1802.06955* (2018). <https://doi.org/10.48550/arXiv.1802.06955>.
 - [232] A.B. Hassanat, A.S. Tarawneh, G.A. Altarawneh, A. Almuhaimeed, Stop Oversampling for Class Imbalance Learning: A Critical Review, *ArXiv Preprint ArXiv:1802.06955* (2022). <https://doi.org/10.48550/arXiv.1802.06955>.
 - [233] R. Van Den Goorbergh, M. Van Smeden, D. Timmerman, Ben Van Calster, The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression, *Journal of the American Medical Informatics Association* 29 (2022) 1525–1534. <https://doi.org/10.1093/jamia/ocac093>.
 - [234] E. Milara, A. Gómez-Grande, S. Tomás-Soler, A.P. Seiffert, R. Alonso, E.J. Gómez, J. Martínez-López, P. Sánchez-González, Bone marrow segmentation and radiomics analysis of [18F]FDG PET/CT images for measurable residual disease assessment in multiple myeloma, *Comput Methods Programs Biomed* 225 (2022) 107083. <https://doi.org/10.1016/J.CMPB.2022.107083>.
 - [235] E. Milara, R. Alonso, L. Maseing, A.P. Seiffert, A. Gómez-Grande, E.J. Gómez, J. Martínez-López, P. Sánchez-González, Radiomics analysis of bone marrow biopsy locations in [18F]FDG PET/CT images for measurable residual disease assessment in multiple myeloma, *Phys Eng Sci Med* 46 (2023) 903–913. <https://doi.org/10.1007/s13246-023-01265-0>.
 - [236] F. Song, Z. Guo, D. Mei, Feature Selection Using Principal Component Analysis, in: *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, IEEE, 2010: pp. 27–30. <https://doi.org/10.1109/ICSEM.2010.14>.
 - [237] N. Pudjihartono, T. Fadason, A.W. Kempa-Liehr, J.M. O’Sullivan, A Review of Feature Selection Methods for Machine Learning-Based Disease Risk

- Prediction, *Frontiers in Bioinformatics* 2 (2022).
<https://doi.org/10.3389/fbinf.2022.927312>.
- [238] E. Milara, P. Sarandeses, A. Gómez-Grande, E.J. Gómez, P. Sánchez-González, Propuesta de modelo basado en redes neuronales para la clasificación de regiones del esqueleto a partir de imágenes TC, in: *XLI Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Cartagena, 2023: pp. 298–301.
- [239] L. Rodríguez-Luque, E. Milara, A. Jiménez-Ubieto, D. Moreno-Blanco, A. Saviatto, M. Poza, E.J. Gómez, A. Gómez-Grande, P. Sarandeses, P. Sánchez-González, Modelo de clasificación de médula ósea en pacientes diagnosticados con linfoma no Hodgkin mediante la cuantificación de médula, bazo e hígado en imágenes [18F]FDG PET/CT, in: *XLII Congreso Anual de La Sociedad Española de Ingeniería Biomédica*, Sevilla, 2024: pp. 325–328.

Annex A. Hyperparameters grid

All information regarding the different hyperparameters and their possible values can be found in the official scikit-learn² documentation.

Decision Tree

```
param_grid = dict(
    criterion = ['gini', 'entropy'],
    splitter = ['best', 'random'],
    max_depth = [None, 1, 3, 5, 10, 15],
    min_samples_split = [0.1, 0.3, 0.5, 0.7, 0.9],
    min_samples_leaf = [1, 5, 10, 20],
    class_weight = ['balanced'])
```

Support Vector Machines – kernel RBF

```
param_grid = dict(
    C = [0.1, 0.5, 1, 5, 10],
    gamma = ['scale', 'auto', 0.001, 0.01, 0.1, 0.5, 0.9],
    tol = [0.00001, 0.0001, 0.001, 0.01, 0.1],
    max_iter = [-1, 500, 1000, 2000, 5000],
    class_weight = ['balanced'],
    kernel = ['rbf'])
```

Support Vector Machines – kernel Linear for two classes

```
param_grid = [
    {'penalty' = ['l2'], 'C' = [0.1, 0.5, 1, 2, 5, 10], 'loss' = ['hinge', 'squared_hinge'],
     'tol' = [0.00001, 0.0001, 0.001, 0.01, 0.1],
     'max_iter' = [500, 1000, 2000, 5000, 7000, 10000], 'class_weight' = ['balanced']},
    {'penalty' = ['l1'], 'C' = [0.1, 0.5, 1, 2, 5, 10], 'loss' = ['hinge'],
     'tol' = [0.00001, 0.0001, 0.001, 0.01, 0.1],
     'max_iter' = [500, 1000, 2000, 5000, 7000, 10000], 'class_weight' = ['balanced']}]
```

² <https://scikit-learn.org/stable/>

Support Vector Machines – kernel Linear for three classes

```
param_grid = [  
    {'penalty' = ['l2'], 'C' = [0.1, 0.5, 1, 2, 5, 10], 'loss' = ['hinge', 'squared_hinge'],  
    'tol' = [0.00001, 0.0001, 0.001, 0.01, 0.1],  
    'max_iter' = [500, 1000, 2000, 5000, 7000, 10000], 'class_weight' = ['balanced'],  
    'multi_class' = ['ovr']}]
```

Support Vector Machines – kernel sigmoidal

```
param_grid = dict(  
    C = [0.1, 0.5, 1, 5, 10],  
    gamma = ['scale', 'auto', 0.001, 0.01, 0.1, 0.5, 0.9],  
    tol = [0.00001, 0.0001, 0.001, 0.01, 0.1],  
    max_iter = [-1, 500, 1000, 2000, 5000],  
    coef0 = [0, 0.1, -0.1, 0.5, -0.5, 1, -1],  
    class_weight = ['balanced'],  
    kernel = ['sigmoid'])
```

Random Forest

```
param_grid = dict(  
    n_estimators = [10, 100, 500, 1000, 2000],  
    criterion = ['gini', 'entropy'],  
    max_depth = [None, 5, 10],  
    min_samples_split = [0.1, 0.3],  
    min_samples_leaf = [2, 10],  
    max_features = [5, 10],  
    class_weight = ['balanced'])
```

Gradient Boosting

```
param_grid = dict(
    n_estimators = [10, 100, 500, 1000],
    max_depth = [None, 1, 5, 10],
    min_samples_split = [0.1, 0.5, 0.9],
    min_samples_leaf = [1, 10, 20],
    max_features = ['sqrt', 'log2'],
    n_iter_no_change = [10],
    validation_fraction = [0.1, 0.2, 0.3],
    tol = [0.001])
```

Logistic Regression

```
param_grid = [
    {'penalty': ['l1', 'l2'], 'C': [0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10], 'solver': ['liblinear'],
     'max_iter': [100000], 'class_weight': ['balanced']},
    {'penalty': [None], 'solver': ['lbfgs', 'saga'],
     'max_iter': [100000], 'class_weight': ['balanced']},
    {'penalty': ['l2'], 'C': [0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10], 'solver': ['lbfgs'],
     'max_iter': [100000], 'class_weight': ['balanced']},
    {'penalty': ['l1', 'l2'], 'C': [0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10], 'solver': ['saga'],
     'max_iter': [100000], 'class_weight': ['balanced']},
    {'penalty': ['elasticnet'], 'C': [0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10], 'solver': ['saga'],
     'max_iter': [100000], 'l1_ratio': [0.25, 0.5, 0.75, 1], 'class_weight': ['balanced']}
```

Naïve Bayes for two classes

```
param_grid = dict(
    alpha=[1e-10, 0.00001, 0.001, 0.1, 0.5, 0.9, 1, 1.1, 1.5, 2],
    binarize = [None],
    class_prior= [[0.8, 0.2], [0.75, 0.25], [0.7, 0.3], [0.65, 0.35], [0.6, 0.4], [0.55, 0.45],
                  [0.5, 0.5], [0.4, 0.6]])
```

Naïve Bayes for three classes

```
param_grid = dict(
    alpha=[1e-10, 0.00001, 0.001, 0.1, 0.5, 0.9, 1, 1.1, 1.5, 2],
    binarize = [None],
    class_prior= [[0.8, 0.1, 0.1], [0.8, 0.15, 0.05], [0.7, 0.2, 0.1], [0.7, 0.15, 0.15], [0.65,
                  0.2, 0.15], [0.6, 0.2, 0.2], [0.6, 0.2, 0.2], [0.6, 0.3, 0.1], [0.55, 0.25, 0.2]])
```

k-Nearest Neighbours

```
param_grid = [  
    {'algorithm' = ['ball_tree', 'kd_tree'], 'n_neighbors' = [5, 10, 15, 20],  
    'weights' = ['uniform', 'distance'], 'leaf_size' = [10, 20, 30, 40, 50],  
    'p' = [1, 2, 4, 6, 8, 10]},  
    {'algorithm' = ['brute'], 'n_neighbors' = [5, 10, 15, 20],  
    'weights' = ['uniform', 'distance'], 'p' = [1, 2, 4, 6, 8, 10]}]
```

Multi-Layer Perceptron with $X = 100$ and 200 neurons in the hidden layer as different models

```
param_grid = dict(  
    hidden_layer_sizes = [X],  
    activation = ['logistic', 'tanh', 'relu'],  
    solver = ['adam', 'lbfgs'],  
    alpha = [0.01, 0.001, 0.0001],  
    learning_rate = ['constant', 'invscaling'],  
    max_iter = [100, 200, 500, 1000],  
    tol = [0.0001, 0.001, 0.01, 0.1],  
    early_stopping = [True],  
    validation_fraction = [0.1, 0.3, 0.5, 0.7, 0.9],  
    n_iter_no_change = [10])
```


Annex B. Instructions of the automatic segmentation tab

In this Annex, the instructions for the ‘Pre-processing’ tab described in 5.1.1.1.1. *Pre-processing tab* are included. First, in Figure 105, the explanation of how to start the segmentation process is described.

The screenshot shows a window titled 'Preprocessing tab instructions' with logos for GBT, ETSIT UPM, CTB (centro de tecnología biomédica), and Hospital Universitario 12 de Octubre. The main heading is 'PREPROCESSING TAB INSTRUCTIONS - BONE AND MARROW AUTOMATIC SEGMENTATION'. Below this is a section titled 'Automatic Segmentation Process (1/3)' containing four numbered instructions:

1. Press the **Select Subjects Folder** button and select the *Subjects* folder containing folders of different subjects, each of them with the PET and CT protocols formed by a set of DICOMs. *Example path:* 'Subjects/Patient1/PET/dicom1.dcm' or 'Subjects/Patient1/CT/dicom1.dcm'.
2. Select the clinical context of the subject folder selected. Dividing subjects folders depending on the clinical context is recommended. Different possibilities are:
 - **Measurable Residual Disease (MRD)**, for studying the possible relapse of Multiple Myeloma patients in the post-treatment evaluation images.
 - **Multiple Myeloma (MM)**, for studying the possible relapse or MRD from their diagnosis images.
 - **Smoldering (SMM)**, for studying the level of risk (high, medium or low).
 - **Non-Hodgkin Lymphoma (NHL)**, for studying the possible bone marrow involvement.
3. Press **Select/Create Results Folder** button and select the folder *Results*. The first time, a individual Folder for each clinical context is created. (DO NOT select other *Results* folder the rest of times if you want to continue a segmentation process in the same subjects folder). If you do not have a *Results* folder in the selected folder, a new one will be created. It is recommended not to create the *Results* folder inside the *Subjects* folder to avoid interpretation problems.
4. Click the **Segment images** button to start segmentation. OPTIONAL: mark the check box below to perform the segmentation of the cases that have already been previously segmented. The processed cases will contain all the necessary files to be able to perform manual editing in the 'Editing' tab. NOT RECOMMENDED: mark the check box to skip the slices location manual process only if you have revised them previously.

Below the instructions is a form with the following fields and options:

- Select Subjects Folder**: A button.
- Clinical Context**: A dropdown menu currently showing 'Measurable Residual Disease (MRD)' with a list of options: 'Measurable Residual Disease (MRD)', 'Multiple Myeloma (MM)', 'Smoldering (SMM)', and 'Non-Hodgkin Lymphoma (NHL)'.
- Select/Create Results Folder**: A button.
- Segment Images**: A button.
- ☐ Mark if you want to resegment subjects previously segmented.
- Non-processed cases:** A label.

At the bottom, there is a note: 'Cases for which segmentation was not possible will be displayed at the bottom, next to the *Non-processed cases* label. It is recommended to check if these cases contain only the mentioned protocols and dicom files inside their folders. REMEMBER NOT TO USE MATLAB DURING THE SEGMENTATION PROCESS.'

Navigation buttons at the bottom right: 'Previous' and 'Next (1/3)'.

Figure 105. Instructions of the ‘Pre-processing’ tab: I.

Secondly, in Figure 106, the explanation about the selection of the slices, only for the platform when the DL model is included as segmentation method, is included.

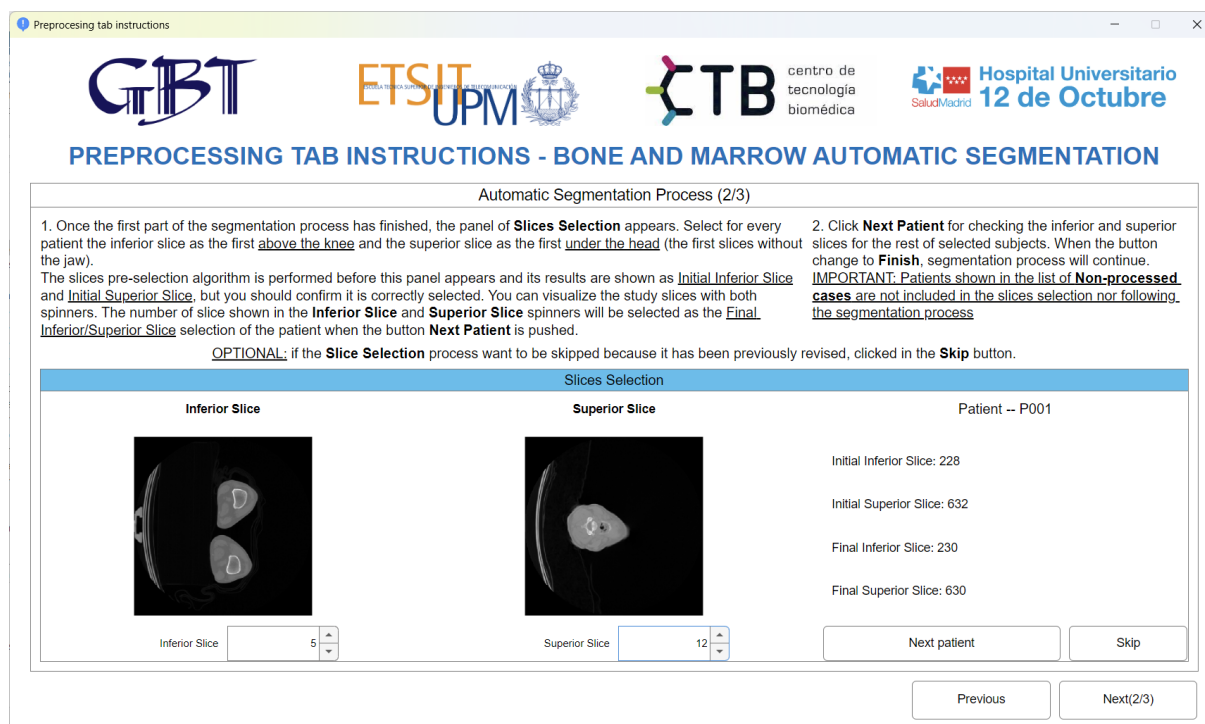


Figure 106. Instructions of the 'Pre-processing' tab: II.

Finally, the explanation of the different information panels, and the panel for the segmentation process are included in Figure 107.

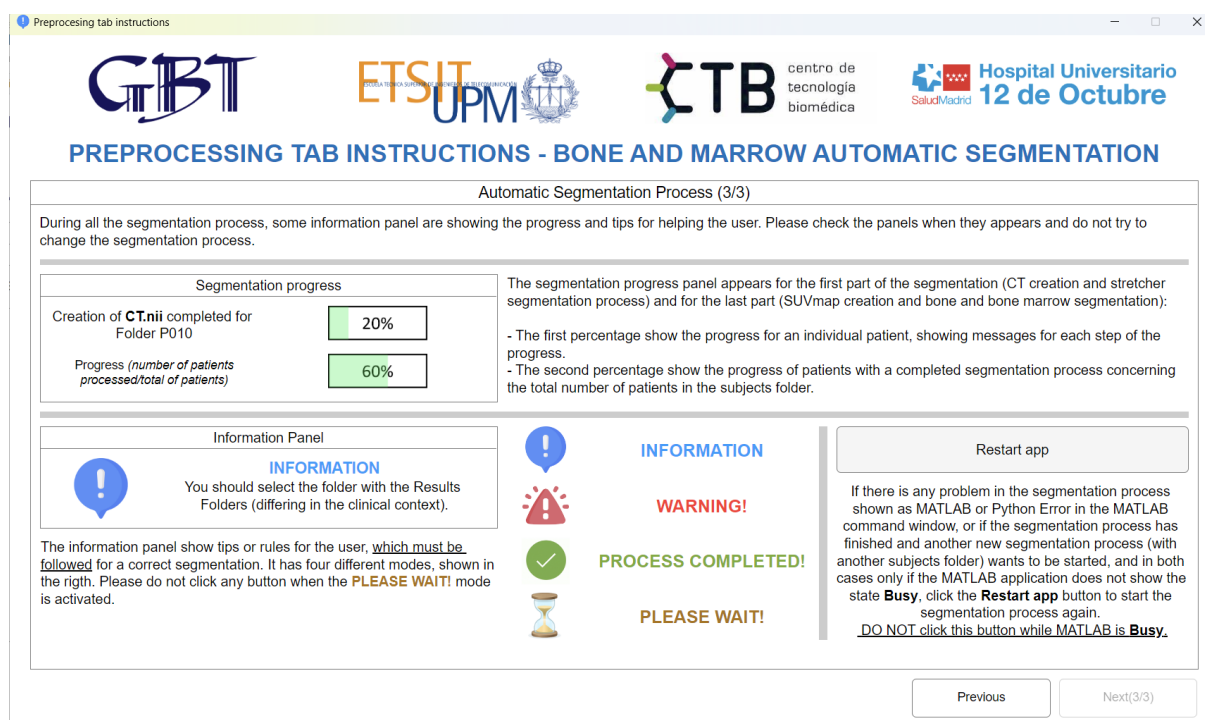


Figure 107. Instructions of the 'Pre-processing' tab: III.

