


Doctoral Dissertation

Doctoral Program in Bioengineering and medical-surgical sciences (33rd Cycle)

Active bone marrow segmentation based on CT imaging in anal cancer patients: a machine-learning approach

Christian Fiandra
* * * * * *

Supervisors

Prof. Gabriella Balestra, Supervisor
Prof. Pierfrancesco Franco, Co-supervisor

Politecnico di Torino
February 11, 2022

Doctoral Examination Committee:

Prof. Marco Fato, Referee, University of Genova
Dott. Nicola Dinapoli, Referee, Agostino Gemelli University Hospital Foundation IRCCS

Summary

Radiation therapy uses x-rays, gamma rays and other sources of radiation to induce cell-killing by breaking up molecules and causing reactions that damage living cells. Since radiation, usually delivered with sophisticated equipment of high energy x-rays beams, does not distinguish between cancer cells and normal tissue, radiation volumes are very carefully planned, during the process of radiation treatment, to protect uninvolved tissue and vital organs of the patient. In the radiation treatment of anal canal patients, the radiation dose received by active bone marrow comprised within the pelvic bones is a predictive factor of acute hematologic toxicity. So, the researchers focused on the implementation of strategies to selectively spare pelvic bone marrow through an accurate identification and delineation of these areas. Although the use of the whole pelvic bones, outlined as the outer contour of bony structures on computed tomography (CT), is the most inclusive method with respect to bone marrow identification, other methods have been developed to identify haematopoietically active bone marrow. A commonly used method relies on the use of ^{18}FDG-PET, based on Standardized Uptake Value (SUV), which is often employed in anal patients during diagnosis and staging work up.

In this thesis, a machine learning technique, based on radiomic features as inputs to different classifiers, was applied to CT images in order to identify haematopoietically active bone marrow, to be used as avoidance structure during radiation planning and delivery. This work has been structured in four phases: i) as first step, we tested an identification method for active bone marrow based on the combination of radiomic features extracted from pelvic CT scans with a Decision Tree on 5 patients; ii) a total of 25 patients, splitting the sample in 5 patients used for training and 20 for validation, was employed for four different classifiers (Decision Tree, K-Nearest neighbor, Neural Network and Support Vector Machine) with random selection of voxels; iii) we then compared the classification performances of classifiers between the random training set and a training set with clustering; iv) as phase iii) with a total of 50 patients, dividing the sample between 40 patients used for training and 10 for validation.

Pelvic bone marrow was delineated on planning CT and then divided into 3 subregions: iliac (IBM), lower pelvis (LPBM) and lumbosacral bone marrow (LSBM). Planning CT was then co-registered with ^{18}FDG PET-images and the active part of the three subregions was obtained and stored as ground-truth. Radiomic features were extracted from planning CT images and used to define the structure of different classifiers.

Four parameters were used for the correspondence analysis: DICE index, Precision, Recall and Mean Distance to Conformity.

In Phase i) the highest indices values are obtained for LSBM and IBM subregions, where we have an overlap above 75% in 4 out of 5 patients. For the LPBM, suboptimal results are achieved, as the Dice index is always below 0.5.

In Phase 2, the results show that the performances of the 4 classifiers are quite similar. Then the voting process helps to improve the results in terms of Dice Index especially for

