






Article

Machine Learning Models Based on [¹⁸F]FDG PET Radiomics for Bone Marrow Assessment in Non-Hodgkin Lymphoma

Eva Milara ^{1,2} , Pilar Sarandeses ^{3,4}, Ana Jiménez-Ubieto ⁵, Adriana Saviatto ³, Alexander P. Seiffert ^{1,2} , F. J. Gárate ^{1,2} , D. Moreno-Blanco ^{1,2} , M. Poza ⁵, Enrique J. Gómez ^{1,2,6}, Adolfo Gómez-Grande ^{3,4} and Patricia Sánchez-González ^{1,2,6,*} 

- ¹ Biomedical Engineering and Telemedicine Centre, ETSI Telecomunicación, Center for Biomedical Technology, Universidad Politécnica de Madrid, 28040 Madrid, Spain; eva.milara.hernando@upm.es (E.M.); ap.seiffert@upm.es (A.P.S.); fj.garate@upm.es (F.J.G.); diego.morenob@upm.es (D.M.-B.); enriquejavier.gomez@upm.es (E.J.G.)
- ² Instituto de Investigación Hospital 12 de Octubre (imas12), Hospital Universitario 12 de Octubre, 28041 Madrid, Spain
- ³ Department of Nuclear Medicine, Hospital Universitario 12 de Octubre, 28041 Madrid, Spain; mariadel Pilar.sarandeses@salud.madrid.org (P.S.); adriana.saviatto.nardi@salud.madrid.org (A.S.); adolfo.gomez@salud.madrid.org (A.G.-G.)
- ⁴ Facultad de Medicina, Universidad Complutense de Madrid, 28040 Madrid, Spain
- ⁵ Department of Hematology, Hospital Universitario 12 de Octubre, 28041 Madrid, Spain; ana.jimenezub@salud.madrid.org (A.J.-U.); mpoza@salud.madrid.org (M.P.)
- ⁶ Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, 28029 Madrid, Spain
- * Correspondence: p.sanchez@upm.es



Citation: Milara, E.; Sarandeses, P.; Jiménez-Ubieto, A.; Saviatto, A.; Seiffert, A.P.; Gárate, F.J.; Moreno-Blanco, D.; Poza, M.; Gómez, E.J.; Gómez-Grande, A.; et al. Machine Learning Models Based on [¹⁸F]FDG PET Radiomics for Bone Marrow Assessment in Non-Hodgkin Lymphoma. *Appl. Sci.* **2024**, *14*, 10291. <https://doi.org/10.3390/app142210291>

Academic Editor: Anca Udristoiu

Received: 3 October 2024

Revised: 6 November 2024

Accepted: 6 November 2024

Published: 8 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Non-Hodgkin lymphoma is a heterogeneous group of cancers that triggers bone marrow infiltration in 20–40% of cases. Bone marrow biopsy in combination with a visual assessment of [¹⁸F]FDG PET/CT images is used to assess the marrow status. Despite the potential of both techniques, they still have limitations due to the subjectivity of visual assessment. The present study aims to develop models based on bone marrow uptake in [¹⁸F]FDG PET/CT images at the time of diagnosis to differentiate bone marrow status. For this purpose, a model trained for skeleton segmentation and based on the U-Net architecture is retrained for bone marrow segmentation from CT images. The mask obtained from this segmentation together with the [¹⁸F]FDG PET image is used to extract radiomics features with which 11 machine learning models for marrow status differentiation are trained. The segmentation model yields very satisfactory results with Jaccard and Dice index values of 0.933 and 0.964, respectively. As for the classification models, a maximum F1_score_weighted and F1_score_macro of 0.962 and 0.747, respectively, are achieved. This highlights the potential of these features for bone marrow assessment, laying the foundation for a new clinical decision support system.

Keywords: machine learning; radiomics; bone marrow involvement; Non-Hodgkin Lymphoma; segmentation

1. Introduction

Non-Hodgkin Lymphoma (NHL) is a lymphoid neoplasm characterized by the inclusion of a heterogeneous group of cancers [1], with Follicular Lymphoma (FL) and Diffuse Large B-cell Lymphoma (DLBCL) being the most common indolent and aggressive subtypes in developed countries [2]. Although NHL usually arises in the lymph nodes, in 20–40% of cases, this pathology triggers a bone marrow infiltration (BMI). This infiltration has many therapeutic and prognostic implications, making an accurate assessment of BMI crucial in the staging of NHL [1–3].

Bone marrow aspiration or trephine biopsy are the first-line methods for BMI assessment. However, the BMI pattern is frequently patchy and heterogeneous, making the small sample taken unrepresentative of the entire bone marrow. Therefore, in addition to the discomfort and potential complications for the patient, there is increasing interest in using [^{18}F]FDG PET/CT for BMI assessment, at least as a complementary test [3–6]. The primary reason for employing [^{18}F]FDG PET/CT imaging is that the image can not only be evaluated qualitatively by nuclear medicine experts but also quantitatively with the maximum standardized uptake value (SUV_{max}) [7–9].

However, there is a discrepancy between the biopsy and [^{18}F]FDG PET/CT BMI evaluation. As outlined in the reviews developed by Pakos et al. (2005) [8], Chen et al. (2011) [10], and Almaimani et al. (2022) [11], the potential of the [^{18}F]FDG PET/CT assessment in the detection of biopsy results depends on the NHL subtype, being notably higher for aggressive NHL subtypes compared to indolent subtypes. Specifically, Almaimani et al. evaluated the median sensitivity and specificity of [^{18}F]FDG PET/CT assessment for the articles developed in four different NHL groups: aggressive, indolent, DLBCL, and FL. They obtained sensitivities of 0.77, 0.585, 0.774, and 0.6 and specificity values of 0.938, 0.851, 0.916, and 0.81, respectively. All subtypes exhibit satisfactory specificities. Nonetheless, those related to indolent types are lower than those related to aggressive subtypes. Furthermore, sensitivities are only acceptable in those that are aggressive. Despite the inclusion of studies that quantitatively analyze [^{18}F]FDG PET/CT images, the methodology is limited to SUV metrics and the ratios between them. None of the studies utilized radiomics for the quantification of [^{18}F]FDG PET/CT, although it is considered a useful tool in the evaluation of bone marrow in other clinical contexts with bone marrow alterations [12–15] or in other studies where the primary disease is lymphoma [16,17].

The studies included in the three previous reviews evaluate the differentiation between positive BMI and negative 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 bone marrow is the main hematopoietic organ [18–21]. No studies have been identified that differentiated this inflammatory status from those previously mentioned.

For these reasons, quantification methodologies beyond SUV metrics are needed to assess the potential of PET in distinguishing positive BMI from negative BMI to support clinical decision making in a less subjective way. Furthermore, these methodologies not only should include the distinction between these two bone marrow states but also the differentiation between inflammatory and infectious states.

The aim of this study is to develop classification models to distinguish three bone marrow states: non-infiltrated (BMI[−]), infiltrated (BMI⁺), and inflammatory or reactive (rBM). To this end, different machine learning models based on bone marrow radiomics features are proposed. In order to obtain the features exclusively from the bone marrow region, a U-Net for segmenting bone marrow from CT images is trained and applied to the [^{18}F]FDG PET images for extracting the metabolic information.

2. Materials and Methods

In this study, the methodology shown in Figure 1 was applied. This methodology consists of two different steps, each of them with its own cohort. The first step is bone marrow segmentation, which finishes with the obtention of the segmentation model. The second step is the quantitative bone marrow assessment, which finishes with the evaluation of the machine learning models for evaluating the bone marrow status in NHL patients.