For the ‘Segmentation’ tab, the first part of the instruction window is substituted by the one shown in Figure 108. However, the other two parts of the instruction window are the same as the ones shown in Figure 106 and Figure 107.

Preprocessing tab instructions

GBT **ETSIT UPM** **CTB** centro de tecnología biomédica **Hospital Universitario 12 de Octubre** SaludMadrid

PREPROCESSING TAB INSTRUCTIONS - BONE AND MARROW AUTOMATIC SEGMENTATION

Automatic Segmentation Process (1/3)

1. Mark the check box **Bone** to obtain the skeleton mask. Mark the check **Bone marrow** to obtain the bone marrow mask. You can selected both or only one of them. The new segmentations will be saved in the same folder where the *CT.mat* file is found.
2. Press the **Select Patients Folder (with CT.mat)** and select the folder which contains the patients folder (Example: 'C:/Desktop/BBDD/Patients/PXXX'). Make sure each patient folder contain a whole-body *CT.mat*.
3. Click the **Segment images** button to start segmentation. **OPTIONAL**: mark the check box below to perform the segmentation of the cases that have already been previously segmented. The processed cases will contain all the necessary files to be able to perform manual editing in the 'Editing' tab.

NOT RECOMMENDED: mark the check box to skip the slices location manual process only if you have revised them previously.

Mask to segment: ☒ Bone ☒ Bone marrow

Select Patients Folder (with CT.mat) Segment images

☐ Mark if you want to resegment subjects previously segmented

☐ Skip manual slices location process (NOT RECOMMENDED)

Non-processed cases:

Cases for which segmentation was not possible will be displayed at the bottom, next to the *Non-processed cases* label. It is recommended to check if these cases contain the CT.mat file.
REMEMBER NOT TO USE MATLAB DURING THE SEGMENTATION PROCESS.

Previous Next(1/3)

Figure 108. Instructions of the ‘Segmentation’ tab: I.

Annex C. Extra information of the Editing tab

In this Annex extra images not essential for the explanation of the ‘Editing’ tab are included. In Figure 109, an example of the complete BM volume with the map of SUV values overlayed in the ‘Volume’ emergent window is shown.

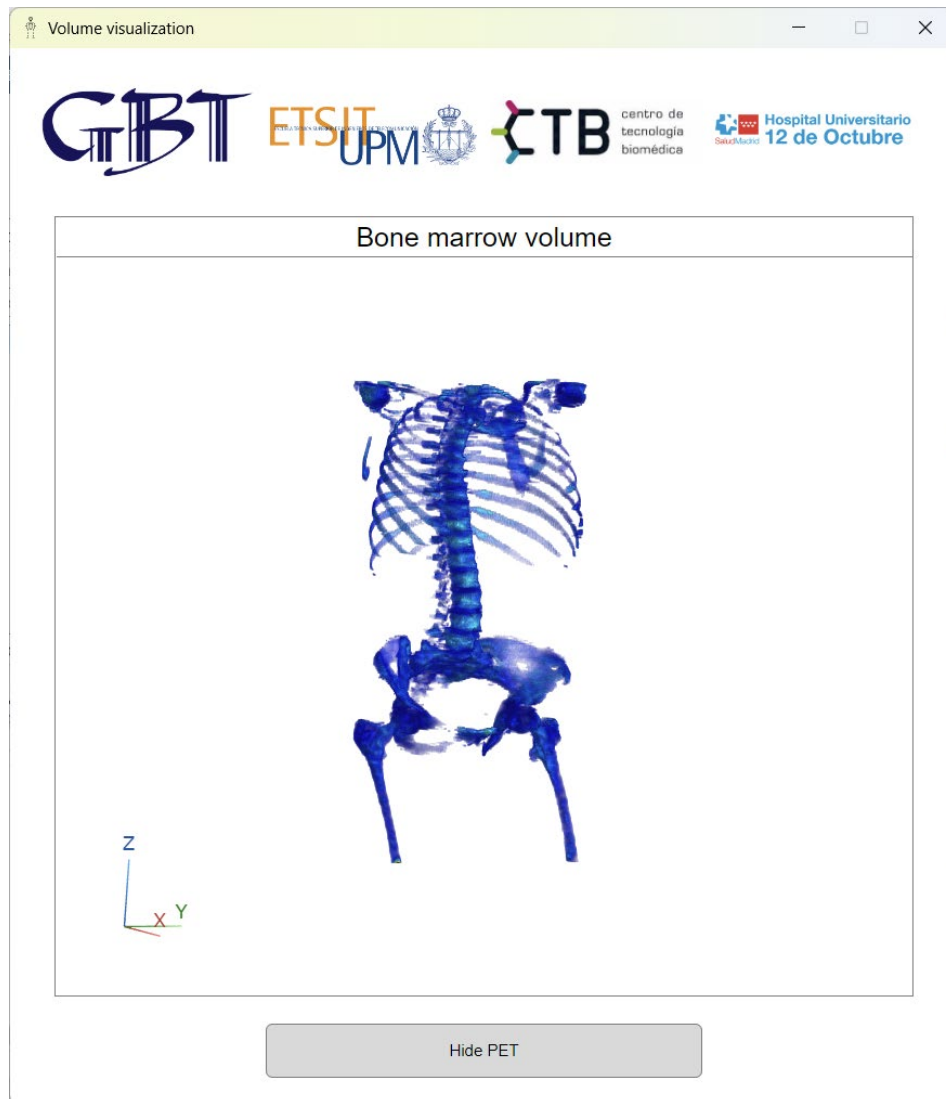


Figure 109. ‘Editing’ tab of the semi-automatic segmentation platform: ‘Volume’ window with PET overlayed.

Figure 110 shows the warning messages which appear in an informative emergent window when an invalid slice value is selected.

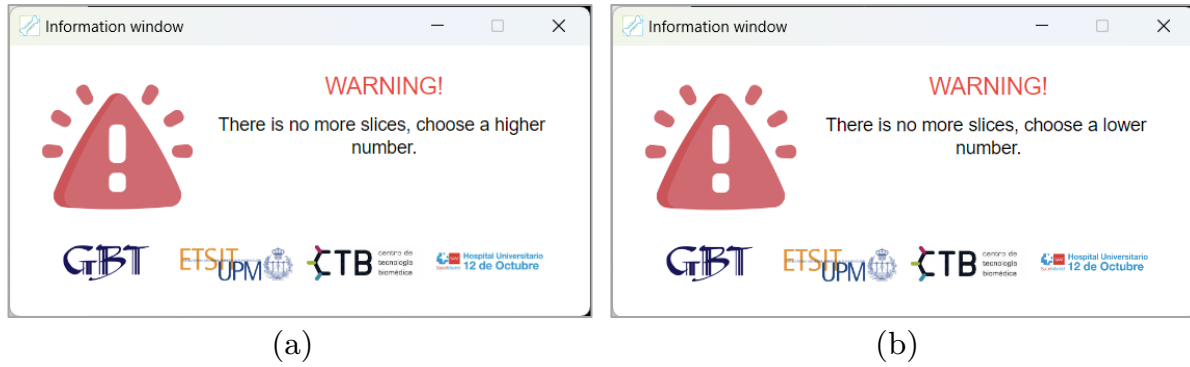


Figure 110. 'Editing' tab of the semi-automatic segmentation platform: 'Information' window with a warning message, result of selecting (a) a lower than 1 for the slices and (b) a higher than the maximum CT and SUVmap slices.

Figure 111 shows the 'Extra Functions Window' when the 'Visualizing equidistant slices from the first to the last slices with mask' option is selected.

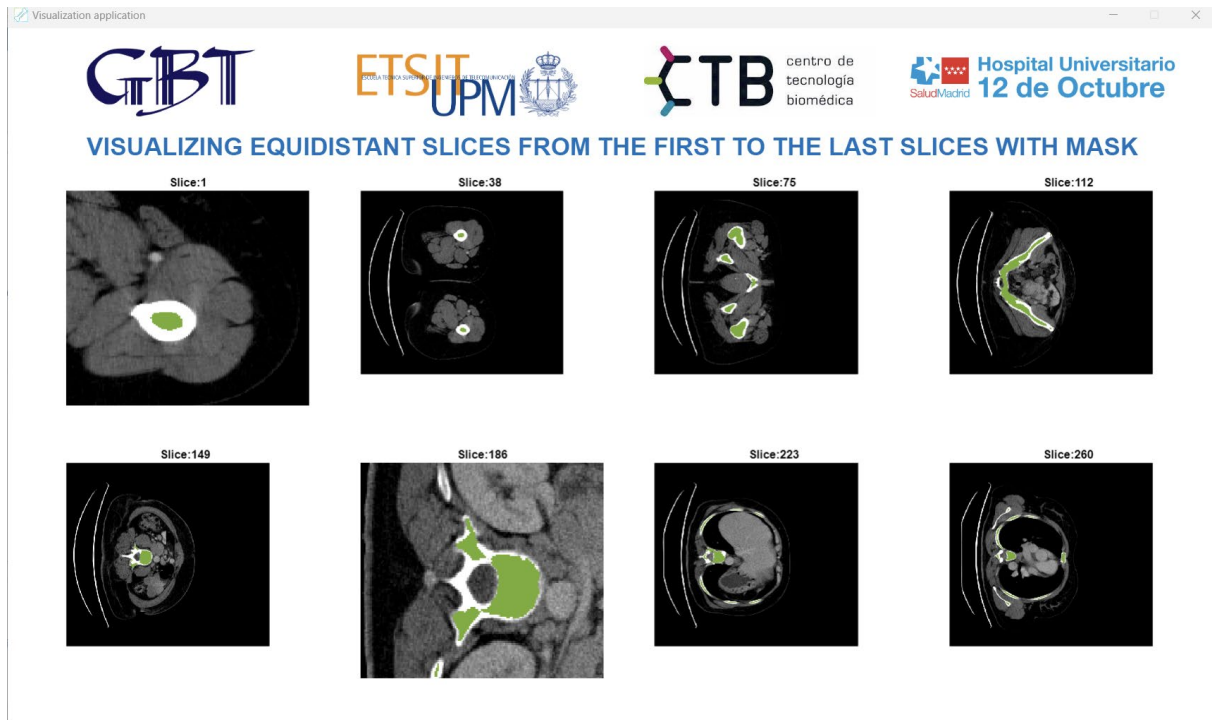


Figure 111. 'Editing' tab of the semi-automatic segmentation platform: 'Extra Functions Window', visualization of slices function.

Annex D. Instructions of the Editing tab

This annex includes the different windows of instructions for the ‘Editing’ tab, divided in instructions of the main window (Figure 112), instructions of the visualization settings window (Figure 113) and instructions of the extra functions window (Figure 114).

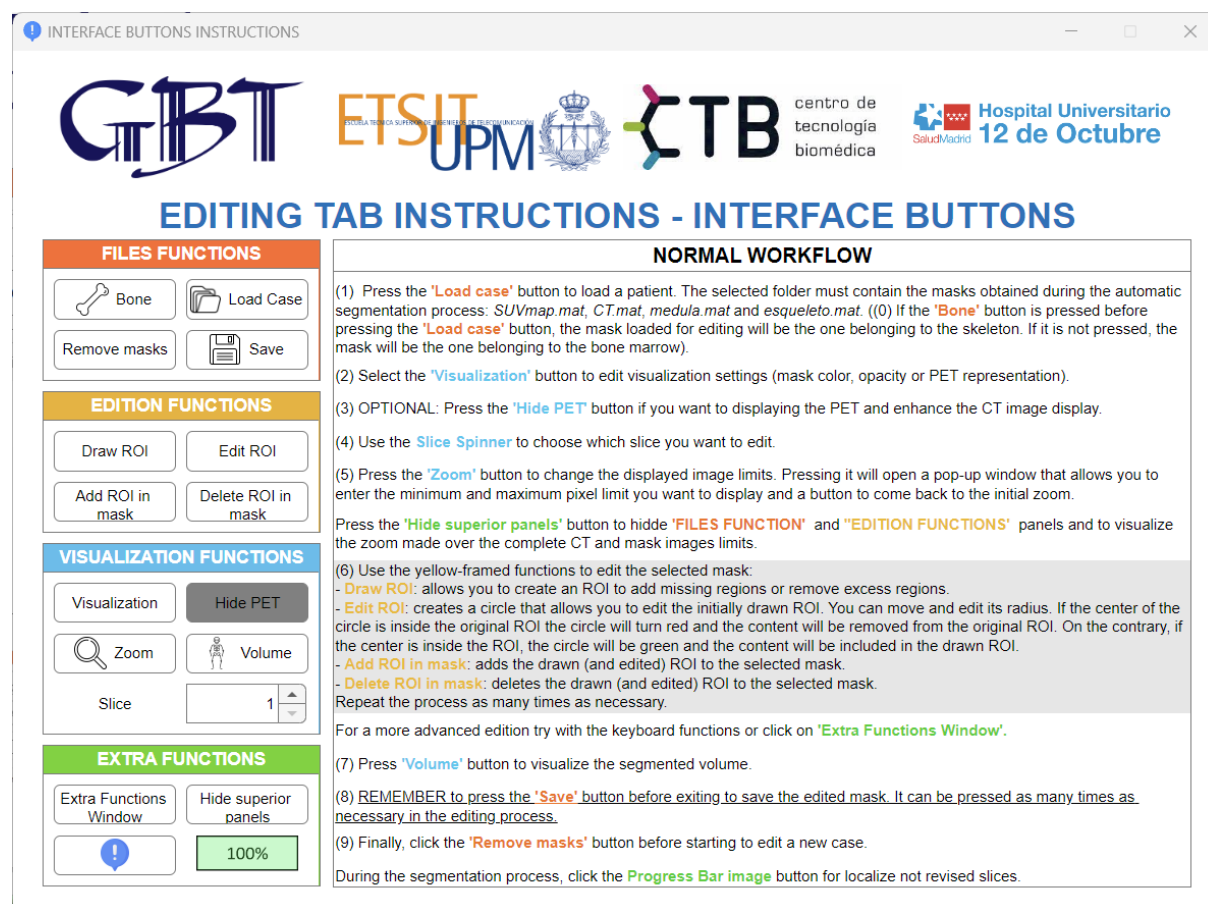


Figure 112. ‘Editing’ tab of the semi-automatic segmentation platform: main window instructions.

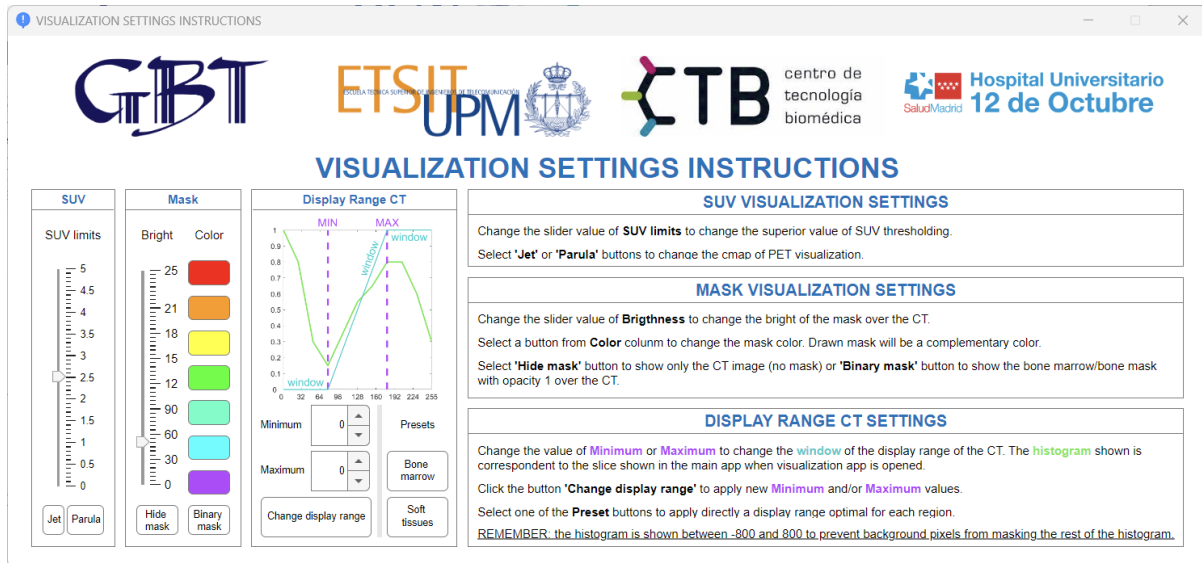


Figure 113. 'Editing' tab of the semi-automatic segmentation platform: 'Visualization settings instructions' window.