IBM and LSBM with significative differences for KNN, NN and SVM, while for LPBM its effect was similar to the other approaches.
Changing the SUV threshold by ±10%, starting values of Dice Index of 0.42 for LPBM went to 0.47-0.49 for the threshold decreased by 10%, but dropped to 0.29 with the same threshold increased by 10%, therefore we felt it was not appropriate to change the reference value of SUV. So, the results, in terms of both Dice Index and MDC are comparable to those obtained in phase 1 where, on average, they were satisfactory for IBM and LSBM subregions (average values of Dice Index between 0.71 and 0.81), even if in these cases there are some patients with poor results (minimum values of Dice Index of 0.23-0.24). Conversely, inadequate results remain on the LPBM subregion (average values of Dice Index between 0.38 e 0.41) with some outliers going towards acceptable values of Dice (maximum values of Dice of 0.62-0.63).
In Phase 3, random extraction performs slightly better in most of the tested classifiers and subregions, but for LPBM and KNN classifier the median value of Dice Index increases from 0.41 with random extraction to median value of 0.61 with clustering.
In Phase 4 on the overall 50 patients, we obtain an average Dice Index of 0.64, 0.70 and 0.39 respectively for IBM, LSBM and LPBM with some patients under-performing for the regions of IBM and LSBM and conversely some patients over-performing for the region of LPBM. A strong correlation was found between the percentage of active bone marrow and the performance of all the classifiers; in particular, for patients having a percentage of active bone marrow below a threshold between 60%-70% the values of Dice Index were below the acceptability values of 0.6 also for IBM and LSBM.
An acceptable degree of overlap was found for 2 out of the 3 subregions within pelvic bones. Mean values for Dice and MDC may be considered satisfactory for both IBM and LSBM even if some cases were found to be underperforming depending on the percentage of active bone marrow relative to the total bone marrow. With respect to LPBM, the percentage of active bone marrow is significantly lower than for the other two sub-regions and the agreement is globally suboptimal; however, for some patients, having more active bone marrow, Dice Index values were close to an acceptability threshold.
While the size of the considered sample may be appropriate with respect to the adopted methodology for classifiers, some limitations in the ^{18}FDG-PET method can be highlighted including: i) the limited spatial resolution; ii) method for identifying the threshold that takes only account of the average value for the SUV, so if the distribution of the SUV would be not normal (in the patient with less active bone marrow for example), the average value could give rise to a threshold with significant subsampling of active bone marrow.

As further development the addition of another imaging technique such as MRI could help us to understand whether the ^{18}FDG-PET can really have limitations or if the problem lies in a more correct tuning of the classification process by CT.

Contents

List of Tables

List of Figures

Figure 14. Box plot showing distribution of Dice index for IBM, LPBM and LSBM structures with all classifiers

Figure 15. Box plot showing distribution of Dice index for IBM, LPBM and LSBM structures with all classifiers

Figure 16. Box plot showing distribution of Dice index for IBM, LPBM and LSBM structures with all classifiers

Figure 17. Box plot showing ratio ActBM/TotalBM for IBM, LPBM and LSBM structures.

Figure 18. Wants to visualize the major difference in favor of clustering extraction from training set data.

Figure 19. Box plot of the Dice Index for the three considered classifiers for the 40 patients used for training set

Figure 20. Box plot of the ratio Precision/Recall for the three considered classifiers for the 40 patients used for training set

Figure 21. Box plot of the Dice Index for the three considered classifiers for the 10 patients used for validation set

Figure 22. Box plot of the Dice Index for the three considered classifiers for the 10 patients used for validation set

Figure 23. Dice Index and the P/R rate reported patient by patient for the DT classifier on the left part, while the values of the values of the Active/Total BM rate for the three structures

Figure 24. Correlation between the amount of Active bone marrow with both Dice Index and ratio P/R for DT classifier for the three structures.

Chapter 1

Introduction

1.1 Radiation therapy overview

Radiation therapy (RT) is a cancer treatment method that uses strong energy beams to kill cancer cells and about 50% of cancer patients will receive radiotherapy during their course of disease. It most often uses X-rays, but protons or other particles can also be used. Cell killing is based on DNA damage and its impact on cell growth and division. Although both healthy cells and cancer cells are damaged, the goal of radiotherapy is to destroy as few normal healthy cells as possible. Normal cells can usually repair most of the damage caused by radiation. Technology has always been crucial in the development of radiotherapy, and, in the early days, technological progresses supported Radiation Therapy improvements. The road to modernity for High-tech radiation oncology continues discovering and integrating innovative ideas and technical solutions from multiple disciplines. The continuous process of integrating other disciplines, including mechanics and electronics, engineering, computer science, mathematics, imaging physics and technology, statistics and data science, it is essential for never-ending improvement. In particular, the rapid progress in computer science allowed for the development and integration of IT hardware for clinical Radiation therapy solutions and at the same time to improve software for optimizing treatment plans and delivery. The main areas where most technology-driven research is being developed include image guidance, adaptive radiation therapy, artificial intelligence integration, heavy particle therapy, and "flash" ultra-high dose rate radiation therapy.

1.2 The Impact of Artificial Intelligence on Radiation Oncology

Artificial intelligence (AI) approaches have focused the attention of many in medicine and in particular in Radiation Oncology and current writings recommend there are numerous potential advantages that could change future clinical work processes and dynamic. Compared to past transformative innovations such as the Monte Carlo method

or intensity-modulated radiotherapy, the development and adoption of AI-based tools is occurring at faster rates and promises to transform practices of the radiation treatment planning team.