2.1. Bone Marrow Segmentation

2.1.1. Subjects

The study cohort is composed of the same patients as the cohort of [22]. Consequently, a total of 77 CT images from two different acquisition protocols are included: 45 cases of the multiple myeloma diagnostic protocol (whole-body) and 32 cases of the NHL diagnostic

protocol (femur-to-head). Siemens Biograph TruePoint 6 PET/CT (Siemens Healthineers, Erlangen, Germany) was used to obtain whole-body CT scans as part of the [^{18}F]FDG PET/CT scans at the Department of Nuclear Medicine of the Hospital Universitario 12 de Octubre in Madrid, Spain. These studies were acquired according to the European Association of Nuclear Medicine (EANM) procedure guidelines [23]. CT images were obtained using helical CTs (120–140 kVp, 25–170 mAs) with a resolution of 512×512 and a voxel size of $0.9766 \times 0.9766 \times 2.5 \text{ mm}^3$. The number of slices of these CT scans varied from 254 to 724.

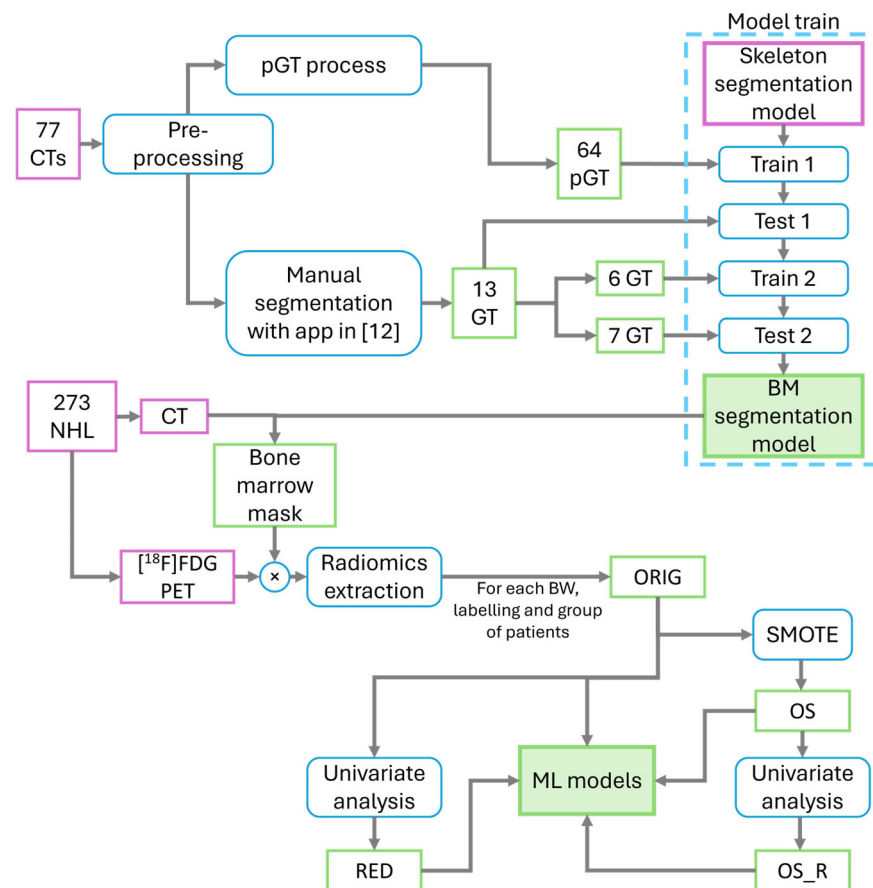


Figure 1. Pipeline of the followed methodology. pGT: pseudo ground truth; GT: ground truth; BM: bone marrow, BW: bin width; ORIG: original; RED: reduced; OS: oversampled; OS_R: oversampled and reduced; ML: machine learning.

2.1.2. Pre-Processing

As in [22], pre-processing consisted of four steps: (I) stretcher removal, (II) thresholding, (III) image clipping, and (IV) intensity normalization. Based on the original CT images and the stretcher removal process described in [22], the U-Net architecture is trained to obtain the stretcher masks. The mask defined as the ground truth for this model is the result of the thresholding of the stretcher mask obtained by the procedure described in [22]. Similarly to that study, different numbers of channels in each feature map ($F = 2, 4,$ and 8) and different numbers of epochs ($E = 5, 10,$ and 20) are tested. The model with the best performance is obtained for $F = 8$ and $E = 5$, with a Jaccard Index of 0.989 and Dice of 0.995.

Once the stretcher is identified and removed, to focus the segmentation on the bone marrow, a threshold between -100 HU and 300 HU is applied to the CT images. Then, following the methodologies applied in previous studies [12,22], all images are cropped from the first visible femur slice to the first slice of the head. Finally, the images are normalized between 0 and 255.

2.1.3. Datasets

Once the pre-processing is applied to the 77 CT images to generate the inputs of the segmentation model, the masks are obtained differently for the training (64 patients) and testing (13 patients) sets. Because the process of manual bone marrow segmentation requires a large amount of time—between 6 and 12 h per patient (approximately 300 slices)—a pseudo ground truth is generated from the manual skeleton segmentations used in [22]. This process, applied over the 21,110 slices of the 64 training patients, includes four steps: (I) dilation of the skeleton mask with a 3-pixel radius followed by subtracting the original mask from this dilated mask to leave only the edges; (II) dilation of the edges of the mask with a radius of 2 pixels to remove the compact bone; (III) removal of all the pixels with a grey value equal to the maximum after normalization (which corresponds to all the pixels over 300 HU in the original CT scan); and (IV) the removal of structures with less than 10 pixels to smooth the resulting mask. Figure 2 shows the different masks. In order to make the test as accurate as possible, 4007 slices from the 13 ground truth masks of the test set are manually segmented.

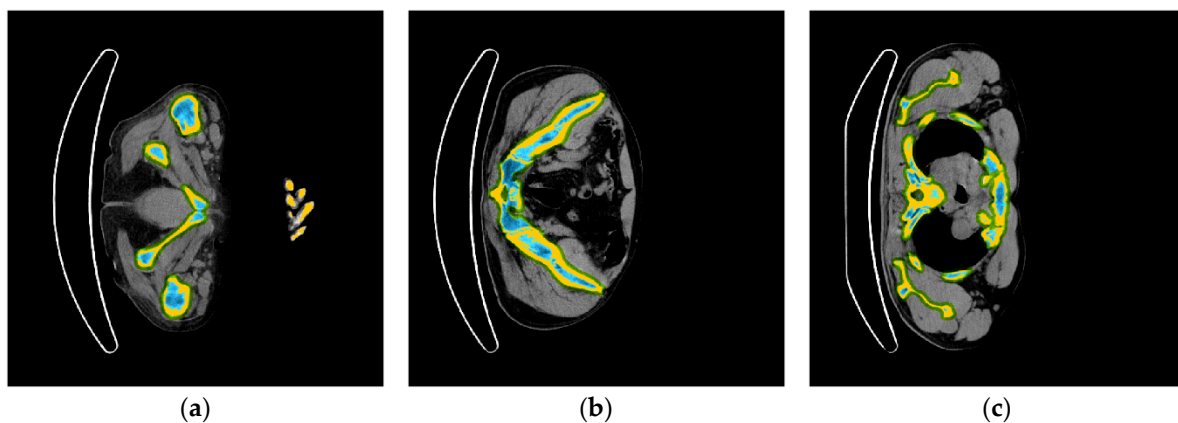


Figure 2. Examples of the pseudo ground truth created for bone marrow segmentation in (a) inferior and (b) superior pelvic area and (c) ribs and shoulder area; (Green) edges removed as a result of step (II); (yellow) pixels over 300 HU in the original CT removed in step (III); (blue) resulting bone marrow mask after applying the four steps.