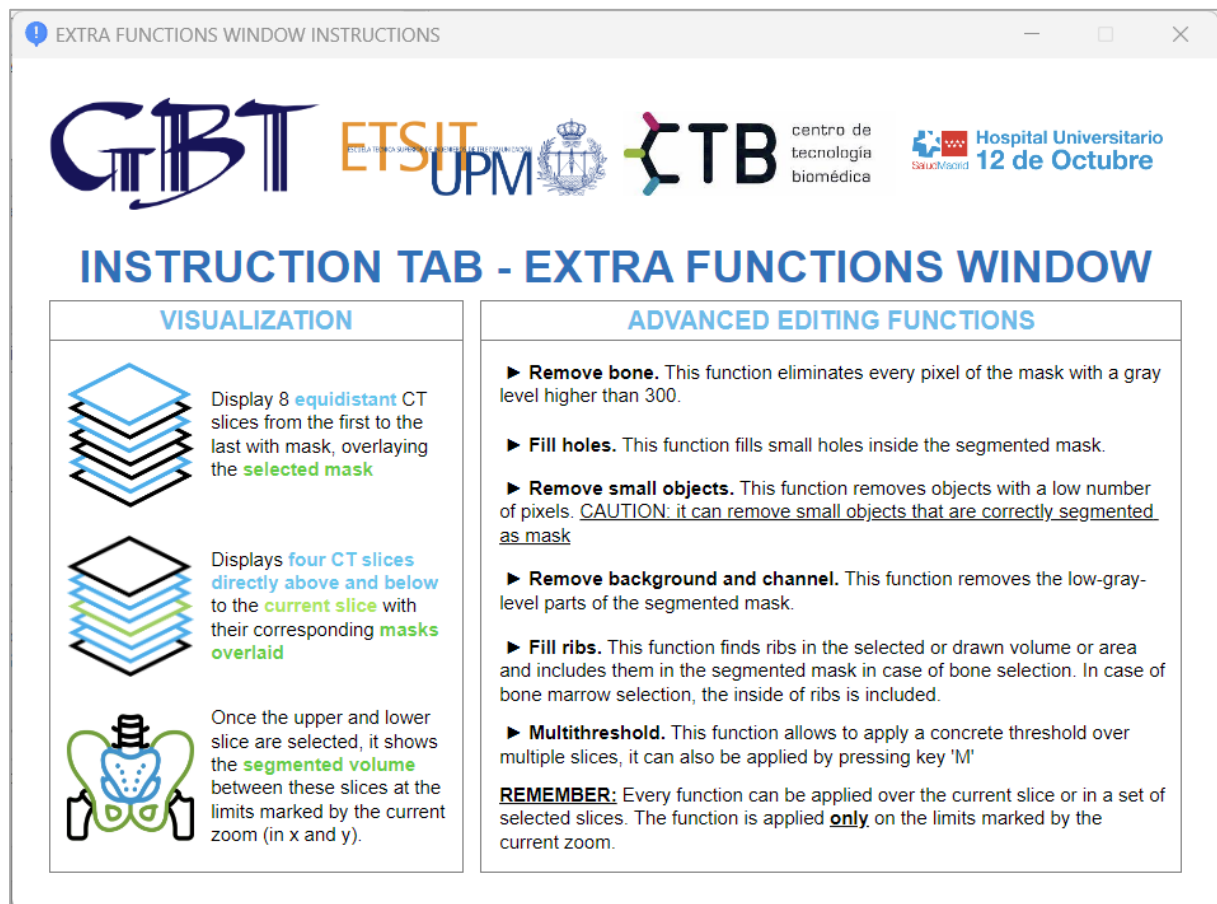


Figure 114. 'Editing' tab of the semi-automatic segmentation platform: 'Extra Functions Window' instructions.

Annex E. Unprocessed bone segmentation

As shown in Figure 31 and Figure 32, the U-Net model trained with an unprocessed CT scans provide unacceptable results. In order to quantify this segmentation, Figure 115 shows the mean Dice results for the different combination of the parameters F, E and order of training. As it could be seen, the best performance obtained is a Dice index of 0.564.

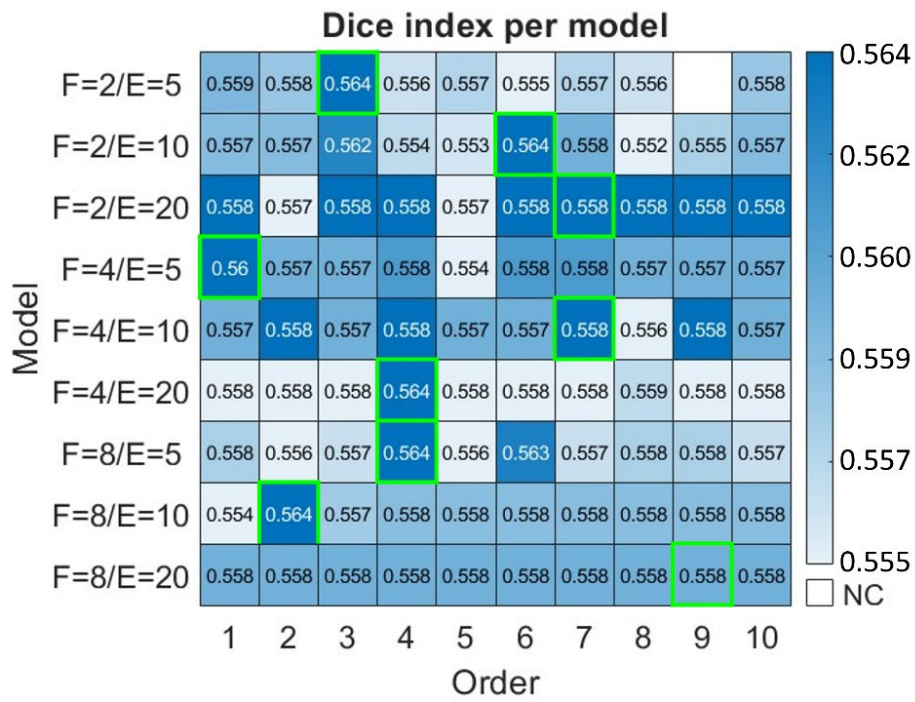


Figure 115. Comparison of training orders for each model, i.e., different pairs of F and E in unprocessed database. NC: Non-Convergent. In lighting green are shown the best result for each pair of feature models: the best Dice Index; a same value of Dice Index, the best Dice (mask) Index; a same value of Dice (mask) Index, the model with higher value of TP.

Annex F. Stretcher removal model results

Figure 116 shows the results of the U-Net models for the different combinations of F and number of epochs for the stretcher segmentation.

F	Epochs	IoU			Dice		
		Mask	Bg	Mean	Mask	Bg	Mean
2	5	0,950	0,999	0,975	0,974	0,999	0,987
	10	0,973	0,999	0,986	0,986	1,000	0,993
	20	0,963	0,999	0,981	0,981	1,000	0,991
4	5	0,965	0,999	0,982	0,982	1,000	0,991
	10	0,971	0,999	0,985	0,985	1,000	0,993
	20	0,971	0,999	0,985	0,985	1,000	0,993
8	5	0,977	1,000	0,989	0,989	1,000	0,995
	10	0,976	1,000	0,988	0,988	1,000	0,994
	20	0,973	0,999	0,986	0,986	1,000	0,993

Figure 116. Results of U-Net architecture trained for the stretcher location, varying the number of channels or filters in the first feature map (F) and the number of epochs. In green is marked the model with the best performance.

Annex G. Performance of ML models MRD detection

Following tables show F1_macro and F1_weighted of the 44 trained ML models for the three types of labelling: PET (Table 33), MFC (Table 34) and pGT (Table 35).

		PET							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
ORIG	DT	0.625	0.675	0.767	0.786	0.686	0.703	0.646	0.660
	RBF	0.767	0.786	0.713	0.739	0.860	0.872	0.395	0.515
	LIN	0.617	0.682	0.233	0.162	0.300	0.245	0.378	0.494
	SIG	0.582	0.648	0.686	0.703	0.776	0.788	0.395	0.515
	RF	0.713	0.739	0.808	0.826	0.767	0.786	0.617	0.652
	GB	0.726	0.745	0.654	0.690	0.732	0.767	0.795	0.819
	LR	0.564	0.570	0.686	0.703	0.767	0.786	0.606	0.616
	NB	0.852	0.867	0.852	0.867	0.852	0.867	0.795	0.819
	k-NN	0.732	0.767	0.692	0.729	0.692	0.729	0.462	0.548
	MLP1	0.852	0.867	0.692	0.729	0.753	0.779	0.686	0.703
	MLP2	0.713	0.739	0.686	0.703	0.673	0.700	0.686	0.703
RED	DT	0.477	0.484	0.726	0.745	0.726	0.745	0.686	0.703
	RBF	0.606	0.616	0.673	0.700	0.726	0.745	0.395	0.515
	LIN	0.361	0.471	0.378	0.494	0.258	0.179	0.753	0.779
	SIG	0.564	0.570	0.713	0.739	0.625	0.675	0.395	0.515
	RF	0.686	0.703	0.726	0.745	0.808	0.826	0.726	0.745
	GB	0.617	0.652	0.726	0.745	0.767	0.786	0.839	0.860
	LR	0.564	0.570	0.686	0.703	0.726	0.745	0.646	0.660
	NB	0.726	0.745	0.726	0.745	0.713	0.739	0.767	0.786
	k-NN	0.625	0.675	0.582	0.648	0.701	0.748	0.701	0.748
	MLP1	0.726	0.745	0.808	0.826	0.686	0.703	0.617	0.652
	MLP2	0.726	0.745	0.808	0.826	0.654	0.690	0.767	0.786
OS	DT	0.606	0.616	0.767	0.786	0.808	0.826	0.686	0.703
	RBF	0.776	0.788	0.713	0.739	0.808	0.826	0.395	0.515
	LIN	0.300	0.245	0.692	0.729	0.300	0.245	0.489	0.577
	SIG	0.233	0.305	0.549	0.615	0.776	0.788	0.395	0.515
	RF	0.606	0.616	0.808	0.826	0.606	0.616	0.686	0.703
	GB	0.732	0.767	0.852	0.867	0.808	0.826	0.808	0.826
	LR	0.646	0.660	0.726	0.745	0.713	0.739	0.596	0.618

OS_RED	NB	0.852	0.867	0.897	0.910	0.767	0.786	0.795	0.819
	k-NN	0.654	0.690	0.713	0.739	0.753	0.779	0.673	0.700
	MLP1	0.686	0.703	0.635	0.659	0.558	0.575	0.635	0.659
	MLP2	0.673	0.700	0.635	0.659	0.635	0.659	0.686	0.703
	DT	0.753	0.779	0.673	0.700	0.808	0.826	0.673	0.700
	RBF	0.635	0.659	0.635	0.659	0.753	0.779	0.395	0.515
	LIN	0.378	0.494	0.378	0.494	0.549	0.615	0.258	0.179
	SIG	0.606	0.616	0.673	0.700	0.776	0.788	0.395	0.515
	RF	0.767	0.786	0.726	0.745	0.767	0.786	0.686	0.703
	GB	0.625	0.675	0.726	0.745	0.767	0.786	0.852	0.867
	LR	0.726	0.745	0.726	0.745	0.726	0.745	0.646	0.660
	NB	0.767	0.786	0.686	0.703	0.726	0.745	0.767	0.786
	k-NN	0.713	0.739	0.544	0.574	0.808	0.826	0.795	0.819
	MLP1	0.726	0.745	0.580	0.614	0.808	0.826	0.654	0.690
	MLP2	0.507	0.533	0.487	0.528	0.662	0.711	0.686	0.703

Table 33. Performance metrics for the ML model for the PET labelling in MRD detection. Outstanding values, i.e. higher than 0.85, are marked in green. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		MFC							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
ORIG	DT	0.377	0.416	0.343	0.400	0.545	0.585	0.364	0.364
	RBF	0.267	0.339	0.313	0.398	0.353	0.449	0.389	0.495
	LIN	0.389	0.495	0.371	0.473	0.267	0.194	0.343	0.285
	SIG	0.545	0.585	0.760	0.775	0.633	0.643	0.389	0.495
	RF	0.313	0.398	0.509	0.546	0.398	0.420	0.267	0.285
	GB	0.313	0.398	0.400	0.466	0.389	0.495	0.371	0.473
	LR	0.664	0.685	0.358	0.374	0.358	0.374	0.569	0.596
	NB	0.509	0.546	0.530	0.553	0.545	0.585	0.569	0.596
	k-NN	0.576	0.634	0.476	0.528	0.455	0.529	0.509	0.564
	MLP1	0.411	0.454	0.476	0.528	0.476	0.528	0.444	0.492
	MLP2	0.530	0.553	0.509	0.546	0.545	0.585	0.581	0.622
OS	DT	0.305	0.331	0.342	0.375	0.436	0.463	0.581	0.622
	RBF	0.317	0.309	0.290	0.370	0.400	0.466	0.389	0.495
	LIN	0.455	0.529	0.248	0.211	0.389	0.495	0.267	0.194
	SIG	0.473	0.506	0.213	0.242	0.583	0.599	0.389	0.495

RF	0.476	0.528	0.313	0.398	0.313	0.398	0.400	0.466
GB	0.444	0.492	0.427	0.499	0.267	0.339	0.343	0.400
LR	0.491	0.509	0.530	0.553	0.583	0.599	0.718	0.732
NB	0.530	0.553	0.473	0.506	0.569	0.596	0.530	0.553
k-NN	0.581	0.622	0.618	0.661	0.542	0.598	0.444	0.492
MLP1	0.371	0.434	0.411	0.454	0.377	0.416	0.377	0.416
MLP2	0.725	0.732	0.444	0.492	0.530	0.553	0.509	0.546

Table 34. Performance metrics for the ML model for the MFC labelling in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		pGT							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
ORIG	DT	0.391	0.559	0.391	0.559	0.391	0.559	0.300	0.429
	RBF	0.417	0.595	0.591	0.653	0.417	0.595	0.626	0.660
	LIN	0.417	0.595	0.417	0.595	0.417	0.595	0.417	0.595
	SIG	0.563	0.589	0.476	0.524	0.429	0.429	0.708	0.726
	RF	0.475	0.571	0.427	0.515	0.427	0.515	0.524	0.626
	GB	0.417	0.595	0.417	0.595	0.635	0.736	0.417	0.595
	LR	0.417	0.595	0.391	0.559	0.364	0.519	0.475	0.571
	NB	0.378	0.454	0.475	0.571	0.427	0.515	0.427	0.515
	k-NN	0.417	0.595	0.417	0.595	0.417	0.595	0.417	0.595
	MLP1	0.417	0.595	0.635	0.736	0.417	0.595	0.364	0.519
	MLP2	0.417	0.595	0.417	0.595	0.576	0.679	0.635	0.736
OS	DT	0.417	0.595	0.391	0.559	0.391	0.559	0.524	0.626
	RBF	0.563	0.589	0.689	0.727	0.650	0.714	0.626	0.660
	LIN	0.326	0.265	0.476	0.524	0.417	0.595	0.417	0.595
	SIG	0.626	0.660	0.626	0.660	0.429	0.429	0.563	0.589
	RF	0.576	0.679	0.524	0.626	0.417	0.595	0.524	0.626
	GB	0.635	0.736	0.576	0.679	0.576	0.679	0.576	0.679
	LR	0.391	0.559	0.391	0.559	0.364	0.519	0.427	0.515
	NB	0.263	0.376	0.300	0.429	0.300	0.429	0.417	0.452
	k-NN	0.391	0.559	0.714	0.775	0.417	0.595	0.364	0.519
	MLP1	0.635	0.736	0.391	0.559	0.417	0.595	0.635	0.736
	MLP2	0.576	0.679	0.364	0.519	0.417	0.595	0.635	0.736

Table 35. Performance metrics for the ML model for the pGT labelling in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

Annex H. Performance of CNNs in MRD detection

Below are the resulting tables for all tested CNN architectures separated by labelling pair and input type. For PET labelling, Table 36 (3D input), Table 37 (pRGB input) and Table 38 (MIP input) are included. For MFC labelling, Table 39 (3D input), Table 40 (pRGB input) and Table 41 (MIP input) are included. For pGT labelling, Table 42 (3D input), Table 43 (pRGB input) and Table 44 (MIP input) are included.