Inserting AI methods in the training and qualification program of medical physicists could be an essential advance in preparing radiation specialists (RTs) to participate in equipped and safe practice as they use clinical innovations.

Given the large volume of various clinical and patient related data, these approaches of AI and its various methods like machine or deep learning, demonstrated unique opportunities to outranks most statistical based existing methods.

However, Machine Learning (ML), for example, requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality; it needs enough time to allow the algorithms to learn. Another big challenge is the ability to accurately interpret the results generated by the algorithms and it may be also highly susceptible to errors, if you train an algorithm with data sets that are not representative of the possible elements. On the other hand, since expert opinions are usually used to train algorithms, if the problem or ground truth is not clearly defined, the use of machine learning is limited. As a result, machines may excel at replicating, automating and standardizing human behavior on manual chores, meanwhile the conceptual clinical challenges relating to definition, evaluation, and judgement remain in the realm of human intelligence and insight [1].

Recently, a survey on the clinical use of AI in radiotherapy revealed that most popular AI supported applications were automatic segmentation and treatment planning, followed by synthetic CT (sCT) generation. Traditionally, the technology was mainly based on a deterministic nature in contrast with the predictable nature of artificial intelligence (AI). In addition, the mathematics education of artificial intelligence technology is absent in a standard curriculum for medical physicists, and they could therefore feel uncomfortable with this new technology. So, it emerges very clear a demand for guidance on the implementation of AI in clinical practice i.e., for commissioning, implementation and quality assurance (QA).

The commissioning procedure is two-fold, to train an AI algorithm/model and to investigate the accuracy and reproducibility of the model prior to clinical use: it can be divided into a training/test phase (first phase) and a validation phase (second phase) with some differences depending on whether the AI model has been built in house, in collaboration with a vendor or was commercially available.

Using locally acquired data offers the advantage of preserving the department's clinical guidelines and (imaging) protocols; well-known, not too large, high-quality datasets are generally preferred over very large datasets of lower quality due to the evolving nature of clinical protocols and guidelines. Validation of the model is accomplished using quantitative metrics by comparing the model output to the clinical data on a smaller set of data representative as possible the data on which the model will be applied clinically. The results of the model are then presented to clinical experts for revision and for eventually incorporating the cases on which the model and the expert disagree. If the models are built in collaboration with the vendor or are commercially available do typically not allow or require this phase but should then be accompanied by proof of FDA/CE marking and a vendor's validation report detailing the performance of the model. The goal of the test phase is to obtain an independent evaluation of the final

performance of the model; a minimum selection of ten patients is recommended as a good starting point, which can be adjusted in case a large variation in the results is present. Optionally, other relevant endpoints such as the reduction of interobserver variation and/or timesaving could be evaluated in this phase as well.

It is important to keep in mind that while AI brings consistency, systematic errors can remain present if undetected during the implementation phase. It is strongly recommended to perform a risk analysis before any model is implemented. A well-known method is a Failure Mode and Effect Analysis (FMEA) or Risk Analysis Matrices, which includes a brainstorm on the potential risks with people from all disciplines dealing with implementation and use of the model. It is important to note that these methods are under investigation and that supervision is currently the main tool.

Radiotherapy treatment planning contains an optimization problem having many degrees of freedom. It typically requires advanced skills, is labor intensive and associated with large user variability. Developments in AI have led to applications in the field of treatment planning to decrease human intervention and workload, improve plan quality and consistency. In terms of clinical acceptability, reviewers should be able to directly compare DVHs, clinical dose goals and dose distributions via correlated scrolling in addition to dose difference maps and should score/rank the ML plans according to criteria as overall approval, target coverage, OAR sparing, high dose conformity, dose gradient, etc. Depending on the optimizer it might be necessary to perform pre-treatment QA of the predicted plans to assess deliverability. if AI is used for the prediction of the agreement between calculated and delivered dose and this agreement shows the plan is unlikely to pass QA, the dosimetrist or physicist may choose to reduce the plan complexity in the optimization process. In this way, failing plans could be potentially eliminated and a possible treatment delay can be avoided.