2.1.4. Model Architecture and Assessment

The 2D model based on the U-Net architecture described in [24] and implemented in [22], considering the model with the best performance, is retrained for bone marrow segmentation. Therefore, thanks to the possibility of transfer learning, a U-Net model already trained for a similar segmentation is used. As shown in Figure 1, the retraining is divided into two different steps. Firstly, the model is retrained with pseudo bone marrow masks and tested with manually segmented bone marrow masks. The input order of the different subsets of bone marrow masks is again randomized in twenty different orders. Secondly, in order to make the training more accurate, the resulting model is retrained again with 6 patients of the 13 manually segmented (1749 slices) and tested with the 7 resting patients (2258 slices). The performance assessment is evaluated using the Jaccard (Intersection over Union, IoU) Index and the Dice Index, as in [22].

Python 3.9.12 software (Python Software Foundation, Wilmington, DE, USA, EEUU) is used for the implementation, training, and testing of the model. The complete development is carried out using an NVIDIA GeForce GTX 1660 GPU with 6 GB of VRAM (NVIDIA Corp., Santa Clara, CA, USA).

2.2. Quantitative Bone Marrow Assessment

2.2.1. Subjects

The quantification cohort is composed of a total of 273 patients diagnosed with NHL at Hospital Universitario 12 de Octubre, Madrid, Spain. Only patients with [¹⁸F]FDG

PET/CT acquired at the Department of Nuclear Medicine according to the European Association of Nuclear Medicine (EANM) procedure guidelines [23] and their respective visual evaluations conducted by Nuclear Medicine experts are included in the cohort. The same device used for the segmentation cohort is used to obtain whole-body PET/CT scans for the quantitative assessment. Of those patients, 234 patients also have a hematological bone marrow assessment based on Multiparametric Flow Cytometry (MFC) of the bone marrow biopsy. The group of NHL patients is divided into those diagnosed with FL (153, 56.04%), DLBCL (116, 42.49%), and transformed (4, 1.46%), i.e., patients who were first diagnosed as FL and then evolved into an aggressive DLBCL.

2.2.2. Labelling

The bone marrow assessment based on MFC and the visual assessment of the [^{18}F]FDG PET image can differ depending on the uptake pattern. A positive PET with a negative MFC often indicates a missed BMI due to focal uptake [8]. In addition, as has been mentioned before, the heterogeneous pattern observed in positive PET images is very similar to the pattern for those patients with an inflammation or infection, i.e., an rBM. For these reasons, the models are trained not only for the mentioned labels but also for pseudo ground truth (pGT) proposed as an additional label to predict. The rules applied to this pGT are defined as follows, from the most to the least important, and shown in Table 1:

1. If there is only an assessment using PET imaging, the pGT is defined as the PET label decided by experts in nuclear medicine.
2. If both MFC and PET bone marrow assessments have the same class, the pGT is defined as this coincident class.
3. If the assessment by MFC is BMI+ or rBM, the pGT is the same class as the MFC, independently of the PET assessment, due to the MFC technique's high sensitivity.
4. If the assessment by MFC is BMI− but the PET indicates a BMI+, the pGT is defined as BMI+ since the sample taken for the biopsy could be not representative due to its small size.
5. If the assessment by MFC is BMI− and the PET indicates rBM, the pGT is defined as BMI− since rBM identification by visual assessment of the PET image is considered tough.

Table 1. Confusion matrix for the pGT definition. X: no MFC test available.

		PET		
		BMI−	BMI+	rBM
MFC	BMI−	BMI−	BMI+	BMI−
	BMI+	BMI+	BMI+	BMI+
	rBM	rBM	rBM	rBM
	X	BMI−	BMI+	rBM

Moreover, considering the differences observed between indolent and aggressive subtypes of NHL, these databases are created and used for three different groups of patients: NHL, FL, and DLBCL. The transformed subtype is not taken into account due to its reduced number of cases.

2.2.3. Radiomics Features Extraction

Using the [^{18}F]FDG PET image and the bone marrow segmentation as the mask, the pyradiomics library [25] is used to extract 124 radiomic features from each patient: 18 first-order statistics; 14 based on 3D shape; 24 Gray Level Co-occurrence Matrix (GLCM) features; 16 Gray Level Run Length Matrix (GLRLM) features; 16 Gray Level Size Zone Matrix (GLSZM) features; 5 Neighboring Gray Tone Difference Matrix (NGTDM) features; 14 Gray Level Dependence Matrix (GLDM) features; and 18 features based on the mean, minimum, and maximum grey values and number of voxels from the original and interpolated

image and mask. For quantization, the fixed bin size approach is selected with the aim of maintaining image contrasts and not losing information about the heterogeneous bone marrow uptake pattern. Bin widths (BW) of 0.5 (BW0), 0.25 (BW1), 0.125 (BW2), and 0.0625 (BW3) SUV are tested, which for a maximum SUV of 37.16 in the training set population results in 75, 149, 298, and 595 levels of grey intensity, respectively.

2.2.4. Databases Creations

From each database (differing in BW extraction and labeling, i.e., PET, MFC, and pGT) formed by the radiomic features extracted from the patients included in the training set, a total of three new datasets are generated. For the training and testing split, the proportion of each class (BMI−, BMI+, and rBM) is respected, maintaining 85% of each class in the training set and 15% of each class in the testing set. The number of patients used in each set changes depending on the evaluated labeling (MFC, PET, or pGT) and disease group (NHL, FL, and DLBCL), with this always being 85% and 15% of the total number of patients for the training and testing sets, respectively.

Firstly, an oversampling of the data is performed to achieve a better representation of the minority classes by means of the SMOTE (Synthetic Minority Oversampling Technique) function [26]. From the actual cases, a set of BMI+ and rBM patients is created, increasing by 33% for each class.

Subsequently, in the original database, the univariate analysis shown in Figure 3 is performed. First of all, a normality test by means of the Shapiro–Wilk method is applied. Those features with a non-normal distribution, i.e., p -value in Shapiro–Wilk test < 0.05 , are studied by means of the Kruskal–Wallis test to evaluate significant differences between the three possible statuses (BMI−, BMI+, and rBM). If the feature results in statistically significant differences between groups with the Kruskal–Wallis test or the ANOVA test, then a post hoc test based on the Mann–Whitney U test or T-test, respectively, is applied in order to study if the statistically significant difference observed is between the three groups or only two of them. Simultaneously to the post hoc tests, a correction by means of the Benjamini–Hochberg (BH) test is applied to the p -values obtained for the Kruskal–Wallis and ANOVA tests. All features with statistically significant differences are included in the Pearson correlation analysis, while the rest are removed from the database. Using Pearson’s correlation, one variable is removed for each pair with a correlation greater than 0.9, removing the feature with the lower p -value in the difference analysis. Finally, those remaining with higher p -values in the Kruskal–Wallis and ANOVA analyses are eliminated until the number of variables is equal to 10% of the number of cases (1 characteristic for every 10 subjects). A p -value of < 0.05 is considered statistically significant in all analyses.