3D - PET													
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.437	0.519	0.518	0.581	0.590	0.638	0.701	0.748	0.487	0.548	0.478	0.484
	2	0.456	0.514	0.544	0.574	0.635	0.659	0.654	0.690	0.606	0.595	0.727	0.745
	4	0.727	0.745	0.767	0.786	0.258	0.179	0.818	0.830	0.487	0.528	0.329	0.267
	8	0.597	0.618	0.625	0.675	0.564	0.570	0.463	0.548	0.581	0.614	0.555	0.602
VGG16	1	0.713	0.739	0.258	0.179	0.258	0.179	0.753	0.779	0.258	0.179	0.258	0.179
	2	0.394	0.442	0.395	0.515	0.597	0.618	0.395	0.515	0.258	0.179	0.558	0.575
	4	0.395	0.515	0.646	0.660	0.654	0.690	0.582	0.648	0.258	0.179	0.258	0.179
	8	0.000	0.000	0.646	0.660	0.258	0.179	0.582	0.648	0.258	0.179	0.258	0.179
ResNet	1	0.395	0.515	0.582	0.648	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
	2	0.395	0.515	0.258	0.179	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
	4	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.258	0.179	0.395	0.515
	8	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
DenseNet	1	0.395	0.515	0.395	0.515	0.258	0.179	0.732	0.767	0.258	0.179	0.692	0.729
	2	0.395	0.515	0.395	0.515	0.258	0.179	0.395	0.515	0.395	0.515	0.258	0.179
	4	0.395	0.515	0.258	0.179	0.662	0.711	0.395	0.515	0.258	0.179	0.517	0.606
	8	0.582	0.648	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
RPET_NET	1	0.425	0.478	0.555	0.602	0.590	0.638	0.411	0.489	0.617	0.652	0.478	0.471
	2	0.357	0.425	0.635	0.659	0.692	0.729	0.597	0.618	0.487	0.548	0.582	0.648
	4	0.487	0.528	0.544	0.574	0.463	0.548	0.558	0.575	0.674	0.700	0.518	0.531
	8	0.597	0.618	0.581	0.614	0.617	0.652	0.652	0.656	0.431	0.415	0.521	0.565

Table 36. Performance metrics for CNNs architectures for PET labelling and 3D input in MRD detection. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.

pRGB - PET													
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.518	0.581	0.517	0.606	0.555	0.602	0.646	0.660	0.558	0.575	0.753	0.779
	2	0.654	0.690	0.437	0.519	0.654	0.690	0.646	0.660	0.521	0.565	0.582	0.648
	4	0.617	0.652	0.767	0.786	0.662	0.711	0.478	0.484	0.361	0.319	0.674	0.700
	8	0.732	0.767	0.549	0.615	0.555	0.602	0.549	0.615	0.418	0.387	0.469	0.448
VGG16	1	0.487	0.548	0.692	0.729	0.581	0.614	0.395	0.515	0.258	0.179	0.258	0.179
	2	0.521	0.565	0.258	0.179	0.549	0.615	0.767	0.786	0.258	0.179	0.652	0.656
	4	0.694	0.702	0.590	0.638	0.654	0.690	0.564	0.570	0.582	0.648	0.687	0.703
	8	0.555	0.602	0.558	0.575	0.590	0.638	0.625	0.675	0.727	0.745	0.522	0.522
ResNet	1	0.395	0.515	0.395	0.515	0.258	0.179	0.395	0.515	0.258	0.179	0.395	0.515
	2	0.395	0.515	0.489	0.577	0.395	0.515	0.395	0.515	0.395	0.515	0.258	0.179
	4	0.463	0.548	0.582	0.648	0.395	0.515	0.395	0.515	0.395	0.515	0.582	0.648
	8	0.489	0.577	0.489	0.577	0.517	0.606	0.395	0.515	0.395	0.515	0.395	0.515
DenseNet	1	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
	2	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.258	0.179
	4	0.395	0.515	0.258	0.179	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
	8	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
RPET_NET	1	0.411	0.489	0.487	0.548	0.384	0.457	0.469	0.490	0.635	0.659	0.558	0.575
	2	0.456	0.514	0.469	0.490	0.521	0.565	0.654	0.690	0.735	0.745	0.635	0.659
	4	0.418	0.448	0.555	0.602	0.625	0.675	0.590	0.638	0.597	0.618	0.544	0.574
	8	0.437	0.519	0.521	0.565	0.646	0.660	0.654	0.690	0.564	0.570	0.564	0.570

Table 37. Performance metrics for CNNs architectures for PET labelling and pRGB input in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

MIP - PET													
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.581	0.614	0.654	0.690	0.652	0.656	0.395	0.515	0.258	0.179	0.258	0.179
	2	0.625	0.675	0.507	0.533	0.258	0.179	0.395	0.515	0.418	0.387	0.258	0.179
	4	0.391	0.399	0.456	0.514	0.617	0.652	0.662	0.711	0.258	0.179	0.564	0.558
	8	0.617	0.652	0.654	0.690	0.487	0.528	0.674	0.700	0.564	0.570	0.258	0.179
VGG16	1	0.581	0.614	0.654	0.690	0.652	0.656	0.395	0.515	0.258	0.179	0.258	0.179
	2	0.625	0.675	0.507	0.533	0.258	0.179	0.395	0.515	0.418	0.387	0.258	0.179
	4	0.391	0.399	0.456	0.514	0.617	0.652	0.662	0.711	0.258	0.179	0.564	0.558
	8	0.617	0.652	0.654	0.690	0.487	0.528	0.674	0.700	0.564	0.570	0.258	0.179
ResNet	1	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
	2	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515

	4	0.617	0.652	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
	8	0.395	0.515	0.625	0.675	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
DenseNet	1	0.489	0.577	0.258	0.179	0.258	0.179	0.395	0.515	0.258	0.179	0.395	0.515
	2	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
	4	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
	8	0.395	0.515	0.258	0.179	0.395	0.515	0.395	0.515	0.395	0.515	0.395	0.515
RPET_NET	1	0.489	0.577	0.555	0.602	0.555	0.602	0.617	0.682	0.654	0.690	0.521	0.565
	2	0.549	0.615	0.487	0.548	0.625	0.675	0.662	0.711	0.452	0.488	0.487	0.548
	4	0.692	0.729	0.463	0.548	0.654	0.690	0.692	0.729	0.692	0.729	0.692	0.729
	8	0.654	0.690	0.625	0.675	0.590	0.638	0.463	0.548	0.654	0.690	0.662	0.711

Table 38. Performance metrics for CNNs architectures for PET labelling and MIP input in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		3D - MFC											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.697	0.738	0.313	0.398	0.618	0.661	0.372	0.473	0.377	0.416	0.545	0.585
	2	0.542	0.598	0.545	0.585	0.509	0.564	0.389	0.495	0.509	0.546	0.267	0.194
	4	0.618	0.661	0.400	0.466	0.618	0.661	0.353	0.449	0.267	0.194	0.267	0.194
	8	0.530	0.553	0.530	0.506	0.267	0.194	0.389	0.495	0.267	0.194	0.267	0.194
VGG16	1	0.389	0.495	0.727	0.758	0.267	0.194	0.389	0.495	0.267	0.194	0.267	0.194
	2	0.389	0.495	0.267	0.194	0.411	0.454	0.389	0.495	0.267	0.194	0.267	0.194
	4	0.389	0.495	0.267	0.194	0.267	0.194	0.389	0.495	0.267	0.194	0.267	0.194
	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ResNet	1	0.389	0.495	0.353	0.449	0.389	0.495	0.389	0.495	0.389	0.495	0.267	0.194
	2	0.389	0.495	0.473	0.506	0.389	0.495	0.267	0.194	0.267	0.194	0.398	0.376
	4	0.389	0.495	0.389	0.495	0.389	0.495	0.372	0.473	0.389	0.495	0.389	0.495
	8	0.389	0.495	0.389	0.495	0.576	0.634	0.389	0.495	0.389	0.495	0.389	0.495
DenseNet	1	0.389	0.495	0.267	0.194	0.267	0.194	0.389	0.495	0.267	0.194	0.267	0.194
	2	0.389	0.495	0.267	0.194	0.267	0.194	0.389	0.495	0.389	0.495	0.389	0.495
	4	0.389	0.495	0.267	0.194	0.267	0.194	0.389	0.495	0.267	0.194	0.267	0.194
	8	0.389	0.495	0.267	0.194	0.267	0.194	0.389	0.495	0.267	0.194	0.389	0.495
RPET_NET	1	0.542	0.598	0.477	0.528	0.334	0.424	0.477	0.528	0.428	0.499	0.607	0.636
	2	0.428	0.499	0.377	0.416	0.576	0.634	0.509	0.564	0.455	0.529	0.444	0.492
	4	0.576	0.634	0.477	0.528	0.509	0.564	0.444	0.492	0.400	0.466	0.530	0.553
	8	0.291	0.370	0.428	0.499	0.353	0.449	0.511	0.590	0.545	0.585	0.343	0.400

Table 39. Performance metrics for CNNs architectures for MFC labelling and 3D input in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

Table 40. Performance metrics for CNNs architectures for MFC labelling and pRGB input in MRD detection. Outstanding values, i.e. above 0.85, are marked in green. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

254

	4	0.389	0.495	0.389	0.495	0.389	0.495	0.389	0.495	0.267	0.194	0.389	0.495
	8	0.267	0.194	0.612	0.670	0.389	0.495	0.389	0.495	0.267	0.194	0.389	0.495
DenseNet	1	0.389	0.495	0.267	0.194	0.389	0.495	0.389	0.495	0.267	0.194	0.267	0.194
	2	0.389	0.495	0.267	0.194	0.267	0.194	0.389	0.495	0.389	0.495	0.389	0.495
	4	0.389	0.495	0.389	0.495	0.267	0.194	0.389	0.495	0.267	0.194	0.267	0.194
	8	0.542	0.598	0.389	0.495	0.267	0.194	0.389	0.495	0.267	0.194	0.389	0.495
RPET_NET	1	0.545	0.585	0.509	0.564	0.509	0.564	0.398	0.420	0.444	0.492	0.581	0.622
	2	0.400	0.466	0.313	0.398	0.334	0.424	0.353	0.449	0.408	0.415	0.444	0.492
	4	0.353	0.449	0.334	0.424	0.313	0.398	0.400	0.466	0.612	0.670	0.428	0.499
	8	0.400	0.466	0.542	0.598	0.455	0.529	0.313	0.364	0.542	0.598	0.511	0.590

Table 41. Performance metrics for CNNs architectures for MFC labelling and MIP input in MRD detection. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		3D - pGT											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.640	0.752	0.640	0.752	0.583	0.700	0.423	0.620	0.423	0.620	0.423	0.620
	2	0.354	0.434	0.423	0.620	0.423	0.620	0.423	0.620	0.211	0.112	0.267	0.267
	4	0.318	0.466	0.400	0.587	0.400	0.587	0.423	0.620	0.423	0.620	0.423	0.620
	8	0.286	0.419	0.423	0.620	0.423	0.620	0.423	0.620	0.400	0.587	0.423	0.620
VGG16	1	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	2	0.348	0.510	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	4	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ResNet	1	0.598	0.611	0.762	0.806	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	2	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	4	0.400	0.587	0.400	0.587	0.400	0.587	0.423	0.620	0.423	0.620	0.423	0.620
	8	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
DenseNet	1	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.211	0.112	0.211	0.112
	2	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	4	0.423	0.620	0.423	0.620	0.211	0.112	0.211	0.112	0.423	0.620	0.423	0.620
	8	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.211	0.112
RPET_NET	1	0.583	0.700	0.318	0.466	0.535	0.650	0.400	0.587	0.375	0.550	0.318	0.466
	2	0.535	0.650	0.348	0.510	0.375	0.550	0.603	0.677	0.498	0.561	0.423	0.620
	4	0.489	0.600	0.348	0.510	0.535	0.650	0.211	0.309	0.445	0.497	0.583	0.700
	8	0.375	0.550	0.400	0.587	0.375	0.550	0.400	0.493	0.400	0.493	0.498	0.561

Table 42. Performance metrics for CNNs architectures for pGT labelling and 3D input in MRD detection. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		pRGB - pGT											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.354	0.434	0.445	0.548	0.525	0.554	0.423	0.620	0.318	0.466	0.348	0.510
	2	0.659	0.733	0.489	0.600	0.306	0.370	0.445	0.497	0.465	0.481	0.583	0.700
	4	0.603	0.677	0.603	0.677	0.584	0.622	0.498	0.561	0.550	0.620	0.331	0.351
	8	0.700	0.747	0.535	0.650	0.445	0.497	0.423	0.620	0.423	0.620	0.445	0.548
VGG16	1	0.445	0.548	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.550	0.620
	2	0.603	0.677	0.375	0.550	0.354	0.434	0.306	0.370	0.423	0.620	0.423	0.620
	4	0.390	0.428	0.723	0.748	0.489	0.600	0.721	0.790	0.423	0.620	0.423	0.620
	8	0.423	0.620	0.423	0.620	0.489	0.600	0.423	0.620	0.423	0.620	0.423	0.620
ResNet	1	0.423	0.620	0.721	0.790	0.423	0.620	0.400	0.587	0.423	0.620	0.423	0.620
	2	0.423	0.620	0.640	0.752	0.423	0.620	0.423	0.620	0.423	0.620	0.211	0.112
	4	0.423	0.620	0.423	0.620	0.318	0.466	0.423	0.620	0.423	0.620	0.535	0.650
	8	0.583	0.700	0.423	0.620	0.535	0.650	0.423	0.620	0.423	0.620	0.348	0.510
DenseNet	1	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	2	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	4	0.423	0.620	0.423	0.620	0.306	0.241	0.423	0.620	0.211	0.112	0.423	0.620
	8	0.211	0.112	0.211	0.112	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
RPET_NET	1	0.348	0.510	0.535	0.650	0.286	0.419	0.445	0.497	0.445	0.497	0.375	0.550
	2	0.792	0.850	0.550	0.620	0.375	0.550	0.584	0.622	0.489	0.600	0.400	0.493
	4	0.423	0.620	0.721	0.790	0.318	0.466	0.603	0.677	0.498	0.561	0.267	0.267
	8	0.423	0.620	0.445	0.548	0.318	0.466	0.498	0.561	0.400	0.587	0.535	0.650

Table 43. Performance metrics for CNNs architectures for pGT labelling and pRGB input in MRD detection. Outstanding values, i.e. above 0.85, are marked in green. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		MIP - pGT											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.348	0.510	0.423	0.620
	2	0.423	0.620	0.535	0.650	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	4	0.535	0.650	0.535	0.650	0.535	0.650	0.423	0.620	0.423	0.620	0.423	0.620
	8	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
VGG16	1	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.348	0.510	0.423	0.620
	2	0.423	0.620	0.535	0.650	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	4	0.535	0.650	0.535	0.650	0.535	0.650	0.423	0.620	0.423	0.620	0.423	0.620
	8	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
Re	1	0.400	0.587	0.423	0.620	0.211	0.112	0.423	0.620	0.423	0.620	0.423	0.620

DenseNet	2	0.489	0.600	0.423	0.620	0.254	0.300	0.423	0.620	0.423	0.620	0.423	0.620
	4	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	8	0.331	0.310	0.423	0.620	0.400	0.587	0.423	0.620	0.423	0.620	0.423	0.620
	1	0.423	0.620	0.423	0.620	0.211	0.112	0.423	0.620	0.423	0.620	0.423	0.620
	2	0.400	0.587	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	4	0.354	0.434	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	8	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620	0.423	0.620
	1	0.423	0.620	0.400	0.587	0.348	0.510	0.659	0.733	0.423	0.620	0.535	0.650
	2	0.318	0.466	0.318	0.466	0.535	0.650	0.400	0.587	0.375	0.550	0.640	0.752
	4	0.400	0.587	0.375	0.550	0.535	0.650	0.423	0.620	0.375	0.550	0.375	0.550
	8	0.583	0.700	0.348	0.510	0.535	0.650	0.375	0.550	0.400	0.587	0.318	0.466
RPET_NET	1	0.423	0.620	0.400	0.587	0.348	0.510	0.659	0.733	0.423	0.620	0.535	0.650
	2	0.318	0.466	0.318	0.466	0.535	0.650	0.400	0.587	0.375	0.550	0.640	0.752
	4	0.400	0.587	0.375	0.550	0.535	0.650	0.423	0.620	0.375	0.550	0.375	0.550
	8	0.583	0.700	0.348	0.510	0.535	0.650	0.375	0.550	0.400	0.587	0.318	0.466

Table 44. Performance metrics for CNNs architectures for pGT labelling and MIP input in MRD detection. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

Annex I. Performance of ML models in MRD prediction

Following tables show F1_macro and F1_weighted of the 44 trained ML models for MRD prediction with the two types of labelling: PET (Table 45) and MFC (Table 46).

		PET							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
ORIG	DT	0,333	0,467	0,697	0,709	0,495	0,515	0,495	0,474
	RBF	0,412	0,577	0,412	0,577	0,333	0,467	0,412	0,577
	LIN	0,412	0,577	0,167	0,233	0,167	0,233	0,375	0,525
	SIG	0,412	0,577	0,762	0,800	0,670	0,710	0,583	0,617
	RF	0,333	0,467	0,600	0,680	0,375	0,525	0,412	0,577
	GB	0,375	0,525	0,375	0,525	0,375	0,525	0,375	0,525
	LR	0,375	0,525	0,333	0,467	0,231	0,323	0,231	0,323
	NB	0,375	0,425	0,670	0,710	0,583	0,617	0,762	0,800
	k-NN	0,600	0,680	0,375	0,525	0,375	0,525	0,375	0,525
	MLP1	0,333	0,467	0,333	0,467	0,583	0,617	0,451	0,516
	MLP2	0,375	0,525	0,375	0,525	0,286	0,400	0,286	0,400
	DT	0,333	0,467	0,333	0,467	0,762	0,800	0,286	0,400
OS	RBF	0,231	0,139	0,670	0,710	0,583	0,617	0,231	0,139
	LIN	0,400	0,400	0,375	0,525	0,400	0,400	0,375	0,525
	SIG	0,231	0,323	0,293	0,321	0,286	0,400	0,583	0,617
	RF	0,333	0,467	0,333	0,467	0,600	0,680	0,375	0,525
	GB	0,375	0,525	0,375	0,525	0,600	0,680	0,375	0,525
	LR	0,524	0,600	0,524	0,600	0,286	0,400	0,333	0,467
	NB	0,524	0,600	0,600	0,680	0,524	0,600	0,762	0,800
	k-NN	0,451	0,516	0,333	0,467	0,762	0,800	0,375	0,525
	MLP1	0,375	0,525	0,600	0,680	0,333	0,467	0,286	0,400
	MLP2	0,600	0,680	0,375	0,525	0,375	0,525	0,375	0,525

Table 45. Performance metrics for the ML model for the PET labelling in MRD prediction. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		MFC							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
ORIG	DT	0,451	0,516	0,600	0,600	0,495	0,515	0,600	0,680
	RBF	0,333	0,467	0,231	0,139	0,231	0,139	0,231	0,139
	LIN	0,412	0,577	0,412	0,577	0,375	0,525	0,600	0,680
	SIG	0,524	0,600	0,524	0,600	0,495	0,515	0,375	0,425
	RF	0,600	0,680	0,600	0,680	0,600	0,680	0,524	0,600
	GB	0,688	0,763	0,375	0,525	0,412	0,577	0,412	0,577
	LR	0,293	0,321	0,293	0,321	0,412	0,577	0,412	0,577
	NB	0,524	0,600	0,670	0,710	0,583	0,617	0,670	0,710
	k-NN	0,412	0,577	0,412	0,577	0,688	0,763	0,688	0,763
	MLP1	0,524	0,600	0,375	0,525	0,375	0,525	0,600	0,680
	MLP2	0,583	0,617	0,670	0,710	0,583	0,617	0,583	0,617
OS	DT	0,670	0,710	0,583	0,617	0,400	0,400	0,524	0,600
	RBF	0,286	0,400	0,375	0,425	0,451	0,516	0,412	0,577
	LIN	0,375	0,525	0,524	0,600	0,451	0,516	0,495	0,474
	SIG	0,293	0,321	0,524	0,600	0,400	0,400	0,375	0,425
	RF	0,600	0,680	0,583	0,617	0,600	0,680	0,600	0,680
	GB	0,375	0,525	0,688	0,763	0,688	0,763	0,688	0,763
	LR	0,293	0,321	0,375	0,425	0,293	0,321	0,293	0,321
	NB	0,762	0,800	0,670	0,710	0,600	0,680	0,670	0,710
	k-NN	0,524	0,600	0,451	0,516	0,688	0,763	0,524	0,600
	MLP1	0,375	0,425	0,670	0,710	0,451	0,516	0,583	0,617
	MLP2	0,375	0,425	0,583	0,617	0,451	0,516	0,495	0,515

Table 46. Performance metrics for the ML model for the PET labelling in MRD prediction. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

Annex J. Univariate analysis results for risk assessment in SMM

In this Annex, the results for SMM quantification for risk of progression to MM prediction are included in Table 47 for IMWG labelling and Table 48 for PETHEMA labelling.