Synthetic CT (sCT) is often used to allow for (improved) accuracy of dose calculations on Cone-Beam CT (CBCT) or MRI images. Synthetic CTs can be generated using voxel, atlas based or hybrid approaches. Thus far, the most practical and common approach is deformable image registration to map planning CT HU to the (daily) treatment image. Recently, it has been shown that CNNs (DL) provide promising methods for synthetic CT generation based on CBCT or MRI images. To establish a paired training set, one should carefully check the (voxel-wise) alignment of patients and possibly improve this alignment by further image processing such as deformable registration. sCTs should be evaluated in terms of image similarity, geometric fidelity and dosimetric accuracy.

QA of linear accelerators is periodically performed to monitor longitudinal stability; due to developments seen in the delivery and monitoring systems, opportunities arise to complement with approaches such as Probabilistic Safety Assessment (PSA) or risk analysis to focus where AI can amplify detection levels and prediction accuracy of potential failure or deviation from intent.

1.3 Radiation therapy of anal cancer

The standard of care for patients with squamous cell carcinoma of the anal canal is concurrent chemo-radiation (RT-CHT), which is a combination of radiotherapy, usually delivered with intensity modulation, and concurrent 5-fluorouracil and Mitomycin C [2;3]. Even with highly conformal techniques, the acute toxicity associated with combined modality treatment is non-negligible and may cause unplanned treatment breaks and reduced treatment intensity, potentially jeopardizing oncological outcomes [4;5]. Hematologic toxicity (HT) is a substantial clinical issue for this subset of patients, as it increases their risk of asthenia, bleeding, or infections and may negatively affect patient's compliance to treatment [6;7]. Chemotherapy is one of the most important HT triggers as it induces myelosuppression, but radiotherapy, delivered to circulating blood cells and precursors within bone marrow (BM), also plays an important role [8]. The observation that the radiation dose received by BM comprised within the pelvic bones is a predictive factor for HT occurrence prompted researchers to implement strategies to selectively spare pelvic BM through the accurate identification and delineation of these areas [9]. This may decrease the probability of developing HT, since in the average adult population, pelvic bones and lumbar vertebrae make up almost 60% of the total BM. Although the use of the whole pelvic bones, outlined as the outer contour of bony structures on computed tomography (CT), is the most inclusive method with respect to BM identification, other methods have been developed to identify haematopoietically active BM (actBM). Even if a widely accepted strategy for the contouring of active BM volumes has yet to be defined, several methods have been proposed such as: single-photon emission computed tomography (SPECT), 18F-fluorodeoxyglucose-labeled positron-emission tomography (^{18}FDG-PET) and 3'-deoxy-3'-^{18}F-fluorothymidine-labeled positron-emission tomography (^{18}FLT-PET), which all have potential advantages and disadvantages. A commonly used method for identifying actBM is ^{18}FDG-PET, which is often prescribed to anal and cervical cancer patients during diagnosis and staging work up [9;10]. Nevertheless, it is considered an optional exam by international guidelines in the diagnostic work-out of anal cancer patients and it is therefore not available for all patients in daily clinical practice. Magnetic resonance (MR) has also been used by Andreychenko A. et al. [11] who developed a tool for the semi-automatic segmentation of the red BM based on water-fat MR imaging in gynecologic cancer patients. Although interesting, this option cannot be used extensively. Conversely, all anal cancer patients undergo a CT scan before receiving radiotherapy. Consequently, an approach based on CT images aimed at defining actBM would prove useful and could be used on a broader scale.

1.4 Automatic segmentation in the radiation treatment planning process

Image segmentation is an important task performed routinely in a RT clinical workflow with the radiation oncologist that will manually segment the Regions of interest