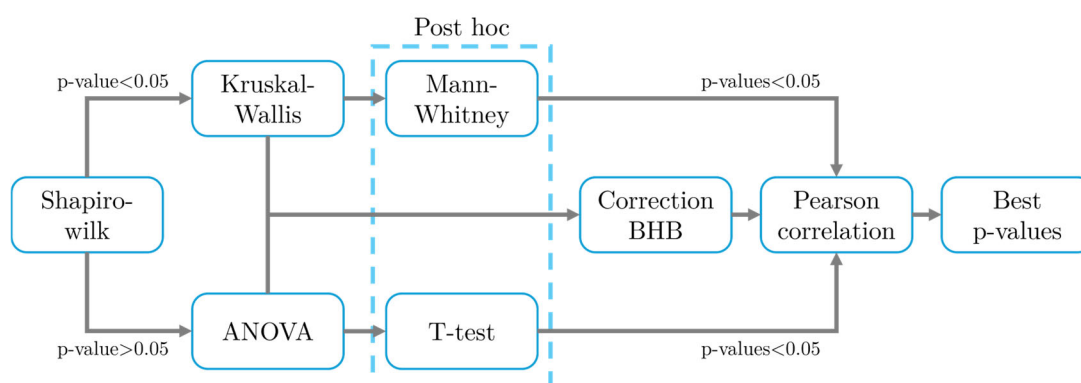


Figure 3. Univariate analysis methodology.

This way, four different databases are created (Figure 1) for each pair of BW and labeling: (ORIG) original; (RED) reduced through the presented methodology (see Figure 3); (OS) oversampled by the SMOTE function; and (OS_R) oversampled database reduced

with the features resulting from the univariate analysis applied to the ORIG database. This way, a total of 48 databases are created.

2.2.5. Machine Learning Models

These databases are employed for the training of 11 distinct machine learning models derived from scikit-learn [27], namely a decision tree (DT), Support Vector Machines (SVMs) with three different kernels called radial basis function (RBF), linear (LIN), and sigmoidal (SIG), a random forest (RF), gradient boosting (GB), logistic regression (LR), Bernoulli naïve bayes (NB), k-nearest neighbor (KNN), and multilayer perceptron with 100 and with 200 neurons in the hidden layer (MLP1 and MLP2). Hyperparameter optimization is applied for each model type with the F1_macro score (mean of F1_score for each group) used to determine the optimal values. Furthermore, with the exception of the DT, SVM, and NB models, a data normalization process is applied to each variable, ensuring a mean of 0 and a standard deviation of 1. In the cases of DT and SVM, no improvement to the models is observed, so it is not applied. In the case of NB, according to multivariate Bernoulli distributions, binary features are necessary to train the model. The optimal cut-off values for features defining high- and low-value sub-groups are obtained through the application of the receiver operating characteristic (ROC) and the closest-to-(0,1) criterion. In order to evaluate the models, the metric employed for its optimization, F1_score_macro, is also used for the model evaluation, along with an F1_macro* calculated as the mean of the F1_score_macro of each label interpreted as a binary problem and the F1_score_weighted. This represents the metrics from less to more important regarding the class imbalance.

3. Results

3.1. Bone Marrow Segmentation

Since the first part of the training is based on a model previously developed in [22] and pseudo bone marrow masks, only the parameter order can be evaluated. From the twenty orders, the best of them obtains a performance characterized by a Dice Index of 0.94 and IoU of 0.892. After training this model with the six manually segmented cases, i.e., a real ground truth, the performance improves in values of these same indexes of 0.964 and 0.933. As can be seen in Figure 4, the vast majority of segmented pixels are correctly assigned as bone marrow masks (green). Concretely, in Figure 5, the slices from Figure 4 are magnified to observe the mistaken pixels. Only some regions are complete or partially misleading, while many pixels belong to misunderstood edges.

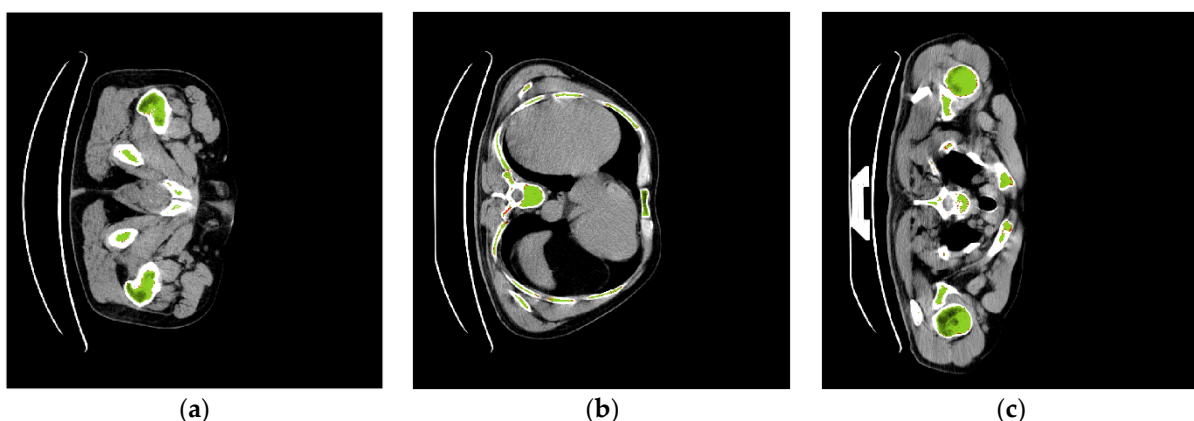


Figure 4. Examples of the results for bone marrow segmentation for different anatomical areas: (a) pelvis, (b) ribs, and (c) shoulders and collarbones; (Green) pixels correctly assigned as bone marrow mask; (red) pixels wrongly assigned as background; (orange) pixels wrongly assigned as bone marrow mask.

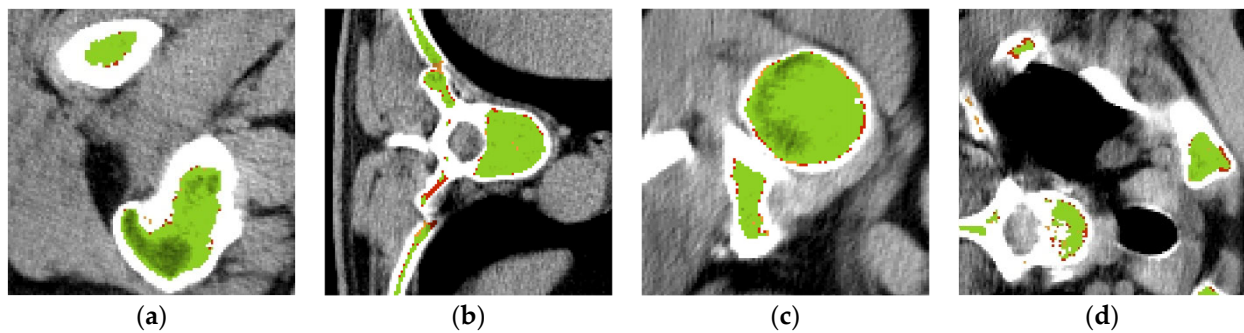


Figure 5. Magnification of the results for bone marrow segmentation shown in Figure 4: (a) end of femur and beginning of iliac crest in the right side, (b) vertebral column, (c) left shoulder, and (d) left collarbone and vertebral column; (Green) pixels correctly assigned as bone marrow mask; (red) pixels wrongly assigned as background; (orange) pixels wrongly assigned as bone marrow mask.

3.2. Quantitative Bone Marrow Assessment

3.2.1. Subjects

As can be seen in Table 2, 37.6% (88 out of 234), 40.29% (56 out of 139), and 31.86% (29 out of 91) of NHL, FL, and DLBCL patients with both MFC and PET assessment belong to different classes. Following the rules described above, the pGT labeling 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.