Gr.	Variable	Bin Width			
		0.5	0.25	0.125	0.0625
Image and ROI descriptors	Image-original_Mean	0.182	0.182	0.182	0.182
	Image-original_Minimum	0.892	0.892	0.892	0.892
	Image-original_Maximum	0.236	0.236	0.236	0.236
	Mask-original_VoxelNum	0.120	0.120	0.120	0.120
	Mask-original_VolumeNum	0.342	0.342	0.342	0.342
	Image-interpolated_Mean	0.645	0.645	0.645	0.645
	Image-interpolated_Minimum	0.892	0.892	0.892	0.892
	Image-interpolated_Maximum	0.236	0.236	0.236	0.236
	Mask-interpolated_VoxelNum	0.120	0.120	0.120	0.120
	Mask-interpolated_VolumeNum	0.342	0.342	0.342	0.342
	Mask-interpolated_Mean	0.257	0.257	0.257	0.257
	Mask-interpolated_Minimum	0.085	0.085	0.085	0.085
	Mask-interpolated_Maximum	0.805	0.805	0.805	0.805
	Mask-resegmented_VoxelNum	0.120	0.120	0.120	0.120
	Mask-resegmented_VolumeNum	0.342	0.342	0.342	0.342
	Mask-resegmented_Mean	0.259	0.259	0.259	0.259
	Mask-resegmented_Minimum	0.085	0.085	0.085	0.085
	Mask-resegmented_Maximum	0.737	0.737	0.737	0.737
	Shape_Elongation	0.410	0.410	0.410	0.410
	Shape_Flatness	0.856	0.856	0.856	0.856
	Shape_LeastAxisLength	0.568	0.568	0.568	0.568
	Shape_MajorAxisLength	0.577	0.577	0.577	0.577
	Shape_Maximum2DDiameterColumn	0.216	0.216	0.216	0.216
	Shape_Maximum2DDiameterRow	0.258	0.258	0.258	0.258
	Shape_Maximum2DDiameterSlice	0.158	0.158	0.158	0.158
	Shape_Maximum3DDiameter	0.334	0.334	0.334	0.334
	Shape_MeshVolume	0.123	0.123	0.123	0.123
	Shape_MinorAxisLength	0.139	0.139	0.139	0.139
	Shape_Sphericity	0.149	0.149	0.149	0.149
	Shape_SurfaceArea	0.182	0.182	0.182	0.182
	Shape_SurfaceVolumeRatio	0.056	0.056	0.056	0.056
	Shape_VoxelVolume	0.120	0.120	0.120	0.120

First Order	10Percentile	0.129	0.129	0.129	0.129
	90Percentile	0.491	0.491	0.491	0.491
	Energy	0.701	0.701	0.701	0.701
	Entropy	0.791	0.793	0.816	0.811
	InterquartileRange	0.745	0.745	0.745	0.745
	Kurtosis	0.164	0.164	0.164	0.164
	Maximum	0.737	0.737	0.737	0.737
	MeanAbsoluteDeviation	0.792	0.792	0.792	0.792
	Mean	0.259	0.259	0.259	0.259
	Median	0.190	0.190	0.190	0.190
	Minimum	0.085	0.085	0.085	0.085
	Range	0.795	0.795	0.795	0.795
	RobustMeanAbsoluteDeviation	0.704	0.704	0.704	0.704
	RootMeanSquared	0.362	0.362	0.362	0.362
	Skewness	0.072	0.072	0.072	0.072
	TotalEnergy	0.701	0.701	0.701	0.701
	Uniformity	0.720	0.718	0.732	0.724
	Variance	0.910	0.910	0.910	0.910
GLCM	Autocorrelation	0.401	0.415	0.557	0.581
	ClusterProminence	0.897	0.854	0.875	0.875
	ClusterShade	0.248	0.203	0.260	0.236
	ClusterTendency	0.895	0.890	0.895	0.895
	Contrast	0.745	0.801	0.820	0.819
	Correlation	0.948	0.802	0.879	0.826
	DifferenceAverage	0.617	0.612	0.624	0.618
	DifferenceEntropy	0.639	0.684	0.674	0.633
	DifferenceVariance	0.787	0.815	0.805	0.805
	Id	0.579	0.505	0.450	0.386
	Idm	0.588	0.533	0.470	0.379
	Idmn	0.711	0.498	0.514	0.499
	Idn	0.577	0.601	0.600	0.582
	Imc1	0.374	0.415	0.558	0.499
	Imc2	0.980	0.752	0.679	0.419
	InverseVariance	0.535	0.293	0.117	0.786
	JointAverage	0.283	0.295	0.476	0.587
	JointEnergy	0.641	0.632	0.537	0.509
	JointEntropy	0.723	0.704	0.702	0.672
	MCC	0.759	0.679	0.687	0.184
	MaximumProbability	0.671	0.682	0.895	0.878
	SumAverage	0.283	0.295	0.476	0.587
	SumEntropy	0.740	0.747	0.811	0.823
	SumSquares	0.893	0.888	0.894	0.894

GLDM	DependenceEntropy	0.551	0.695	0.804	0.463
	DependenceNonUniformity	0.058	0.074	0.245	0.358
	DependenceNonUniformityNormalized	0.155	0.536	0.281	0.251
	DependenceVariance	0.835	0.410	0.348	0.363
	GrayLevelNonUniformity	0.132	0.135	0.137	0.142
	GrayLevelVariance	0.906	0.904	0.909	0.909
	HighGrayLevelEmphasis	0.377	0.403	0.554	0.581
	LargeDependenceEmphasis	0.299	0.280	0.239	0.209
	LargeDependenceHighGrayLevelEmphasis	0.429	0.448	0.635	0.627
	LargeDependenceLowGrayLevelEmphasis	0.141	0.207	0.485	0.748
	LowGrayLevelEmphasis	0.122	0.280	0.464	0.604
	SmallDependenceEmphasis	0.296	0.360	0.456	0.439
	SmallDependenceHighGrayLevelEmphasis	0.473	0.612	0.796	0.742
	SmallDependenceLowGrayLevelEmphasis	0.163	0.055	0.389	0.699
GLRLM	GrayLevelNonUniformity	0.218	0.253	0.232	0.191
	GrayLevelNonUniformityNormalized	0.854	0.893	0.872	0.817
	GrayLevelVariance	0.920	0.956	0.987	0.979
	HighGrayLevelRunEmphasis	0.447	0.466	0.676	0.669
	LongRunEmphasis	0.174	0.206	0.259	0.264
	LongRunHighGrayLevelEmphasis	0.943	0.716	0.931	0.800
	LongRunLowGrayLevelEmphasis	0.111	0.159	0.440	0.626
	LowGrayLevelRunEmphasis	0.098	0.117	0.381	0.437
	RunEntropy	0.591	0.545	0.668	0.822
	RunLengthNonUniformity	0.582	0.671	0.434	0.294
	RunLengthNonUniformityNormalized	0.419	0.451	0.337	0.262
	RunPercentage	0.303	0.308	0.273	0.230
	RunVariance	0.137	0.151	0.237	0.328
	ShortRunEmphasis	0.471	0.489	0.371	0.274
	ShortRunHighGrayLevelEmphasis	0.517	0.549	0.705	0.707
	ShortRunLowGrayLevelEmphasis	0.099	0.081	0.245	0.263
GLSZM	GrayLevelNonUniformity	0.946	0.916	0.989	0.977
	GrayLevelNonUniformityNormalized	0.694	0.672	0.612	0.851
	GrayLevelVariance	0.711	0.500	0.399	0.533
	HighGrayLevelZoneEmphasis	0.635	0.771	0.905	0.882
	LargeAreaEmphasis	0.808	0.569	0.590	0.474
	LargeAreaHighGrayLevelEmphasis	0.965	0.807	0.671	0.556
	LargeAreaLowGrayLevelEmphasis	0.506	0.335	0.484	0.565
	LowGrayLevelZoneEmphasis	0.030	0.057	0.225	0.440
	SizeZoneNonUniformity	0.724	0.770	0.792	0.757
	SizeZoneNonUniformityNormalized	0.952	0.754	0.828	0.977
	SmallAreaEmphasis	0.894	0.845	0.792	0.996
	SmallAreaHighGrayLevelEmphasis	0.635	0.718	0.725	0.910

NGTDM	SmallAreaLowGrayLevelEmphasis	0.123	0.104	0.332	0.572
	ZoneEntropy	0.873	0.385	0.190	0.199
	ZonePercentage	0.305	0.377	0.309	0.467
	ZoneVariance	0.823	0.577	0.595	0.474
	Busyness	0.627	0.634	0.715	0.805
	Coarseness	0.350	0.265	0.315	0.523
	Complexity	0.779	0.754	0.827	0.805
	Contrast	0.491	0.417	0.283	0.186
	Strength	0.872	0.824	0.795	0.679

Table 47. P-values for ANOVA or Kruskal-Wallis tests for univariate analysis to assess the risk of progression to MM in patients with SMM in IMGW labelling. Variables with a p-value lower than 0.05, i.e. with significant differences between classes, for any of the bin widths are marked in green and bold type. P-values corrected by Benjamini-Hochberg correction are not shown, as none obtain values which means significant differences between classes. Gr: group of features.

Gr.	Variable	Bin Width			
		0.5	0.25	0.125	0.0625
Image and ROI descriptors	Image-original_Mean	0.882	0.882	0.882	0.882
	Image-original_Minimum	0.820	0.820	0.820	0.820
	Image-original_Maximum	0.622	0.622	0.622	0.622
	Mask-original_VoxelNum	0.891	0.891	0.891	0.891
	Mask-original_VolumeNum	0.379	0.379	0.379	0.379
	Image-interpolated_Mean	0.671	0.671	0.671	0.671
	Image-interpolated_Minimum	0.820	0.820	0.820	0.820
	Image-interpolated_Maximum	0.622	0.622	0.622	0.622
	Mask-interpolated_VoxelNum	0.891	0.891	0.891	0.891
	Mask-interpolated_VolumeNum	0.379	0.379	0.379	0.379
	Mask-interpolated_Mean	0.602	0.602	0.602	0.602
	Mask-interpolated_Minimum	0.045	0.045	0.045	0.045
	Mask-interpolated_Maximum	0.768	0.768	0.768	0.768
	Mask-resegmented_VoxelNum	0.891	0.891	0.891	0.891
	Mask-resegmented_VolumeNum	0.378	0.378	0.378	0.378
	Mask-resegmented_Mean	0.602	0.602	0.602	0.602
	Mask-resegmented_Minimum	0.045	0.045	0.045	0.045
	Mask-resegmented_Maximum	0.698	0.698	0.698	0.698
	Shape_Elongation	0.517	0.517	0.517	0.517
	Shape_Flatness	0.064	0.064	0.064	0.064
	Shape_LeastAxisLength	0.072	0.072	0.072	0.072
	Shape_MajorAxisLength	0.111	0.111	0.111	0.111
	Shape_Maximum2DDiameterColumn	0.704	0.704	0.704	0.704
	Shape_Maximum2DDiameterRow	0.174	0.174	0.174	0.174

First Order	Shape_Maximum2DDiameterSlice	0.875	0.875	0.875	0.875
	Shape_Maximum3DDiameter	0.145	0.145	0.145	0.145
	Shape_MeshVolume	0.888	0.888	0.888	0.888
	Shape_MinorAxisLength	0.171	0.171	0.171	0.171
	Shape_Sphericity	0.207	0.207	0.207	0.207
	Shape_SurfaceArea	0.821	0.821	0.821	0.821
	Shape_SurfaceVolumeRatio	0.082	0.082	0.082	0.082
	Shape_VoxelVolume	0.891	0.891	0.891	0.891
	10Percentile	0.311	0.311	0.311	0.311
	90Percentile	0.780	0.780	0.780	0.780
	Energy	0.572	0.572	0.572	0.572
	Entropy	0.577	0.560	0.563	0.577
	InterquartileRange	0.539	0.539	0.539	0.539
	Kurtosis	0.787	0.787	0.787	0.787
	Maximum	0.698	0.698	0.698	0.698
	MeanAbsoluteDeviation	0.658	0.658	0.658	0.658
	Mean	0.602	0.602	0.602	0.602
	Median	0.855	0.855	0.855	0.855
	Minimum	0.045	0.045	0.045	0.045
	Range	0.783	0.783	0.783	0.783
	RobustMeanAbsoluteDeviation	0.495	0.495	0.495	0.495
	RootMeanSquared	0.851	0.851	0.851	0.851
	Skewness	0.417	0.417	0.417	0.417
	TotalEnergy	0.572	0.572	0.572	0.572
	Uniformity	0.505	0.428	0.489	0.490
	Variance	0.817	0.817	0.817	0.817
GLCM	Autocorrelation	0.640	0.658	0.906	0.969
	ClusterProminence	0.934	0.936	0.934	0.934
	ClusterShade	0.818	0.721	0.772	0.750
	ClusterTendency	0.759	0.738	0.747	0.745
	Contrast	0.869	0.938	0.945	0.946
	Correlation	0.209	0.103	0.063	0.053
	DifferenceAverage	0.797	0.838	0.840	0.839
	DifferenceEntropy	0.762	0.822	0.814	0.796
	DifferenceVariance	0.872	0.965	0.973	0.974
	Id	0.776	0.777	0.735	0.692
	Idm	0.782	0.797	0.748	0.687
	Idmn	0.840	0.764	0.808	0.767
	Idn	0.861	0.805	0.843	0.845
	Imc1	0.125	0.175	0.255	0.262
	Imc2	0.386	0.035	0.021	0.019
	InverseVariance	0.745	0.604	0.431	0.906

GLDM	JointAverage	0.669	0.675	0.861	0.913
	JointEnergy	0.600	0.694	0.799	0.771
	JointEntropy	0.641	0.706	0.737	0.820
	MCC	0.190	0.045	0.057	0.210
	MaximumProbability	0.575	0.693	0.977	0.955
	SumAverage	0.669	0.675	0.861	0.913
	SumEntropy	0.573	0.574	0.488	0.500
	SumSquares	0.770	0.747	0.755	0.753
	DependenceEntropy	0.541	0.534	0.311	0.213
	DependenceNonUniformity	0.508	0.664	0.628	0.572
	DependenceNonUniformityNormalized	0.236	0.044	0.467	0.413
	DependenceVariance	0.482	0.538	0.726	0.521
	GrayLevelNonUniformity	0.417	0.396	0.402	0.418
	GrayLevelVariance	0.832	0.812	0.820	0.818
	HighGrayLevelEmphasis	0.629	0.632	0.868	0.928
	LargeDependenceEmphasis	0.556	0.592	0.554	0.535
	LargeDependenceHighGrayLevelEmphasis	0.951	0.921	0.754	0.457
	LargeDependenceLowGrayLevelEmphasis	0.465	0.513	0.470	0.442
	LowGrayLevelEmphasis	0.464	0.509	0.476	0.391
	SmallDependenceEmphasis	0.522	0.572	0.543	0.678
GLRLM	SmallDependenceHighGrayLevelEmphasis	0.563	0.646	0.787	0.836
	SmallDependenceLowGrayLevelEmphasis	0.176	0.090	0.156	0.551
	GrayLevelNonUniformity	0.612	0.566	0.513	0.459
	GrayLevelNonUniformityNormalized	0.627	0.659	0.623	0.548
	GrayLevelVariance	0.901	0.955	0.979	0.942
	HighGrayLevelRunEmphasis	0.606	0.617	0.881	0.927
	LongRunEmphasis	0.388	0.550	0.595	0.648
	LongRunHighGrayLevelEmphasis	0.896	0.839	0.833	0.986
	LongRunLowGrayLevelEmphasis	0.394	0.308	0.476	0.417
	LowGrayLevelRunEmphasis	0.345	0.294	0.362	0.319
	RunEntropy	0.690	0.456	0.416	0.453
	RunLengthNonUniformity	0.318	0.430	0.552	0.700
	RunLengthNonUniformityNormalized	0.651	0.700	0.637	0.560
	RunPercentage	0.581	0.633	0.600	0.560
	RunVariance	0.319	0.535	0.618	0.714
	ShortRunEmphasis	0.657	0.702	0.636	0.557
GLSZM	ShortRunHighGrayLevelEmphasis	0.582	0.626	0.875	0.918
	ShortRunLowGrayLevelEmphasis	0.230	0.152	0.347	0.244
	GrayLevelNonUniformity	0.547	0.520	0.610	0.341
	GrayLevelNonUniformityNormalized	0.965	0.826	0.934	0.849
	GrayLevelVariance	0.989	0.984	0.960	0.950
	HighGrayLevelZoneEmphasis	0.688	0.860	0.938	0.658

NGTDM	LargeAreaEmphasis	0.824	0.658	0.711	0.462
	LargeAreaHighGrayLevelEmphasis	0.879	0.864	0.721	0.358
	LargeAreaLowGrayLevelEmphasis	0.790	0.538	0.652	0.756
	LowGrayLevelZoneEmphasis	0.030	0.013	0.166	0.268
	SizeZoneNonUniformity	0.948	0.741	0.500	0.460
	SizeZoneNonUniformityNormalized	0.702	0.458	0.658	0.867
	SmallAreaEmphasis	0.704	0.567	0.442	0.796
	SmallAreaHighGrayLevelEmphasis	0.761	0.898	0.920	0.955
	SmallAreaLowGrayLevelEmphasis	0.035	0.135	0.206	0.388
	ZoneEntropy	0.637	0.704	0.862	0.458
	ZonePercentage	0.773	0.560	0.356	0.566
	ZoneVariance	0.824	0.700	0.704	0.451
	Busyness	0.637	0.677	0.809	0.803
	Coarseness	0.389	0.330	0.312	0.408
	Complexity	0.776	0.749	0.833	0.839
	Contrast	0.993	0.893	0.879	0.935
	Strength	0.884	0.866	0.890	0.890

Table 48. P-values for ANOVA or Kruskal-Wallis tests for univariate analysis to assess the risk of progression to MM in patients with SMM in PETHEMA labelling. Variables with a p-value lower than 0.05, i.e. with significant differences between classes, for any of the bin widths are marked in green and bold type. P-values corrected by Benjamini-Hochberg correction are not shown, as none obtain values which means significant differences between classes. Gr: group of features.