(ROIs) on RT simulation scan. The manual segmentation of these ROIs is a time-consuming process with inconsistencies in target and organs-at-risk segmentations that have been reported having both inter- and interobserver segmentation variability [12;13]; inconsistencies lay on the inherently subjective process with decisions on which voxels to include or not based on prior knowledge and available imaging. The initial approach for automatic delineation was mainly atlas-based; it performs segmentation on a new image set using the knowledge of a prior segmentation based on deformable registration algorithms, but it still requires editing and review by an expert [14]. In particular, multiatlas segmentation has been shown to minimize the effects of intrasubject variability and improve segmentation accuracy from single atlas approaches. With many available contoured images, different shape or appearance of structures of interest could be used to train statistical shape models or statistical appearance models for auto-segmentation restricting the final segmentation results to anatomically plausible shapes [15]. To improve performance of atlas-based segmentation, hierarchical approach was used for selection of the best training cluster set that performs the best contour [16]. With many available contoured images, ML approaches can aid in segmentation by learning appropriate priors for structures and organs or image context and tissue appearance for voxel classification [17-19]. Support vector machines [20-22] and tree ensemble (i.e., random forests) [23-28] algorithms have shown promising results in thoracic, abdominal, and pelvic tumor and normal tissue segmentation. These generally employ human-engineered features, usually derived from the image intensity histograms, from a large patient database as inputs to train the segmentation model. The data used to build the final model usually comes from multiple datasets: a training dataset is a dataset of examples used during the learning process and is used to fit the parameters of the supervised learning algorithm, for example, to determine, or learn, the optimal combinations of variables that will generate a good predictive model.; however it is well-understood that the training set performance tends to overestimate the validation set accuracy, so it is advantageous to use a test set to evaluate the trained model's performance during hyper-parameter search and model optimization. One of the most popular methods is "k-fold cross validation" or "leave-x-out cross validation" where the training set is divided in k parts and each of the n original cases has been left out exactly once (x=n/k). The resulting algorithm will then be tailored specifically to the training data. There are many possible ways to split the original dataset and this task could be dependent on data availability. A split providing 20% validation and 30% test from the remaining 80% has been shown to produce good generalization from the test to validation set accuracy [29]. Deep learning (DL) is part of the field of machine learning where algorithms can learn data representation on their own.

1.5 Objective and Challenges

Organ at risks (OARs) and target volume segmentation is an essential point for radiation treatment planning; however manual segmentation is a time-consuming task with high intra and interobserver variability both within [30] and across [31] radiotherapy

centers. The initial approach for automatic delineation was mainly atlas based; it performs segmentation on a new image set using the knowledge of a prior segmentation based on deformable registration algorithms, but it still requires editing and review by an expert [14]. Later, when more contour images are available, machine learning methods can help segmentation by learning appropriate prior structure and organ or image context and tissue appearance for voxel classification. Support vector machines and tree ensemble (ie random forest) algorithms have shown encouraging results in chest, abdomen and pelvic tumors and normal tissue segmentation. These generally employ human-engineered features, usually derived from the image intensity histograms, from a large patient database as inputs to train the segmentation model.

This work has been structured in four phases:

i) As first step, we investigated an identification method for active bone marrow (actBM) combining radiomic features extracted from pelvic CT scans with a Decision Tree classifier proving the feasibility of such approach on 5 patients.

ii) a total of 25 patients dividing the sample between 5 patients used for training and 20 for validation; the selection of voxels to train the classifiers was basically random and the results of 4 different classifiers (Decision Tree, K-Nearest neighbor, Neural Network and Support Vector Machine) were analyzed.

iii) Then, we compared the classification performances of the architectures obtained starting from the GA solutions with the random training set compared to those obtained through the training set with clustering

iv) a total of 50 patients, dividing the sample between 40 patients used for training and 10 for validation; the selection of voxel to train the classifiers was made with clustering through unsupervised learning algorithms and the results of 3 classifiers were analyzed.

To our current knowledge, there are no other specific studies on actBM automatic identification in the literature, so the aim of this study was to improve this new method based on radiomic features extracted from pelvic CT scans on patients undergoing RT-CHT for anal cancer as inputs for Supervised learning algorithms in the field of machine learning task of learning a function that maps an input to an output based on example input-output pairs.

Chapter 2

2. Materials and methods

2.1 Machine Learning for automatic segmentation

Different Machine learning approaches are available in the context of automatic segmentation; the next few paragraphs will provide a general description of those used for this work.

2.1.1 Decision tree classifier

Decision trees (DT) is a type of supervised machine learning method in which data is continuously split according to specific parameters. Every node corresponds to a certain feature, and the arcs connecting successive nodes represent certain values of that feature. Starting from an initial node, the various nodes are covered according to the values of the variables and according to the rules of separation of each node until arriving at a final node that matches the target class. The structure is built iteratively during the training: each iteration searches for the best splitting rule, that is the rule that allows you to divide the training set into partitions that are as pure as possible, where pure partitions are intended as with elements belonging to a single class. Each best slitting rule found corresponds to a node in the structure. The construction of the tree stops when all pure partitions have been obtained and there are no more variables that can be used to derive splitting rule. At the end of the training, the end nodes are associated to the membership classes; in particular, the node class is the class with greater representation within the partition.