Table 2. Confusion matrix comparing PET and MFC labeling. X: no MFC test available.

Class		PET Assessment								
		NHL			FL			DLBCL		
		BMI [−]	BMI ⁺	rBM	BMI [−]	BMI ⁺	rBM	BMI [−]	BMI ⁺	rBM
MFC assessment	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

Taking MFC and pGT labeling as the ground truth, the F1_score for the PET visual assessment is summarized in Table 3. As can be seen, taking MFC as the ground truth, no acceptable values (greater than 0.7) are obtained. As pGT is obtained partially from PET labeling, these results are better than those obtained with MFC. However, the F1_score for the rBM class is only acceptable for the BMI[−] class and the BMI⁺ class in the FL and DLBCL groups.

Table 3. (F1_0) F1_score for BMI[−] class; (F1_1) F1_score for BMI⁺ class; (F1_2) 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 used PET labeling as predicted and MFC or pGT labeling as ground truth.

	MFC			pGT		
	NHL	FL	DLBCL	NHL	FL	DLBCL
F1_0	0.741	0.806	0.797	0.783	0.763	0.722
F1_1	0.521	0.463	0.480	0.696	0.705	0.815
F1_2	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

3.2.2. Machine Learning Models

Tables 4 and 5 show the F1_score metrics of the machine learning models with the best F1_macro * and F1_weighted, respectively, for the PET, MFC, and pGT labeling for NHL, FL, and DLBCL patients. The lack of significant features should be noted for the NHL and DLBCL groups in MFC labeling. Both the FL group in MFC labeling and the DLBCL group in pGT labeling have significant features only for BW3, BW2, and BW3, respectively.

As can be seen in Table 5, the best F1_macro * results are obtained for the PET labeling with maximums of 0.822, 0.836, and 0.759 for NHL, FL, and DLBCL, respectively, since the features studied are directly extracted from the [¹⁸F]FDG PET image without any clinical aspect. Similarly, pGT labeling obtains better performances of 0.685, 0.739, and 0.744 than MFC with 0.659, 0.619, and 0.714 since it partially takes into account the PET labeling.

Table 4. Metrics for the models with the best performance (maximum F1_m *) for each group (ORIG, OS, RED, or OS_R), group of patients (NHL, FL, and DLBCL), and labeling (PET, MFC, and pGT). (F1_0) F1_score for BMI− class; (F1_1) F1_score for BMI+ class; (F1_2) 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. Grey cells represent the lack of those groups, i.e., no significant features after reduction. Green cells represent crossover between the model with the best F1_m* and the best F1_w. The measures marked in bold are the best of its metric for the pair labeling and group of patients; those marked in italics are the best of its labeling; and the underline indicates the best for its group of patients.

F1_macro *		NHL				FL				DLBCL			
		ORIG	RED	OS	OS_R	ORIG	RED	OS	OS_R	ORIG	RED	OS	OS_R
PET	F1_0	0.900	0.852	<u>0.918</u>	0.873	0.903	<u>1.000</u>	0.815	0.897	0.897	0.897	0.857	<u>0.929</u>
	F1_1	0.778	<u>0.842</u>	0.824	0.700	0.727	<u>0.923</u>	0.667	0.727	0.500	0.500	0.500	0.500
	F1_2	0.500	0.222	0.500	0.571	0.000	0.000	<u>1.000</u>	0.500	<u>0.667</u>	<u>0.667</u>	0.500	0.500
	F1_m	0.726	0.639	<u>0.747</u>	0.715	0.543	0.641	<u>0.827</u>	0.708	<u>0.688</u>	<u>0.688</u>	0.619	0.643
	F1_m *	0.803	0.748	<u>0.822</u>	0.792	0.710	0.812	<u>0.836</u>	0.797	0.757	0.757	0.706	<u>0.759</u>
	F1_w	0.818	0.792	<u>0.843</u>	0.803	0.815	<u>0.962</u>	0.768	0.835	0.741	0.741	0.690	<u>0.807</u>
MFC	F1_0	0.727		0.744		0.500	0.421	0.727	0.455	0.750		0.842	
	F1_1	0.444		0.167		0.421	0.471	0.625	0.353	<u>0.667</u>		<u>0.667</u>	
	F1_2	0.500		<u>0.667</u>		<u>0.667</u>	0.333	0.000	<u>0.667</u>	0.444		0.333	
	F1_m	0.557		0.526		0.529	0.408	0.451	0.492	0.620		0.614	
	F1_m *	0.659		0.652		0.605	0.546	0.620	0.569	0.705		0.715	
	F1_w	0.644		0.637		0.532	0.517	0.683	0.476	0.699		0.724	
pGT	F1_0	0.682	0.682	0.846	0.698	0.667	0.696	0.727	0.737	0.833	0.727	0.727	0.700
	F1_1	0.615	0.615	0.720	0.710	0.667	0.632	0.696	0.720	0.500	0.500	0.500	0.500
	F1_2	0.333	0.333	0.000	0.000	0.333	0.333	0.000	0.500	<u>0.667</u>	0.500	0.500	0.600
	F1_m	0.543	0.543	0.522	0.469	0.556	0.554	0.474	0.652	<u>0.667</u>	0.576	0.576	0.600
	F1_m *	0.662	0.662	0.686	0.632	0.673	0.674	0.641	0.740	<u>0.744</u>	0.668	0.668	0.690
	F1_w	0.672	0.672	0.754	0.682	0.694	0.696	0.707	0.745	0.732	0.643	0.643	0.666

Table 5. Metrics for the models with the best performance (maximum F1_w) for each group (ORIG, OS, RED, or OS_R), group of patients (NHL, FL, and DLBCL), and labeling (PET, MFC, and pGT). (F1_0) F1_score for BMI− class; (F1_1) F1_score for BMI+ class; (F1_2) 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. Grey cells represent the lack of those groups, i.e., no significant features after reduction. Green cells represent crossover between the model with the best F1_m* and the best F1_w. The measures marked in bold are the best of its metric for the pair labeling and group of patients; those marked in italics are the best of its labeling, i.e., the best of its column; and the underline indicates the best for its group of patients.

F1_weighted	NHL				FL				DLBCL				
	ORIG	RED	OS	OS_R	ORIG	RED	OS	OS_R	ORIG	RED	OS	OS_R	
PET	F1_0	0.918	0.852	0.918	0.873	0.903	1.000	0.933	0.968	0.897	0.889	0.897	0.929
	F1_1	0.800	0.842	0.824	0.700	0.727	0.923	0.667	0.833	0.500	0.500	0.400	0.500
	F1_2	0.333	0.222	0.500	0.571	0.000	0.000	0.000	0.000	0.667	0.400	0.000	0.500
	F1_m	0.684	0.639	0.747	0.715	0.543	0.641	0.533	0.600	0.688	0.596	0.432	0.643
	F1_m *	0.786	0.748	0.822	0.792	0.710	0.812	0.714	0.774	0.757	0.716	0.619	0.759
	F1_w	0.834	0.792	0.843	0.803	0.815	0.962	0.842	0.909	0.741	0.754	0.691	0.807
MFC	F1_0	0.722		0.743		0.720	0.500	0.727	0.545	0.889		0.842	
	F1_1	0.593		0.538		0.533	0.375	0.625	0.444	0.000		0.400	
	F1_2	0.000		0.000		0.000	0.333	0.000	0.000	0.500		0.500	
	F1_m	0.438		0.427		0.418	0.403	0.451	0.330	0.463		0.581	
	F1_m *	0.609		0.601		0.589	0.543	0.620	0.504	0.652		0.700	
	F1_w	0.663		0.667		0.631	0.521	0.683	0.515	0.748		0.728	
pGT	F1_0	0.723	0.698	0.846	0.698	0.727	0.750	0.727	0.737	0.833	0.667	0.783	0.700
	F1_1	0.690	0.727	0.720	0.710	0.750	0.667	0.696	0.720	0.500	0.571	0.500	0.500
	F1_2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.667	0.444	0.286	0.600
	F1_m	0.471	0.475	0.522	0.469	0.492	0.472	0.474	0.652	0.667	0.561	0.523	0.600
	F1_m *	0.632	0.639	0.686	0.632	0.659	0.640	0.641	0.740	0.744	0.666	0.644	0.690
	F1_w	0.679	0.686	0.754	0.682	0.727	0.706	0.707	0.745	0.732	0.657	0.654	0.666