Annex K. BM quantification in NHL patients by ML models

Following tables show F1_macro and F1_weighted of the 44 trained ML models for the three cohorts: NHL in Table 49, Table 50 and Table 51 for PET, MFC and pGT labelling methods, respectively; FL in Table 52, Table 53 and Table 54 for PET, MFC and pGT labelling methods, respectively; DLBCL in Table 55 Table 56 and Table 57 for PET, MFC and pGT labelling methods, respectively.

		PET - NHL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.661	0.709	0.674	0.711	0.579	0.596	0.706	0.722
	RBF	0.445	0.425	0.445	0.425	0.562	0.585	0.445	0.425
	LIN	0.545	0.507	0.614	0.661	0.673	0.708	0.611	0.672
	SIG	0.451	0.477	0.561	0.599	0.638	0.647	0.613	0.648
	RF	0.704	0.751	0.724	0.761	0.681	0.701	0.691	0.717
	GB	0.678	0.745	0.664	0.716	0.744	0.824	0.667	0.718
	LR	0.786	0.834	0.803	0.818	0.773	0.800	0.687	0.749
	NB	0.652	0.688	0.626	0.656	0.636	0.666	0.574	0.585
	k-NN	0.648	0.692	0.683	0.740	0.648	0.692	0.648	0.692
	MLP1	0.637	0.677	0.621	0.655	0.637	0.677	0.761	0.786
	MLP2	0.739	0.779	0.724	0.761	0.755	0.781	0.689	0.763
R	DT	0.694	0.739	0.724	0.716	0.600	0.666	0.472	0.453
	RBF	0.530	0.581	0.509	0.547	0.562	0.588	0.539	0.571
	LIN	0.548	0.559	0.498	0.520	0.390	0.342	0.296	0.236
	SIG	0.452	0.471	0.497	0.525	0.638	0.650	0.457	0.468
	RF	0.721	0.753	0.710	0.724	0.647	0.644	0.659	0.675
	GB	0.705	0.766	0.667	0.718	0.660	0.712	0.705	0.772
	LR	0.713	0.742	0.714	0.746	0.748	0.792	0.744	0.783
	NB	0.621	0.671	0.583	0.611	0.623	0.651	0.610	0.634
	k-NN	0.662	0.727	0.650	0.695	0.636	0.696	0.725	0.728
	MLP1	0.632	0.649	0.660	0.706	0.638	0.680	0.719	0.785
	MLP2	0.701	0.746	0.731	0.761	0.738	0.775	0.745	0.788
OS	DT	0.694	0.739	0.612	0.629	0.525	0.556	0.625	0.629
	RBF	0.595	0.624	0.523	0.521	0.608	0.620	0.445	0.425
	LIN	0.327	0.309	0.618	0.665	0.522	0.544	0.552	0.583
	SIG	0.452	0.471	0.491	0.517	0.638	0.650	0.245	0.262

OS_R	RF	0.745	0.788	0.730	0.770	0.647	0.644	0.726	0.745
	GB	0.695	0.755	0.822	0.843	0.787	0.810	0.713	0.779
	LR	0.770	0.805	0.770	0.805	0.803	0.818	0.752	0.781
	NB	0.652	0.688	0.602	0.623	0.245	0.262	0.245	0.262
	k-NN	0.695	0.755	0.601	0.625	0.553	0.562	0.677	0.735
	MLP1	0.696	0.758	0.733	0.805	0.687	0.749	0.678	0.745
	MLP2	0.649	0.706	0.691	0.769	0.649	0.706	0.677	0.735
	DT	0.678	0.724	0.590	0.595	0.584	0.576	0.532	0.537
	RBF	0.518	0.565	0.461	0.459	0.562	0.588	0.539	0.571
	LIN	0.289	0.288	0.577	0.616	0.296	0.236	0.296	0.236
	SIG	0.452	0.471	0.558	0.606	0.562	0.576	0.557	0.596
	RF	0.712	0.744	0.712	0.720	0.623	0.619	0.695	0.703
	GB	0.661	0.710	0.758	0.773	0.661	0.710	0.688	0.679
	LR	0.654	0.712	0.666	0.678	0.612	0.636	0.615	0.611
	NB	0.606	0.650	0.596	0.627	0.605	0.612	0.624	0.633
	k-NN	0.628	0.660	0.641	0.679	0.732	0.738	0.757	0.742
	MLP1	0.621	0.693	0.688	0.762	0.792	0.803	0.690	0.693
	MLP2	0.612	0.678	0.685	0.674	0.672	0.683	0.638	0.680

Table 49. Performance metrics for the ML model for the PET labelling in NHL patients for BM classification. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		MFC - NHL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.483	0.487	0.466	0.479	0.482	0.481	0.435	0.449
	RBF	0.328	0.300	0.328	0.300	0.328	0.300	0.328	0.300
	LIN	0.566	0.541	0.435	0.387	0.409	0.392	0.370	0.350
	SIG	0.464	0.484	0.281	0.308	0.281	0.308	0.460	0.440
	RF	0.539	0.567	0.484	0.507	0.508	0.534	0.435	0.446
	GB	0.578	0.610	0.518	0.545	0.526	0.544	0.596	0.657
	LR	0.611	0.622	0.561	0.598	0.544	0.593	0.596	0.661
	NB	0.607	0.643	0.616	0.619	0.593	0.587	0.659	0.644
	k-NN	0.524	0.560	0.506	0.533	0.603	0.652	0.524	0.552
	MLP1	0.586	0.630	0.495	0.518	0.546	0.580	0.560	0.606
	MLP2	0.539	0.567	0.564	0.610	0.609	0.663	0.582	0.631
OS	DT	0.480	0.482	0.502	0.509	0.485	0.512	0.442	0.399

RBF	0.478	0.483	0.504	0.482	0.493	0.459	0.459	0.447
LIN	0.544	0.539	0.441	0.413	0.533	0.552	0.395	0.366
SIG	0.477	0.482	0.468	0.460	0.522	0.558	0.652	0.637
RF	0.523	0.557	0.469	0.485	0.515	0.550	0.465	0.481
GB	0.552	0.602	0.560	0.602	0.542	0.595	0.540	0.585
LR	0.601	0.667	0.585	0.636	0.547	0.589	0.565	0.607
NB	0.565	0.590	0.594	0.603	0.599	0.608	0.619	0.624
k-NN	0.508	0.546	0.548	0.586	0.484	0.507	0.484	0.507
MLP1	0.574	0.624	0.555	0.600	0.520	0.557	0.586	0.634
MLP2	0.539	0.577	0.583	0.643	0.560	0.606	0.549	0.595

Table 50. Performance metrics for the ML model for the MFC labelling in NHL patients for BM classification. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		pGT - NHL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.554	0.538	0.509	0.524	0.477	0.470	0.491	0.464
	RBF	0.444	0.442	0.513	0.511	0.543	0.549	0.486	0.484
	LIN	0.533	0.552	0.540	0.569	0.446	0.445	0.338	0.303
	SIG	0.556	0.571	0.502	0.512	0.561	0.590	0.579	0.592
	RF	0.588	0.603	0.589	0.599	0.570	0.570	0.608	0.632
	GB	0.614	0.660	0.605	0.647	0.632	0.679	0.607	0.645
	LR	0.532	0.538	0.505	0.523	0.517	0.557	0.491	0.499
	NB	0.502	0.515	0.539	0.571	0.626	0.663	0.562	0.590
	k-NN	0.566	0.596	0.566	0.596	0.549	0.570	0.562	0.596
	MLP1	0.605	0.617	0.662	0.672	0.524	0.537	0.585	0.604
	MLP2	0.543	0.558	0.585	0.618	0.551	0.581	0.570	0.601
R	DT	0.536	0.570	0.468	0.483	0.462	0.473	0.545	0.556
	RBF	0.493	0.492	0.469	0.477	0.514	0.533	0.485	0.488
	LIN	0.583	0.609	0.411	0.438	0.588	0.621	0.498	0.499
	SIG	0.517	0.552	0.412	0.448	0.605	0.615	0.501	0.528
	RF	0.551	0.581	0.521	0.544	0.535	0.534	0.617	0.631
	GB	0.629	0.679	0.639	0.686	0.585	0.621	0.627	0.672
	LR	0.568	0.583	0.505	0.528	0.518	0.554	0.533	0.554
	NB	0.556	0.587	0.516	0.548	0.551	0.574	0.564	0.588
	k-NN	0.549	0.570	0.587	0.619	0.597	0.632	0.619	0.656
	MLP1	0.526	0.546	0.569	0.601	0.522	0.520	0.476	0.490
	MLP2	0.447	0.445	0.591	0.595	0.579	0.596	0.662	0.672

OS	DT	0.502	0.526	0.555	0.568	0.538	0.542	0.573	0.582
	RBF	0.460	0.471	0.508	0.499	0.527	0.536	0.430	0.440
	LIN	0.686	0.754	0.474	0.484	0.509	0.514	0.439	0.410
	SIG	0.400	0.391	0.456	0.454	0.405	0.402	0.594	0.614
	RF	0.589	0.599	0.610	0.637	0.578	0.584	0.612	0.632
	GB	0.570	0.602	0.599	0.645	0.587	0.628	0.625	0.678
	LR	0.555	0.594	0.571	0.577	0.579	0.596	0.535	0.547
	NB	0.508	0.535	0.600	0.632	0.572	0.593	0.553	0.567
	k-NN	0.583	0.618	0.597	0.643	0.571	0.624	0.565	0.596
	MLP1	0.512	0.532	0.558	0.597	0.542	0.576	0.568	0.573
	MLP2	0.541	0.534	0.553	0.591	0.532	0.562	0.508	0.529
OS_R	DT	0.506	0.532	0.481	0.496	0.526	0.545	0.486	0.526
	RBF	0.454	0.465	0.568	0.578	0.507	0.515	0.549	0.569
	LIN	0.533	0.552	0.468	0.500	0.527	0.537	0.509	0.514
	SIG	0.406	0.380	0.445	0.476	0.607	0.651	0.485	0.527
	RF	0.546	0.577	0.506	0.521	0.502	0.523	0.611	0.642
	GB	0.632	0.682	0.578	0.621	0.562	0.599	0.580	0.613
	LR	0.524	0.554	0.572	0.600	0.570	0.589	0.523	0.555
	NB	0.469	0.505	0.507	0.530	0.542	0.563	0.565	0.595
	k-NN	0.514	0.530	0.596	0.634	0.545	0.558	0.540	0.564
	MLP1	0.595	0.603	0.478	0.481	0.568	0.543	0.534	0.515
	MLP2	0.596	0.609	0.484	0.495	0.521	0.510	0.469	0.487

Table 51. Performance metrics for the ML model for the pGT labelling in NHL patients for BM classification. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		PET - FL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.572	0.628	0.367	0.327	0.627	0.707	0.691	0.783
	RBF	0.507	0.588	0.522	0.582	0.560	0.648	0.439	0.414
	LIN	0.439	0.414	0.652	0.738	0.683	0.782	0.627	0.705
	SIG	0.575	0.637	0.575	0.637	0.550	0.608	0.236	0.282
	RF	0.683	0.782	0.652	0.738	0.710	0.815	0.663	0.750
	GB	0.683	0.782	0.683	0.782	0.683	0.782	0.683	0.782
	LR	0.679	0.771	0.663	0.750	0.663	0.750	0.710	0.815
	NB	0.611	0.711	0.648	0.722	0.598	0.661	0.622	0.692
	k-NN	0.612	0.656	0.439	0.414	0.674	0.743	0.439	0.414

	MLP1	0.663	0.750	0.644	0.719	0.616	0.679	0.608	0.675
	MLP2	0.606	0.669	0.644	0.719	0.616	0.679	0.623	0.710
R	DT	0.580	0.631	0.741	0.774	0.669	0.754	0.638	0.762
	RBF	0.439	0.414	0.439	0.414	0.526	0.618	0.585	0.679
	LIN	0.561	0.614	0.236	0.282	0.536	0.549	0.652	0.738
	SIG	0.602	0.669	0.575	0.637	0.550	0.608	0.560	0.648
	RF	0.812	0.962	0.699	0.802	0.740	0.860	0.690	0.786
	GB	0.683	0.782	0.699	0.802	0.683	0.782	0.731	0.849
	LR	0.784	0.783	0.810	0.799	0.731	0.790	0.784	0.819
	NB	0.585	0.679	0.536	0.549	0.538	0.597	0.508	0.507
	k-NN	0.612	0.656	0.612	0.656	0.612	0.656	0.683	0.782
	MLP1	0.710	0.817	0.641	0.719	0.679	0.771	0.650	0.730
	MLP2	0.652	0.738	0.691	0.783	0.690	0.786	0.720	0.829
OS	DT	0.552	0.603	0.506	0.502	0.544	0.588	0.714	0.842
	RBF	0.507	0.588	0.522	0.582	0.560	0.648	0.439	0.414
	LIN	0.627	0.705	0.326	0.347	0.403	0.422	0.418	0.387
	SIG	0.592	0.608	0.585	0.679	0.558	0.630	0.344	0.420
	RF	0.624	0.697	0.836	0.768	0.784	0.819	0.663	0.750
	GB	0.683	0.782	0.683	0.782	0.699	0.802	0.683	0.782
	LR	0.767	0.791	0.663	0.750	0.663	0.750	0.633	0.708
	NB	0.585	0.679	0.622	0.692	0.648	0.722	0.622	0.692
	k-NN	0.521	0.554	0.536	0.549	0.642	0.699	0.428	0.398
	MLP1	0.644	0.719	0.624	0.697	0.616	0.679	0.606	0.669
	MLP2	0.606	0.669	0.680	0.793	0.623	0.710	0.606	0.669
OS_R	DT	0.652	0.734	0.497	0.486	0.662	0.745	0.618	0.696
	RBF	0.507	0.588	0.522	0.582	0.526	0.618	0.585	0.706
	LIN	0.403	0.372	0.439	0.414	0.315	0.246	0.699	0.802
	SIG	0.659	0.671	0.565	0.579	0.537	0.618	0.550	0.595
	RF	0.670	0.782	0.774	0.909	0.797	0.835	0.695	0.810
	GB	0.683	0.782	0.683	0.782	0.731	0.849	0.683	0.782
	LR	0.794	0.795	0.640	0.690	0.608	0.675	0.644	0.685
	NB	0.585	0.679	0.538	0.597	0.622	0.692	0.503	0.547
	k-NN	0.624	0.697	0.624	0.697	0.627	0.705	0.683	0.782
	MLP1	0.720	0.700	0.563	0.604	0.679	0.771	0.679	0.771
	MLP2	0.669	0.759	0.650	0.730	0.679	0.771	0.672	0.756

Table 52. Performance metrics for the ML model for the PET labelling in FL patients for BM classification. Outstanding values, i.e. higher than 0.85, are marked in green. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are

marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		MFC - FL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.497	0.528	0.458	0.472	0.545	0.509	0.495	0.513
	RBF	0.387	0.392	0.425	0.461	0.398	0.394	0.401	0.371
	LIN	0.484	0.492	0.415	0.387	0.558	0.522	0.340	0.330
	SIG	0.451	0.430	0.391	0.367	0.436	0.416	0.387	0.353
	RF	0.525	0.554	0.538	0.572	0.493	0.504	0.481	0.492
	GB	0.429	0.413	0.445	0.428	0.478	0.481	0.462	0.452
	LR	0.533	0.567	0.517	0.544	0.482	0.500	0.505	0.526
	NB	0.416	0.391	0.436	0.418	0.436	0.418	0.436	0.418
	k-NN	0.495	0.503	0.440	0.428	0.435	0.428	0.452	0.448
	MLP1	0.429	0.413	0.570	0.622	0.544	0.582	0.605	0.532
	MLP2	0.589	0.631	0.492	0.516	0.524	0.557	0.489	0.506
R	DT							0.543	0.521
	RBF							0.491	0.508
	LIN							0.468	0.466
	SIG							0.455	0.467
	RF							0.546	0.517
	GB							0.462	0.452
	LR							0.504	0.515
	NB							0.445	0.428
	k-NN							0.429	0.413
	MLP1							0.445	0.428
	MLP2							0.421	0.397
OS	DT	0.386	0.375	0.452	0.452	0.530	0.553	0.443	0.438
	RBF	0.474	0.474	0.485	0.466	0.452	0.442	0.401	0.371
	LIN	0.445	0.428	0.381	0.389	0.425	0.412	0.358	0.319
	SIG	0.377	0.385	0.471	0.465	0.513	0.477	0.387	0.353
	RF	0.508	0.531	0.469	0.477	0.450	0.454	0.481	0.492
	GB	0.545	0.580	0.463	0.476	0.472	0.478	0.474	0.474
	LR	0.457	0.465	0.583	0.634	0.536	0.582	0.497	0.518
	NB	0.458	0.470	0.436	0.418	0.436	0.418	0.436	0.418
	k-NN	0.468	0.466	0.454	0.454	0.489	0.506	0.469	0.477