Comparing different groups of patients (NHL, FL, and DLBCL), the best performance in PET labeling corresponds to the FL with an F1_score_macro* of 0.836, i.e., the indolent subgroup, followed by the NHL group with a value of 0.822 for the same metric, which includes both FL and DLBCL. However, it is remarkable that for the MFC and pGT labeling, the aggressive subtype of lymphoma, i.e., DLBCL, obtains better performance than FL or NHL with a metric of 0.715 and 0.744, respectively. Furthermore, the best F1_score for the BMI− class (F1_0) is obtained in the three groups of patients for the PET labeling, with values of 0.918, 1, and 0.929, respectively. However, in the case of the BMI+ class (F1_1) for the NHL and the FL groups and in the case of the rBM class (F1_2) for the FL and the DLBCL groups, the same behavior is observed with values of 0.842 and 0.923, and 1 and 0.667, respectively. However, for the resting groups of both classes, i.e., the BMI+ class for the DLBCL group and the rBM class for the NHL group, MFC labeling obtains better performance with values of 0.667 in both cases.

Similarly, in Table 5, the three groups of patients with the best F1_score_weighted results are obtained for PET labeling with a maximum of 0.843, 0.962, and 0.807 for NHL, FL, and DLBCL, respectively. Moreover, these models coincide with the models of maximum F1_score_macro* for the pair label and group of patients. The behavior observed for the group of patients and labeling is very similar to that observed in Table 4. However, since the F1_score_weighted metric gives more importance to class imbalance, the performance

improves in those models, obtaining good classification of the BMI− and BMI+, although no rBM cases are detected.

Table 6 shows the model and BW corresponding to the models whose performance is shown in Tables 4 and 5. Regarding the selected machine learning model, there is a notable difference between some models that result in the best performance in multiple cases, such as RF, MLP1, GB, or LR (best models in 17, 13, 11, and 11 cases, respectively), compared to others that obtain the best result in only one combination (RBF, SIG, NB, or KNN). Finally, smaller BWs more often obtain the best performance, achieving this 14, 13, 18, and 25 times for BW0, BW1, BW2, and BW3, respectively. Nevertheless, no clear tendency of improvement depending on the BW is observed.

Table 6. Machine learning model (Md) and BW for the metrics represented in Tables 4 and 5. Grey cells represent the lack of those groups, i.e., no significant features after reduction. Green cells represent crossover between the model with the best F1_m* and the best F1_w. The BW marked in bold is the only possible one for those groups and labeling due to feature reduction.

		PET				MFC				pGT			
		F1_m*		F1_w		F1_m*		F1_w		F1_m*		F1_w	
		Md	BW	Md	BW	Md	BW	Md	BW	Md	BW	Md	BW
NHL	ORIG	LR	0	LR	0	NB	3	MLP2	2	MLP1	1	GB	2
	RED	LR	2	LR	2					MLP2	3	GB	1
	OS	GB	1	GB	1	SIG	3	LR	0	LIN	0	LIN	0
	OS_R	MLP1	2	MLP1	2					GB	0	GB	0
FL	ORIG	RF, LR	2,3	RF, LR	2,3	MLP1	3	MLP2	0	DT	0	MLP2	1
	RED	RF	0	RF	0	RF	3	DT	3	LR	2	RF	1
	OS	RF	1	DT	3	MLP1	2	MLP1	2	MLP1	3	MLP1	3
	OS_R	RF	2	RF	1	RF	3	LR	3	RF	2	RF	2
DLBCL	ORIG	GB	3	GB	3	LR	1,2,3	MLP1	1	RF	0	RF	0
	RED	GB, MLP2	1,3	MLP1	3					LR, KNN	3	RBF	3
	OS	MLP1	2	GB	1	RF	0,2	DT	1	DT	2	GB	3
	OS_R	MLP1	3	MLP1	3					RF	3	RF	3

4. Discussion

Bone marrow assessment during NHL diagnosis is of great importance due to its many therapeutic and prognostic implications. In this study, bone marrow from an [¹⁸F]FDG PET/CT image dataset of patients with NHL is segmented and metabolically quantified through radiomics features.

To perform bone marrow segmentation, the U-Net network trained for skeleton segmentation in [22] is used and retrained in two steps. First, it is retrained with pseudo-real bone marrow masks and then retrained with six manually segmented bone marrow masks. The bone marrow masks obtained from this retrained model imply better results than those previously obtained in [12], with a Dice index of 0.964 and an IoU of 0.933 compared to values of 0.867 and 0.79 for the same metrics with the [12] algorithm. In addition, the results show the location of the main regions of bone marrow, failing mainly at the edges of the segmentation. As for other studies in which bone marrow segmentation is performed, examples have only been found for specific regions, such as the spinal column [28], distal tibia [29], or pelvis [30]. No studies of whole-body bone marrow segmentation have been found. Although many other architectures can be used for this segmentation, the U-Net model is selected for the possibility of transfer learning from the model in [22]. In this article, this architecture was selected since it has been proven to be efficient in biomedical image segmentation, such as the studies of Klein et al. [31] and Noguchi et al. [32], which

achieved outstanding bone segmentation. In addition, and compared to other architectures, this handles small datasets better than other complex architectures.

The quantitative bone marrow assessment methodology is also based on [12,13], taking CT bone marrow segmentation and extracting radiomics features from this region in the [¹⁸F]FDG PET to quantify the functional image to evaluate bone marrow status in NHL instead of in MM. However, the feature extraction is extended to 124 features, compared to the 32 original ones, and the process of feature reduction is more precise. Furthermore, the number of machine learning models for posterior classification is increased, and hyperparameter optimization is performed. Taking into account the studies included in the reviews by Pakos et al. [8], Chen et al. [10], and Almainani et al. [11], where the methodology was limited to SUV metrics and ratios between them and no classification methodologies, this study evaluates an innovative and objective quantification using radiomics features. In addition, no previous studies that included the rBM status have been identified.

Based on the labels obtained for the different groups of patients, i.e., MFC and PET, it can be seen in Table 3 that the F1_score values of the PET visual assessment do not reach acceptable values (greater than 0.7) in hardly any of the cases. However, in line with what was observed in the studies included in the reviews of Pakos et al. [8], Chen et al. [10], and Almainani et al. [11], in aggressive subtypes (in this case, DLBCL) BMI+ is easier to visualize without any type of quantification, as well as rBM. On the other hand, the machine learning models for PET labeling obtain better results in any group of patients than the rest of the labels since the features are extracted from the same image evaluated for the PET assessment. However, it can be seen that, in contrast to the studies without quantification, the indolent subgroup, i.e., FL, obtains better results than the aggressive one, i.e., DLBCL, with F1_score_macro * of 0.836 versus 0.759 and F1_score_weighted of 0.962 versus 0.807. 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 labeling (Table 4 with no imbalance class influence), with F1_scores for the BMI+ class of 0.667 versus 0.625, while for PET labeling, the behavior is the opposite, with F1_scores of 0.923 for the indolent type versus 0.5 for the aggressive one. The rBM class is not evaluable for its reduced number of samples in the test set. For pGT labeling, the tendency is similar to that observed in MFC labeling since the majority of patients remain equally labeled. However, due to the influence of PET labeling in those cases of BMI+ for PET and BMI− for MFC, pGT obtains a better result than MFC, observing cases with diffuse uptake but negative biopsy due to the size of the sample taken.

The limitations of this study must be separated into segmentation and quantitative assessment issues. In terms of segmentation, the ground truth used for the first bone marrow training is not manually checked, which may lead to bias. In addition, although the total number of slices for training is 21,110 and 1749 and testing is 4007 and 2258, the number of volumes they come from is 64, 13, 6, and 7, respectively. Additionally, all the images used come from the same center, which limits the applicability to any whole-body CT images from other centers or scanners since there are no other centers or scanner images available to generalize the segmentation model. Finally, only one type of network is trained, despite the fact that other networks, even variations of the same U-Net, have been proven to be as good as or even better than the U-Net for biomedical image segmentation. In terms of bone marrow assessment, the main limitation is the labeling used. Both MFC and PET are labeling approaches used in clinical practice to diagnose patients. However, a real ground truth based on follow-up of the patients would be a better and more realistic ground truth for this study, especially taking into account the discrepancies between both labels for the three groups of patients: 37.6%, 40.29%, and 31.86% for NHL, FL, and DLBCL, respectively. Furthermore, the number of patients of any subtype of NHL, i.e., FL and DLBCL, only allows us to develop proof-of-concept models since there are not enough cases to state their validity for inclusion in clinical practice. Moreover, for any group of patients, there is a disbalance between the number of cases for any class, especially for rBM since there are 20, 34, and 36 classified as rBM for PET, MFC, and pGT labeling, 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. This class imbalance, especially the low number of cases for the rBM class, prevents the models from generalizing adequately. As can be seen in Tables 4 and 5, the models encounter many difficulties in classifying this minority class, with almost all models having an F1_score value for class 2 (F1_2 in the tables) under 0.5. Although attempts to control this problem through the use of oversampling are made, there remain a very small number of example patients for the creation of new data and is insufficient to create a generalized model. Finally, as mentioned for segmentation, only images from the same acquisition scanner and center are used, limiting the generalizability of the developed models. In this case, inter-institutional validation would be necessary to confirm the potential of the machine learning models.

Finally, future works should focus on increasing the number of patients included in any subtype of NHL and even include more subtypes to evaluate the bone marrow status. This increase in the number of patients should also consider the possibility of including patients from other centers and images acquired with other scanners in order to generalize the segmentation and quantification models as much as possible. In terms of segmentation, this increase in patients can also be accompanied by experiments with other model architectures, such as U-Net variations, in order to obtain more accurate segmentation. In terms of quantification, follow-ups of the patients should be included to obtain a realistic ground truth to train the machine learning models adequately, as well as trying to obtain more patients classified as minority classes, i.e., BMI+ and rBM. Once these future works are developed and the models stop being only a proof of concept, the models of the present study could be included as part of a clinical decision support system to help nuclear medicine and hematology experts manage bone marrow status for NHL diagnosis.

5. Conclusions

In clinical routine practice, visual interpretation of [¹⁸F]FDG PET/CT images should be performed, not only for main lesion assessment but also for BMI evaluation and its differentiation from rBM status due to its importance in the diagnosis of NHL patients. In this study, a methodology for the segmentation of bone marrow and its quantification based on radiomics features is proposed. First, a fully automatic bone marrow segmentation was obtained, with remarkable results. Then, we developed machine learning models to demonstrate the possible usefulness of this quantification with a maximum F1_score of 0.843, 0.962, and 0.807 for the NHL, FL, and DLBCL subgroups, respectively. In this way, the models can be used to help clinicians in this diagnosis because the performance of the models exceeds visual assessment.

Author Contributions: All authors have contributed to the conceptualization of the work, the investigation, and review and editing of the manuscript. E.M., P.S., A.J.-U., A.S. and M.P. were responsible for the data curation. E.M., P.S., A.J.-U., F.J.G., D.M.-B. and P.S.-G. were responsible for the methodology. E.M. was responsible for the software. E.M., F.J.G. and D.M.-B. were responsible for the formal analysis. E.M., A.P.S. and P.S.-G. were responsible for the writing—original draft. E.M., P.S., A.J.-U. and P.S.-G. were responsible for validation, assuring the reproducibility of the results. E.M., A.P.S. and P.S.-G. were responsible for visualization of the published work. P.S., A.J.-U., A.S., M.P., A.G.-G., P.S. and E.J.G. were responsible for providing the resources. P.S.-G. was responsible for the supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Hospital Universitario 12 de Octubre, Madrid, Spain (protocol code 24/363 and date of approval 8 October 2024).

Informed Consent Statement: No specific informed consent is required as this is a retrospective study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to clinical patient information.