OS_R	MLP1	0.489	0.506	0.554	0.594	0.620	0.683	0.545	0.580
	MLP2	0.431	0.426	0.508	0.531	0.544	0.582	0.460	0.466
	DT							0.540	0.510
	RBF							0.504	0.490
	LIN							0.441	0.460
	SIG							0.421	0.397
	RF							0.569	0.476
	GB							0.474	0.474
	LR							0.504	0.515
	NB							0.445	0.428
	k-NN							0.452	0.452
	MLP1							0.454	0.461
	MLP2							0.449	0.446

Table 53. Performance metrics for the ML model for the MFC labelling in FL patients for BM classification. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		pGT - FL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.673	0.694	0.495	0.520	0.639	0.705	0.557	0.591
	RBF	0.388	0.405	0.403	0.446	0.450	0.480	0.382	0.345
	LIN	0.469	0.527	0.382	0.345	0.419	0.470	0.432	0.438
	SIG	0.434	0.429	0.584	0.609	0.438	0.441	0.382	0.345
	RF	0.580	0.622	0.569	0.603	0.520	0.536	0.563	0.603
	GB	0.547	0.573	0.547	0.573	0.556	0.590	0.557	0.591
	LR	0.555	0.602	0.563	0.603	0.622	0.686	0.563	0.603
	NB	0.487	0.495	0.471	0.468	0.423	0.403	0.471	0.468
	k-NN	0.544	0.568	0.462	0.456	0.464	0.458	0.530	0.550
	MLP1	0.554	0.601	0.571	0.619	0.613	0.668	0.537	0.567
	MLP2	0.538	0.570	0.659	0.727	0.585	0.629	0.603	0.650
R	DT	0.642	0.676	0.537	0.567	0.612	0.609	0.567	0.585
	RBF	0.388	0.405	0.403	0.446	0.550	0.596	0.525	0.514
	LIN	0.260	0.328	0.260	0.328	0.417	0.413	0.577	0.614
	SIG	0.497	0.516	0.583	0.596	0.494	0.547	0.417	0.454
	RF	0.592	0.643	0.640	0.706	0.560	0.539	0.566	0.608
	GB	0.565	0.598	0.516	0.530	0.487	0.491	0.630	0.688
	LR	0.572	0.622	0.600	0.645	0.674	0.696	0.556	0.590

	NB	0.382	0.345	0.382	0.345	0.440	0.425	0.440	0.425
	k-NN	0.518	0.533	0.516	0.530	0.576	0.612	0.576	0.612
	MLP1	0.470	0.472	0.502	0.515	0.492	0.497	0.527	0.549
	MLP2	0.576	0.612	0.459	0.452	0.462	0.456	0.495	0.506
OS	DT	0.544	0.583	0.575	0.611	0.592	0.658	0.538	0.570
	RBF	0.429	0.451	0.438	0.485	0.515	0.550	0.382	0.345
	LIN	0.577	0.614	0.382	0.345	0.436	0.420	0.382	0.345
	SIG	0.490	0.526	0.422	0.404	0.498	0.510	0.382	0.345
	RF	0.640	0.706	0.611	0.665	0.557	0.591	0.613	0.668
	GB	0.629	0.685	0.613	0.668	0.520	0.536	0.492	0.497
	LR	0.564	0.620	0.584	0.643	0.584	0.643	0.594	0.646
	NB	0.469	0.490	0.471	0.468	0.423	0.403	0.471	0.468
	k-NN	0.520	0.535	0.462	0.456	0.437	0.421	0.434	0.417
	MLP1	0.584	0.629	0.503	0.521	0.622	0.686	0.641	0.707
	MLP2	0.443	0.444	0.527	0.554	0.611	0.680	0.613	0.668
OS_R	DT	0.618	0.678	0.592	0.660	0.573	0.515	0.533	0.516
	RBF	0.388	0.405	0.438	0.485	0.534	0.557	0.553	0.540
	LIN	0.382	0.345	0.260	0.328	0.486	0.549	0.382	0.345
	SIG	0.473	0.503	0.456	0.450	0.541	0.564	0.484	0.491
	RF	0.607	0.659	0.603	0.649	0.740	0.745	0.562	0.523
	GB	0.549	0.580	0.514	0.528	0.544	0.568	0.512	0.524
	LR	0.565	0.606	0.537	0.567	0.622	0.685	0.657	0.631
	NB	0.382	0.345	0.382	0.345	0.440	0.425	0.440	0.425
	k-NN	0.560	0.600	0.565	0.598	0.544	0.568	0.491	0.500
	MLP1	0.515	0.543	0.516	0.530	0.520	0.536	0.512	0.524
	MLP2	0.462	0.470	0.464	0.459	0.574	0.610	0.503	0.521

Table 54. Performance metrics for the ML model for the pGT labelling in FL patients for BM classification. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		PET - DLBCL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.621	0.629	0.605	0.564	0.571	0.584	0.443	0.531
	RBF	0.480	0.483	0.453	0.458	0.579	0.539	0.449	0.431
	LIN	0.483	0.527	0.597	0.605	0.437	0.416	0.595	0.638
	SIG	0.512	0.526	0.483	0.527	0.606	0.573	0.597	0.605
	RF	0.606	0.573	0.598	0.567	0.598	0.567	0.598	0.567

	GB	0.577	0.558	0.577	0.558	0.717	0.690	0.757	0.741
	LR	0.606	0.573	0.635	0.610	0.579	0.539	0.579	0.539
	NB	0.643	0.608	0.582	0.531	0.674	0.646	0.674	0.646
	k-NN	0.667	0.647	0.449	0.431	0.667	0.647	0.573	0.613
	MLP1	0.595	0.650	0.595	0.650	0.635	0.610	0.541	0.572
	MLP2	0.595	0.650	0.511	0.512	0.500	0.505	0.500	0.505
R	DT	0.606	0.573	0.606	0.573	0.635	0.610	0.593	0.620
	RBF	0.510	0.521	0.505	0.525	0.505	0.525	0.570	0.597
	LIN	0.449	0.431	0.258	0.244	0.449	0.431	0.465	0.338
	SIG	0.471	0.415	0.440	0.405	0.498	0.523	0.541	0.563
	RF	0.606	0.573	0.606	0.573	0.667	0.647	0.531	0.544
	GB	0.667	0.647	0.757	0.741	0.667	0.644	0.474	0.539
	LR	0.598	0.567	0.606	0.573	0.598	0.567	0.606	0.573
	NB	0.635	0.610	0.598	0.567	0.606	0.573	0.562	0.598
	k-NN	0.595	0.650	0.625	0.600	0.577	0.573	0.577	0.558
	MLP1	0.539	0.494	0.706	0.690	0.577	0.558	0.716	0.754
	MLP2	0.577	0.558	0.626	0.568	0.411	0.381	0.757	0.741
OS	DT	0.645	0.613	0.566	0.555	0.453	0.503	0.611	0.616
	RBF	0.413	0.389	0.453	0.458	0.593	0.592	0.449	0.431
	LIN	0.532	0.561	0.608	0.569	0.583	0.585	0.625	0.600
	SIG	0.556	0.548	0.446	0.468	0.461	0.431	0.277	0.233
	RF	0.606	0.573	0.598	0.567	0.598	0.567	0.598	0.567
	GB	0.573	0.613	0.619	0.691	0.573	0.613	0.562	0.598
	LR	0.598	0.567	0.667	0.647	0.606	0.573	0.606	0.573
	NB	0.643	0.608	0.582	0.531	0.643	0.608	0.674	0.646
	k-NN	0.667	0.647	0.635	0.610	0.552	0.579	0.560	0.563
	MLP1	0.577	0.573	0.595	0.650	0.706	0.690	0.573	0.613
	MLP2	0.667	0.647	0.667	0.647	0.573	0.613	0.552	0.579
OS_R	DT	0.593	0.592	0.608	0.569	0.635	0.610	0.675	0.651
	RBF	0.612	0.570	0.505	0.525	0.505	0.525	0.505	0.525
	LIN	0.449	0.431	0.597	0.605	0.258	0.225	0.449	0.431
	SIG	0.372	0.375	0.532	0.561	0.442	0.403	0.369	0.277
	RF	0.606	0.573	0.606	0.573	0.606	0.573	0.606	0.573
	GB	0.667	0.647	0.667	0.647	0.667	0.644	0.518	0.612
	LR	0.598	0.567	0.635	0.603	0.598	0.563	0.532	0.561
	NB	0.635	0.610	0.570	0.532	0.635	0.610	0.562	0.598
	k-NN	0.667	0.647	0.706	0.690	0.667	0.644	0.542	0.522
	MLP1	0.667	0.644	0.706	0.690	0.512	0.491	0.759	0.807

MLP2	0.667	0.644	0.654	0.607	0.542	0.522	0.625	0.600

Table 55. Performance metrics for the ML model for the PET labelling in DLBCL patients for BM classification. Good values, i.e. between 0.8 and 0.85, are marked in yellow. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		MFC - DLBCL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.388	0.336	0.548	0.560	0.596	0.555	0.521	0.450
	RBF	0.436	0.431	0.460	0.450	0.325	0.290	0.301	0.273
	LIN	0.415	0.391	0.318	0.267	0.585	0.542	0.676	0.671
	SIG	0.408	0.351	0.490	0.511	0.301	0.273	0.301	0.273
	RF	0.658	0.667	0.477	0.517	0.582	0.588	0.563	0.607
	GB	0.431	0.412	0.431	0.412	0.431	0.412	0.431	0.412
	LR	0.631	0.659	0.705	0.699	0.705	0.699	0.705	0.699
	NB	0.436	0.426	0.381	0.352	0.381	0.352	0.318	0.267
	k-NN	0.563	0.607	0.540	0.539	0.540	0.539	0.540	0.539
	MLP1	0.531	0.552	0.652	0.748	0.522	0.508	0.563	0.559
	MLP2	0.491	0.497	0.528	0.460	0.423	0.382	0.464	0.471
OS	DT	0.533	0.514	0.700	0.728	0.663	0.642	0.606	0.567
	RBF	0.448	0.441	0.460	0.450	0.442	0.352	0.301	0.273
	LIN	0.431	0.412	0.436	0.455	0.398	0.327	0.431	0.412
	SIG	0.452	0.412	0.490	0.511	0.449	0.343	0.318	0.267
	RF	0.715	0.724	0.563	0.607	0.715	0.724	0.536	0.566
	GB	0.431	0.412	0.431	0.412	0.431	0.412	0.415	0.391
	LR	0.491	0.497	0.502	0.507	0.428	0.423	0.399	0.380
	NB	0.486	0.493	0.436	0.426	0.436	0.426	0.381	0.352
	k-NN	0.560	0.595	0.525	0.471	0.606	0.612	0.499	0.518
	MLP1	0.572	0.604	0.531	0.552	0.531	0.552	0.531	0.552
	MLP2	0.464	0.471	0.531	0.552	0.522	0.508	0.543	0.561

Table 56. Performance metrics for the ML model for the MFC labelling in DLBCL patients for BM classification. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

		pGT - DLBCL							
		BW = 0.5		BW = 0.25		BW = 0.125		BW = 0.0625	
		F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w	F1_m*	F1_w
ORIG	DT	0.659	0.653	0.584	0.576	0.601	0.607	0.404	0.342
	RBF	0.439	0.454	0.432	0.414	0.432	0.414	0.432	0.414
	LIN	0.501	0.566	0.501	0.566	0.573	0.564	0.432	0.414
	SIG	0.507	0.543	0.537	0.525	0.486	0.420	0.288	0.275
	RF	0.744	0.732	0.629	0.604	0.635	0.603	0.599	0.567
	GB	0.573	0.564	0.573	0.564	0.432	0.414	0.404	0.377
	LR	0.643	0.640	0.616	0.651	0.643	0.640	0.616	0.651
	NB	0.442	0.452	0.503	0.474	0.547	0.532	0.363	0.355
	k-NN	0.702	0.689	0.663	0.644	0.663	0.644	0.704	0.686
	MLP1	0.616	0.651	0.584	0.576	0.676	0.683	0.614	0.614
	MLP2	0.539	0.535	0.567	0.549	0.575	0.571	0.562	0.568
R	DT					0.550	0.565	0.616	0.631
	RBF					0.537	0.525	0.666	0.657
	LIN					0.432	0.414	0.479	0.514
	SIG					0.573	0.564	0.573	0.564
	RF					0.601	0.607	0.605	0.624
	GB					0.537	0.525	0.506	0.490
	LR					0.540	0.528	0.668	0.643
	NB					0.467	0.433	0.389	0.313
	k-NN					0.520	0.504	0.668	0.643
	MLP1					0.488	0.463	0.495	0.448
	MLP2					0.479	0.458	0.453	0.426
OS	DT	0.585	0.610	0.585	0.610	0.668	0.643	0.659	0.653
	RBF	0.616	0.631	0.582	0.583	0.420	0.350	0.432	0.414
	LIN	0.399	0.377	0.439	0.454	0.432	0.414	0.573	0.564
	SIG	0.507	0.543	0.653	0.612	0.515	0.458	0.288	0.275
	RF	0.629	0.604	0.599	0.567	0.599	0.567	0.599	0.567
	GB	0.534	0.505	0.562	0.568	0.629	0.604	0.644	0.654
	LR	0.602	0.593	0.643	0.640	0.497	0.456	0.566	0.548
	NB	0.442	0.452	0.508	0.488	0.508	0.488	0.399	0.396
	k-NN	0.663	0.644	0.488	0.463	0.599	0.567	0.534	0.571
	MLP1	0.563	0.587	0.495	0.448	0.571	0.531	0.520	0.504
	MLP2	0.571	0.531	0.448	0.404	0.571	0.531	0.571	0.531
OS_R	DT					0.601	0.607	0.585	0.610
	RBF					0.668	0.643	0.374	0.361

LIN			0.497	0.513	0.596	0.629
SIG			0.401	0.336	0.361	0.315
RF			0.601	0.607	0.690	0.666
GB			0.591	0.563	0.453	0.426
LR			0.540	0.528	0.668	0.643
NB			0.467	0.433	0.427	0.364
k-NN			0.562	0.526	0.668	0.643
MLP1			0.420	0.370	0.453	0.426
MLP2			0.453	0.426	0.479	0.458

Table 57. Performance metrics for the ML model for the pGT labelling in DLBCL patients for BM classification. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type.

Annex L. BM quantification in NHL patients by CNNs

Below are the resulting tables for all tested CNN architectures separated by labelling pair and input type. For PET labelling, Table 58 (3D input), Table 59 (pRGB input) and Table 60 (MIP input) are included. For MFC labelling, Table 61(3D input), Table 62 (pRGB input) and Table 63 (MIP input) are included. For pGT labelling, Table 64 (3D input), Table 65 (pRGB input) and Table 66 (MIP input) are included.

3D - PET													
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.602	0.657	0.497	0.566	0.604	0.686	0.693	0.741	0.346	0.173	0.441	0.302
	2	0.710	0.753	0.360	0.090	0.560	0.636	0.627	0.702	0.701	0.752	0.301	0.062
	4	0.513	0.589	0.638	0.724	0.566	0.652	0.644	0.723	0.677	0.753	0.256	0.017
	8	0.677	0.697	0.608	0.690	0.629	0.647	0.592	0.670	0.724	0.741	0.361	0.146
VGG16	1	0.440	0.555	0.256	0.017	0.379	0.214	0.640	0.690	0.256	0.017	0.373	0.129
	2	0.671	0.738	0.256	0.017	0.256	0.017	0.440	0.555	0.256	0.017	0.256	0.017
	4	0.720	0.790	0.256	0.017	0.256	0.017	0.598	0.689	0.256	0.017	0.256	0.017
	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.256	0.017
ResNet	1	0.440	0.555	0.638	0.724	0.540	0.632	0.501	0.607	0.373	0.129	0.440	0.555
	2	0.671	0.738	0.553	0.651	0.545	0.622	0.501	0.607	0.373	0.129	0.440	0.555
	4	0.598	0.689	0.571	0.640	0.553	0.651	0.440	0.555	0.598	0.689	0.540	0.600
	8	0.638	0.724	0.620	0.701	0.522	0.615	0.598	0.689	0.502	0.579	0.501	0.607
DenseNet	1	0.553	0.651	0.539	0.623	0.541	0.643	0.553	0.651	0.256	0.017	0.373	0.129
	2	0.698	0.764	0.256	0.017	0.553	0.651	0.598	0.689	0.340	0.098	0.464	0.563
	4	0.598	0.689	0.484	0.571	0.368	0.112	0.598	0.689	0.652	0.733	0.454	0.358
	8	0.440	0.555	0.498	0.462	0.616	0.603	0.565	0.580	0.335	0.082	0.440	0.555
RPET_NET	1	0.513	0.589	0.612	0.683	0.705	0.749	0.501	0.507	0.536	0.550	0.527	0.524
	2	0.558	0.631	0.583	0.669	0.557	0.641	0.626	0.638	0.654	0.731	0.557	0.641
	4	0.566	0.652	0.590	0.627	0.552	0.636	0.491	0.590	0.573	0.583	0.497	0.580
	8	0.507	0.596	0.507	0.596	0.497	0.580	0.463	0.563	0.498	0.443	0.521	0.547

Table 58. Performance metrics for CNNs architectures for PET labelling and 3D input for BM status classification in NHL patients. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.

		pRGB - PET											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.550	0.606	0.558	0.613	0.625	0.649	0.650	0.724	0.299	0.021	0.517	0.400
	2	0.654	0.707	0.638	0.693	0.388	0.194	0.612	0.648	0.457	0.428	0.482	0.406
	4	0.589	0.626	0.544	0.593	0.631	0.690	0.583	0.631	0.592	0.635	0.513	0.489
	8	0.562	0.586	0.620	0.650	0.611	0.688	0.440	0.555	0.479	0.431	0.255	0.017
VGG16	1	0.440	0.555	0.358	0.086	0.256	0.017	0.440	0.555	0.256	0.017	0.727	0.764
	2	0.440	0.555	0.256	0.017	0.388	0.133	0.440	0.555	0.256	0.017	0.536	0.548
	4	0.487	0.564	0.654	0.679	0.516	0.455	0.440	0.555	0.256	0.017	0.256	0.017
	8	0.440	0.555	0.678	0.690	0.000	0.000	0.440	0.555	0.256	0.017	0.000	0.000
ResNet	1	0.440	0.555	0.291	0.072	0.554	0.607	0.440	0.555	0.283	0.064	0.484	0.571
	2	0.440	0.555	0.563	0.607	0.501	0.607	0.440	0.555	0.301	0.062	0.435	0.542
	4	0.440	0.555	0.553	0.651	0.553	0.651	0.501	0.607	0.501	0.607	0.501	0.607
	8	0.501	0.607	0.501	0.607	0.440	0.555	0.501	0.607	0.686	0.764	0.553	0.651
DenseNet	1	0.440	0.555	0.440	0.555	0.256	0.017	0.440	0.555	0.440	0.555	0.256	0.017
	2	0.440	0.555	0.440	0.555	0.440	0.555	0.440	0.555	0.256	0.017	0.256	0.017
	4	0.440	0.555	0.440	0.555	0.440	0.555	0.440	0.555	0.440	0.555	0.256	0.017
	8	0.440	0.555	0.440	0.555	0.256	0.017	0.440	0.555	0.440	0.555	0.299	0.079
RPET_NET	1	0.510	0.599	0.451	0.520	0.559	0.609	0.711	0.751	0.492	0.455	0.532	0.506
	2	0.596	0.627	0.442	0.525	0.449	0.538	0.560	0.628	0.623	0.585	0.557	0.548
	4	0.500	0.571	0.522	0.537	0.577	0.651	0.632	0.686	0.545	0.468	0.578	0.534
	8	0.593	0.653	0.507	0.596	0.568	0.649	0.596	0.623	0.527	0.443	0.566	0.579

Table 59. Performance metrics for CNNs architectures for PET labelling and pRGB input for BM status classification in NHL patients. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.