Acknowledgments: The author E.M. received financial support through a predoctoral Fellowship (*ayuda del Programa Propio de I+D+i 2020*) from Universidad Politécnica de Madrid.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Shankland, K.R.; Armitage, J.O.; Hancock, B.W. Non-Hodgkin Lymphoma. *Lancet* **2012**, *380*, 848–857. [[CrossRef](#)] [[PubMed](#)]
- Martelli, M.; Ferreri, A.J.M.; Agostinelli, C.; Di Rocco, A.; Pfreundschuh, M.; Pileri, S.A. Diffuse Large B-Cell Lymphoma. *Crit. Rev. Oncol. Hematol.* **2013**, *87*, 146–171. [[CrossRef](#)] [[PubMed](#)]
- Wu, L.-M.; Chen, F.-Y.; Jiang, X.-X.; Gu, H.-Y.; Yin, Y.; Xu, J.-R. 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.* **2012**, *81*, 303–311. [[CrossRef](#)] [[PubMed](#)]
- Çetin, G.; Çıkrıkçıoğlu, M.A.; Özkan, T.; Karatoprak, C.; Ar, M.C.; Eşkazan, A.E.; Ayer, M.; Cerit, A.; Gözübenli, K.; Börkü Uysal, B.; et al. 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? *Turk. J. Hematol.* **2015**, *32*, 213–219. [[CrossRef](#)]
- Doma, A.; Zevnik, K.; Studen, A.; Prevodnik, V.K.; Gasljevic, G.; Novakovic, B.J. Detection Performance and Prognostic Value of Initial Bone Marrow Involvement in Diffuse Large B-Cell Lymphoma: A Single Centre ¹⁸F-FDG PET/CT and Bone Marrow Biopsy Evaluation Study. *Radiol. Oncol.* **2024**, *58*, 15–22. [[CrossRef](#)]
- Alyamany, R.; El Fakih, R.; Alnughmush, A.; Albabtain, A.; Kharfan-Dabaja, M.A.; Aljurf, M. A Comprehensive Review of the Role of Bone Marrow Biopsy and PET-CT in the Evaluation of Bone Marrow Involvement in Adults Newly Diagnosed with DLBCL. *Front. Oncol.* **2024**, *14*, 1301979. [[CrossRef](#)]
- Asenbaum, U.; Nolz, R.; Karanikas, G.; Furtner, J.; Woitek, R.; Simonitsch-Klupp, I.; Raderer, M.; Mayerhoefer, M.E. Bone Marrow Involvement in Malignant Lymphoma. *Acad. Radiol.* **2018**, *25*, 453–460. [[CrossRef](#)]
- Pakos, E.E.; Fotopoulos, A.D.; Ioannidis, J.P.A. 18F-FDG PET for Evaluation of Bone Marrow Infiltration in Staging of Lymphoma: A Meta-Analysis. *J. Nucl. Med.* **2005**, *46*, 958–963.
- El-Najjar, I.; Montoto, S.; McDowell, A.; Matthews, J.; Gribben, J.; Szyszko, T.A. The Value of Semiquantitative Analysis in Identifying Diffuse Bone Marrow Involvement in Follicular Lymphoma. *Nucl. Med. Commun.* **2014**, *35*, 311–315. [[CrossRef](#)]
- Chen, Y.-K.; Yeh, C.-L.; Tsui, C.-C.; Liang, J.-A.; Chen, J.-H.; Kao, C.-H. F-18 FDG PET for Evaluation of Bone Marrow Involvement in Non-Hodgkin Lymphoma. *Clin. Nucl. Med.* **2011**, *36*, 553–559. [[CrossRef](#)]
- Almairani, J.; Tsoumpas, C.; Feltbower, R.; Polycarpou, I. FDG PET/CT versus Bone Marrow Biopsy for Diagnosis of Bone Marrow Involvement in Non-Hodgkin Lymphoma: A Systematic Review. *Appl. Sci.* **2022**, *12*, 540. [[CrossRef](#)]
- Milara, E.; Gómez-Grande, A.; Tomás-Soler, S.; Seiffert, A.P.; Alonso, R.; Gómez, E.J.; Martínez-López, J.; Sánchez-González, P. Bone Marrow Segmentation and Radiomics Analysis of [¹⁸F]FDG PET/CT Images for Measurable Residual Disease Assessment in Multiple Myeloma. *Comput. Methods Programs Biomed.* **2022**, *225*, 107083. [[CrossRef](#)] [[PubMed](#)]
- Milara, E.; Alonso, R.; Maseing, L.; Seiffert, A.P.; Gómez-Grande, A.; Gómez, E.J.; Martínez-López, J.; Sánchez-González, P. 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.* **2023**, *46*, 903–913. [[CrossRef](#)] [[PubMed](#)]
- Mesguich, C.; Hindie, E.; De Senneville, B.D.; Tlili, G.; Pinaquy, J.B.; Marit, G.; Saut, O. Improved 18-FDG PET/CT Diagnosis of Multiple Myeloma Diffuse Disease by Radiomics Analysis. *Nucl. Med. Commun.* **2021**, *42*, 1135–1143. [[CrossRef](#)]
- Jamet, B.; Morvan, L.; Nanni, C.; Michaud, A.-V.; Bailly, C.; Chauvie, S.; Moreau, P.; Touzeau, C.; Zamagni, E.; Bodet-Milin, C.; et al. 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* **2021**, *48*, 1005–1015. [[CrossRef](#)]
- Leithner, D.; Flynn, J.R.; Devlin, S.M.; Mauguen, A.; Fei, T.; Zeng, S.; Zheng, J.; Imber, B.S.; Hubbeling, H.; Mayerhoefer, M.E.; et al. Conventional and Novel [¹⁸F]FDG PET/CT Features as Predictors of CAR-T Cell Therapy Outcome in Large B-Cell Lymphoma. *J. Hematol. Oncol.* **2024**, *17*, 21. [[CrossRef](#)]
- Chen, M.; Rong, J.; Zhao, J.; Teng, Y.; Jiang, C.; Chen, J.; Xu, J. 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.* **2024**, *14*, 1394450. [[CrossRef](#)]
- Matutes, E.; Bain, B.J.; Wotherspoon, A. *Lymphoid Malignancies: An Atlas of Investigation and Diagnosis*, 1st ed.; Evidence-Based Networks Ltd.: Oxford, UK, 2007.
- Hollinger, E.F.; Alibazoglu, H.; Ali, A.; Green, A.; Lamonica, G. Hematopoietic Cytokine-Mediated FDG Uptake Simulates the Appearance of Diffuse Metastatic Disease on Whole-Body PET Imaging. *Clin. Nucl. Med.* **1998**, *23*, 93–98. [[CrossRef](#)]
- Rosenbaum, S.J.; Lind, T.; Antoch, G.; Bockisch, A. False-Positive FDG PET Uptake—The Role of PET/CT. *Eur. Radiol.* **2006**, *16*, 1054–1065. [[CrossRef](#)]
- Salaun, P.Y.; Gastinne, T.; Bodet-Milin, C.; Champion, L.; Cambefort, P.; Moreau, A.; Le Gouill, S.; Berthou, C.; Moreau, P.; Kraeber-Bodéré, F. 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* **2009**, *36*, 1813–1821. [[CrossRef](#)]

22. Milara, E.; Gómez-Grande, A.; Sarandeses, P.; Seiffert, A.P.; Gómez, E.J.; Sánchez-González, P. Automatic Skeleton Segmentation in CT Images Based on U-Net. *J. Imaging Inform. Med.* **2024**, *37*, 2390–2400. [[CrossRef](#)] [[PubMed](#)]
23. Boellaard, R.; Delgado-Bolton, R.; Oyen, W.J.G.; Giammarile, F.; Tatsch, K.; Eschner, W.; Verzijlbergen, F.J.; Barrington, S.F.; Pike, L.C.; Weber, W.A.; et al. FDG PET/CT: EANM Procedure Guidelines for Tumour Imaging: Version 2.0. *Eur. J. Nucl. Med. Mol. Imaging* **2015**, *42*, 328–354. [[CrossRef](#)] [[PubMed](#)]
24. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
25. Van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [[CrossRef](#)] [[PubMed](#)]
26. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
27. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
28. Nguyen, C.; Havlicek, J.; Duong, Q.; Vesely, S.; Gress, R.; Lindenberg, L.; Choyke, P.; Chakrabarty, J.H.; Williams, K. An Automatic 3D CT/PET Segmentation Framework for Bone Marrow Proliferation Assessment. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
29. Chen, C.; Jin, D.; Zhang, X.; Levy, S.M.; Saha, P.K. Segmentation of Trabecular Bone for In Vivo CT Imaging Using a Novel Approach of Computing Spatial Variation in Bone and Marrow Intensities. In *Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2017; pp. 3–15. ISBN 0302-9743.
30. Fiandra, C.; Rosati, S.; Arcadipane, F.; Dinapoli, N.; Fato, M.; Franco, P.; Gallio, E.; Scaffidi Gennarino, D.; Silveti, P.; Zara, S.; et al. Active Bone Marrow Segmentation Based on Computed Tomography Imaging in Anal Cancer Patients: A Machine-Learning-Based Proof of Concept. *Phys. Medica* **2023**, *113*, 102657. [[CrossRef](#)]
31. Klein, A.; Warszawski, J.; Hillengaß, J.; Maier-Hein, K.H. Automatic Bone Segmentation in Whole-Body CT Images. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 21–29. [[CrossRef](#)]
32. Noguchi, S.; Nishio, M.; Yakami, M.; Nakagomi, K.; Togashi, K. Bone Segmentation on Whole-Body CT Using Convolutional Neural Network with Novel Data Augmentation Techniques. *Comput. Biol. Med.* **2020**, *121*, 103767. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.