		MIP - PET											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.577	0.651	0.529	0.591	0.627	0.691	0.615	0.648	0.597	0.641	0.535	0.473
	2	0.684	0.730	0.618	0.685	0.541	0.609	0.638	0.724	0.530	0.466	0.595	0.537
	4	0.563	0.632	0.685	0.687	0.412	0.221	0.555	0.619	0.462	0.363	0.564	0.521
	8	0.526	0.589	0.491	0.530	0.531	0.575	0.527	0.601	0.330	0.070	0.256	0.017
VGG16	1	0.440	0.555	0.416	0.128	0.256	0.017	0.677	0.753	0.256	0.017	0.324	0.084
	2	0.552	0.636	0.466	0.513	0.672	0.698	0.440	0.555	0.329	0.074	0.256	0.017
	4	0.725	0.770	0.732	0.731	0.402	0.123	0.632	0.686	0.399	0.137	0.479	0.312
	8	0.645	0.714	0.592	0.670	0.532	0.563	0.501	0.607	0.256	0.017	0.256	0.017
Re	1	0.638	0.724	0.440	0.555	0.628	0.711	0.501	0.607	0.482	0.561	0.256	0.017

RPET_NET	DenseNet	2	0.464	0.563	0.618	0.690	0.501	0.607	0.440	0.555	0.299	0.079	0.440	0.555
		4	0.440	0.555	0.406	0.298	0.464	0.563	0.501	0.607	0.440	0.555	0.553	0.651
		8	0.574	0.588	0.291	0.072	0.391	0.297	0.502	0.579	0.501	0.607	0.601	0.678
		1	0.299	0.079	0.440	0.555	0.440	0.555	0.440	0.555	0.256	0.017	0.440	0.555
	DenseNet	2	0.440	0.555	0.501	0.607	0.440	0.555	0.440	0.555	0.440	0.555	0.440	0.555
		4	0.650	0.715	0.440	0.555	0.440	0.555	0.440	0.555	0.440	0.555	0.440	0.555
		8	0.440	0.555	0.440	0.555	0.440	0.555	0.440	0.555	0.440	0.555	0.256	0.017
		1	0.552	0.636	0.649	0.719	0.557	0.641	0.554	0.630	0.479	0.386	0.483	0.514
	RPET_NET	2	0.604	0.645	0.532	0.619	0.491	0.590	0.553	0.651	0.553	0.596	0.632	0.608
		4	0.592	0.670	0.552	0.636	0.608	0.690	0.550	0.622	0.564	0.608	0.587	0.612
		8	0.518	0.611	0.598	0.689	0.608	0.690	0.501	0.607	0.612	0.637	0.553	0.651

Table 60. Performance metrics for CNNs architectures for PET labelling and MIP input for BM status classification in NHL patients. Acceptable values, i.e. between 0.7 and 0.8, are marked in orange. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.

3D - MFC													
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.617	0.537	0.546	0.456	0.512	0.432	0.573	0.500	0.630	0.557	0.543	0.420
	2	0.476	0.358	0.494	0.390	0.502	0.386	0.397	0.334	0.623	0.518	0.512	0.411
	4	0.506	0.423	0.379	0.219	0.456	0.370	0.564	0.523	0.302	0.063	0.554	0.423
	8	0.577	0.510	0.531	0.358	0.446	0.364	0.529	0.484	0.302	0.063	0.411	0.167
VGG16	1	0.338	0.143	0.555	0.472	0.320	0.066	0.397	0.334	0.302	0.063	0.302	0.063
	2	0.397	0.334	0.302	0.063	0.497	0.390	0.397	0.334	0.302	0.063	0.302	0.063
	4	0.397	0.334	0.338	0.143	0.302	0.063	0.397	0.334	0.302	0.063	0.302	0.063
	8	0.397	0.334	0.302	0.063	0.302	0.063	0.397	0.334	0.302	0.063	0.302	0.063
ResNet	1	0.397	0.334	0.310	0.065	0.500	0.423	0.397	0.334	0.485	0.437	0.381	0.130
	2	0.497	0.403	0.543	0.449	0.465	0.368	0.490	0.440	0.281	0.041	0.438	0.188
	4	0.500	0.440	0.361	0.253	0.412	0.272	0.446	0.391	0.337	0.143	0.421	0.372
	8	0.518	0.437	0.490	0.440	0.595	0.539	0.548	0.450	0.404	0.365	0.404	0.365
DenseNet	1	0.499	0.472	0.559	0.533	0.404	0.365	0.539	0.517	0.291	0.048	0.371	0.144
	2	0.452	0.386	0.404	0.182	0.404	0.365	0.452	0.386	0.337	0.143	0.291	0.048
	4	0.423	0.334	0.404	0.365	0.291	0.048	0.454	0.423	0.607	0.569	0.291	0.048
	8	0.499	0.472	0.404	0.365	0.291	0.048	0.404	0.365	0.339	0.136	0.291	0.048
RPET_NET	1	0.504	0.446	0.538	0.482	0.529	0.480	0.505	0.456	0.497	0.427	0.459	0.380
	2	0.543	0.463	0.467	0.385	0.545	0.471	0.553	0.466	0.527	0.446	0.496	0.428
	4	0.591	0.517	0.531	0.450	0.536	0.467	0.500	0.435	0.557	0.491	0.497	0.427
	8	0.494	0.439	0.456	0.388	0.417	0.334	0.461	0.403	0.452	0.349	0.409	0.308

Table 61. Performance metrics for CNNs architectures for MFC labelling and 3D input for BM status classification in NHL patients. The maximum value for each group and labelling set is

ResNet	4	0.397	0.334	0.446	0.353	0.510	0.431	0.446	0.391	0.302	0.063	0.423	0.183
	8	0.500	0.379	0.456	0.360	0.487	0.352	0.397	0.334	0.352	0.117	0.302	0.063
	1	0.404	0.365	0.333	0.100	0.404	0.365	0.499	0.472	0.291	0.048	0.291	0.048
	2	0.404	0.365	0.454	0.423	0.454	0.423	0.454	0.423	0.454	0.423	0.291	0.048
	4	0.603	0.542	0.586	0.533	0.454	0.361	0.404	0.365	0.470	0.404	0.539	0.517
DenseNet	8	0.524	0.495	0.391	0.239	0.404	0.365	0.413	0.358	0.591	0.515	0.559	0.533
	1	0.337	0.143	0.404	0.365	0.337	0.143	0.337	0.143	0.337	0.143	0.291	0.048
	2	0.404	0.365	0.404	0.365	0.337	0.143	0.404	0.365	0.291	0.048	0.337	0.143
	4	0.404	0.365	0.291	0.048	0.404	0.365	0.404	0.365	0.404	0.365	0.404	0.365
	8	0.457	0.340	0.404	0.365	0.337	0.143	0.404	0.365	0.291	0.048	0.291	0.048
RPET_NET	1	0.542	0.451	0.569	0.521	0.555	0.482	0.454	0.423	0.465	0.337	0.480	0.341
	2	0.471	0.372	0.443	0.366	0.492	0.399	0.498	0.391	0.503	0.423	0.484	0.369
	4	0.402	0.337	0.430	0.352	0.495	0.400	0.573	0.512	0.493	0.400	0.510	0.417
	8	0.508	0.443	0.440	0.363	0.523	0.446	0.560	0.512	0.552	0.480	0.407	0.294

Table 63. Performance metrics for CNNs architectures for MFC labelling and MIP input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.

		3D - pGT											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
F	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	
AlexNet	1	0.513	0.447	0.552	0.527	0.562	0.537	0.505	0.440	0.592	0.543	0.347	0.174
	2	0.581	0.557	0.600	0.562	0.542	0.486	0.610	0.593	0.393	0.181	0.300	0.074
	4	0.560	0.530	0.369	0.167	0.418	0.201	0.666	0.681	0.274	0.026	0.361	0.169
	8	0.595	0.583	0.615	0.563	0.274	0.026	0.405	0.375	0.274	0.026	0.274	0.026
VGG16	1	0.603	0.582	0.388	0.152	0.274	0.026	0.405	0.375	0.308	0.072	0.274	0.026
	2	0.530	0.501	0.561	0.456	0.274	0.026	0.405	0.375	0.274	0.026	0.274	0.026
	4	0.405	0.375	0.274	0.026	0.274	0.026	0.405	0.375	0.274	0.026	0.274	0.026
	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ResNet	1	0.517	0.436	0.536	0.478	0.308	0.072	0.405	0.375	0.274	0.026	0.274	0.026
	2	0.661	0.662	0.583	0.587	0.512	0.485	0.405	0.375	0.405	0.375	0.405	0.375
	4	0.477	0.423	0.439	0.413	0.501	0.397	0.405	0.375	0.491	0.422	0.568	0.515
	8	0.595	0.597	0.568	0.556	0.638	0.642	0.511	0.436	0.558	0.535	0.436	0.387
DenseNet	1	0.486	0.472	0.274	0.026	0.553	0.552	0.553	0.552	0.347	0.174	0.276	0.019
	2	0.499	0.452	0.274	0.026	0.521	0.514	0.448	0.426	0.274	0.026	0.274	0.026
	4	0.448	0.426	0.486	0.472	0.274	0.026	0.448	0.426	0.274	0.026	0.281	0.027
	8	0.486	0.472	0.350	0.168	0.418	0.210	0.553	0.552	0.494	0.267	0.274	0.026
RPET_NET	1	0.510	0.451	0.446	0.373	0.488	0.402	0.631	0.623	0.613	0.553	0.515	0.412
	2	0.523	0.473	0.503	0.449	0.561	0.496	0.481	0.433	0.436	0.362	0.529	0.464
	4	0.574	0.549	0.423	0.352	0.518	0.483	0.545	0.492	0.338	0.106	0.561	0.518
	8	0.456	0.403	0.451	0.389	0.492	0.451	0.548	0.511	0.484	0.421	0.398	0.300

Table 64. Performance metrics for CNNs architectures for pGT labelling and PET input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.

		pRGB - pGT											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.465	0.386	0.506	0.449	0.513	0.443	0.586	0.553	0.474	0.337	0.274	0.026
	2	0.504	0.447	0.578	0.508	0.570	0.483	0.566	0.551	0.274	0.026	0.395	0.174
	4	0.616	0.600	0.585	0.503	0.555	0.519	0.578	0.549	0.308	0.072	0.274	0.026
	8	0.543	0.508	0.528	0.483	0.636	0.573	0.405	0.375	0.348	0.157	0.381	0.155
VGG16	1	0.576	0.541	0.436	0.305	0.274	0.026	0.553	0.552	0.274	0.026	0.274	0.026
	2	0.477	0.398	0.274	0.026	0.597	0.553	0.414	0.370	0.498	0.392	0.595	0.515
	4	0.462	0.379	0.461	0.400	0.490	0.429	0.405	0.375	0.406	0.181	0.274	0.026
	8	0.405	0.375	0.274	0.026	0.274	0.026	0.405	0.375	0.274	0.026	0.274	0.026
ResNet	1	0.521	0.514	0.556	0.519	0.440	0.335	0.405	0.375	0.440	0.226	0.274	0.026
	2	0.405	0.375	0.274	0.026	0.405	0.375	0.405	0.375	0.421	0.381	0.274	0.026
	4	0.588	0.567	0.448	0.426	0.347	0.174	0.405	0.375	0.405	0.375	0.467	0.230
	8	0.581	0.558	0.476	0.456	0.509	0.496	0.448	0.426	0.421	0.381	0.499	0.478
DenseNet	1	0.405	0.375	0.347	0.174	0.274	0.026	0.405	0.375	0.274	0.026	0.274	0.026
	2	0.405	0.375	0.405	0.375	0.405	0.375	0.405	0.375	0.274	0.026	0.274	0.026
	4	0.405	0.375	0.274	0.026	0.347	0.174	0.405	0.375	0.274	0.026	0.274	0.026
	8	0.405	0.375	0.274	0.026	0.405	0.375	0.405	0.375	0.274	0.026	0.274	0.026
RPET_NET	1	0.598	0.550	0.476	0.414	0.528	0.446	0.519	0.421	0.340	0.164	0.451	0.361
	2	0.546	0.465	0.512	0.427	0.516	0.446	0.644	0.614	0.561	0.524	0.428	0.325
	4	0.484	0.412	0.529	0.472	0.498	0.444	0.451	0.385	0.519	0.434	0.514	0.453
	8	0.486	0.420	0.542	0.471	0.515	0.432	0.405	0.375	0.557	0.522	0.514	0.331

Table 65. Performance metrics for CNNs architectures for pGT labelling and pRGB input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_{macro}; F1_w: F1_{weighted}.

		MIP - pGT											
		Without DA						With DA					
		CW = A		CW = B		CW = C		CW = A		CW = B		CW = C	
		F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w	F1_m	F1_w
AlexNet	1	0.577	0.545	0.556	0.529	0.503	0.447	0.529	0.508	0.424	0.196	0.423	0.229
	2	0.645	0.612	0.477	0.421	0.492	0.436	0.534	0.496	0.394	0.166	0.506	0.383
	4	0.511	0.429	0.574	0.544	0.546	0.507	0.549	0.503	0.438	0.280	0.488	0.355
	8	0.534	0.462	0.482	0.416	0.463	0.405	0.588	0.585	0.380	0.161	0.405	0.177

VGG16	1	0.405	0.375	0.347	0.174	0.352	0.095	0.405	0.375	0.274	0.026	0.413	0.304
	2	0.405	0.375	0.274	0.026	0.441	0.212	0.521	0.514	0.274	0.026	0.274	0.026
	4	0.416	0.310	0.503	0.441	0.453	0.380	0.494	0.490	0.274	0.026	0.274	0.026
	8	0.405	0.375	0.427	0.337	0.486	0.412	0.488	0.407	0.274	0.026	0.274	0.026
ResNet	1	0.405	0.375	0.545	0.528	0.448	0.426	0.405	0.375	0.421	0.381	0.274	0.026
	2	0.657	0.576	0.405	0.375	0.448	0.426	0.405	0.375	0.301	0.064	0.553	0.552
	4	0.546	0.513	0.611	0.583	0.438	0.374	0.553	0.552	0.486	0.472	0.514	0.480
	8	0.486	0.472	0.475	0.402	0.495	0.434	0.521	0.514	0.521	0.514	0.462	0.432
DenseNet	1	0.405	0.375	0.274	0.026	0.405	0.375	0.405	0.375	0.274	0.026	0.405	0.375
	2	0.405	0.375	0.405	0.375	0.274	0.026	0.405	0.375	0.405	0.375	0.274	0.026
	4	0.405	0.375	0.405	0.375	0.405	0.375	0.405	0.375	0.405	0.375	0.274	0.026
	8	0.405	0.375	0.405	0.375	0.405	0.375	0.405	0.375	0.274	0.026	0.405	0.375
RPET_NET	1	0.512	0.473	0.518	0.439	0.428	0.347	0.531	0.456	0.495	0.394	0.557	0.523
	2	0.479	0.388	0.601	0.502	0.582	0.500	0.554	0.479	0.550	0.507	0.511	0.414
	4	0.617	0.576	0.495	0.462	0.563	0.532	0.518	0.477	0.510	0.443	0.586	0.503
	8	0.629	0.587	0.509	0.470	0.573	0.515	0.570	0.558	0.588	0.537	0.594	0.545

Table 66. Performance metrics for CNNs architectures for pGT labelling and MIP input for BM status classification in NHL patients. The maximum value for each group and labelling set is surrounded by continuous black borders and bold type. DA: data augmentation; CW: class weights; F: filters or channels for the first feature map, F1_m: F1_macro; F1_w: F1_weighted.