2.1.2 K-Nearest neighbor

The K-Nearest neighbor (KNN) is a classification algorithm based on a distance between the various elements of the training set and the element to be classified. This distance can be calculated in various ways and one of the most used is the Euclidean distance. So, when a new element must be classified, it is necessary to calculate the

distance between the elements to be classified and the training set: the elements are ordered according to the distance and the first k elements are taken, and the class to which they belong will be the one most represented. So, the parameter to choose and optimize is the value of K.

2.1.3 Neural network

Artificial neural networks are computing systems vaguely, inspired by the biological neural networks; they receive data and train themselves to recognize pattern in these data, then predicts the outputs for a new set of similar data. Used for the resolution of problems of various kinds. It is an adaptive system, whose structure varies according to the elements passed during the learning phase. The neural network used in our case is the Feedforward Neural Network (FNN). The FNN was the first and simplest type of artificial neural network wherein connections between the nodes do not form a cycle, but the information moves in only one direction—forward—from the input nodes, through the hidden nodes (if any) and to the output nodes. It is designed by a series of neurons, connected to each other by arches having different weights. Neurons are organized into layers, so neurons of a given layer are connected to neurons of next layers, but they are not connected to other neurons of the same layer. The network is organized with: i) an input layer, containing a neuron for each input variable; ii) an output layer, containing a variable number of neurons based on the number of classes; iii) a certain number of hidden layers, between input and output, for processing the incoming elements; as the number well as the number of neurons they contain may change according to the problem.

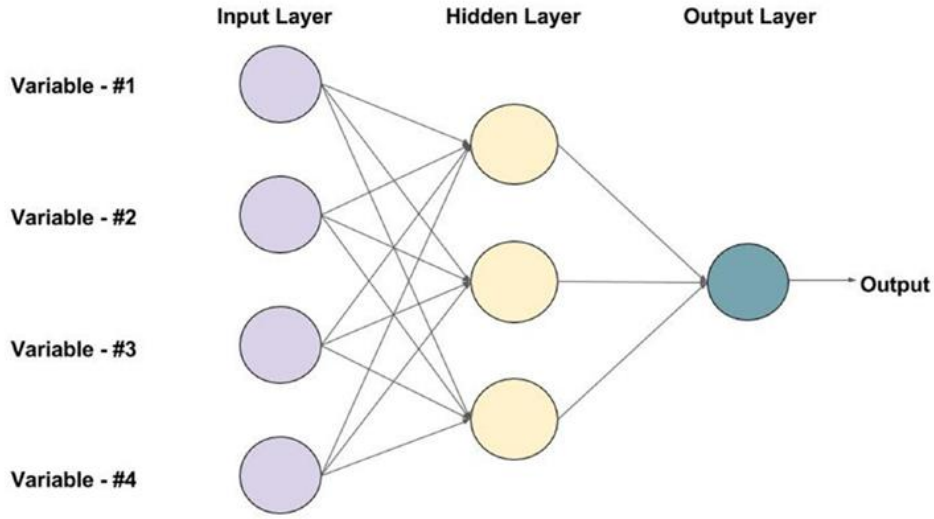


Figure 1. Structure of an FNN neural network

2.1.4 Support Vector Machine

The support-vector machine (SVM) is a machine learning method for two-group classification problems. SVM maps training examples to points in space so as to maximize the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In this space, a linear decision surface is constructed with special properties of the decision surface that ensure high generalization ability of the SVM [32]. The hyperplane is constructed in such a way as to maximize the distance between the plane itself and the support vectors (called margin), that is the elements of the two closer classes to the separation plane. Creating a new space allows you to make any problem out of classification solvable through a linear classifier. The prediction is made at this point simply by observing which side each new element falls on. In SVM, unlike other classifiers, there are two parameters on which to perform optimization; i) the kernel function that is the mathematical transfer function that allows you to map elements in the new multi-dimensional space and thus to allow their classification. Such function can be linear or non-linear, depending on the problem; ii) parameter C, adjusts the wide margin balance and a low classification error. To have wide margins, in fact, creates a very precise distinction between the elements of the training set, but increases the likelihood of misclassification of new elements. So, to have fewer wide margins can improve the overall accuracy of the model.

2.1.5 Self-Organizing maps for clustering

With the concept of clustering, the main aim it is to categorize the data into clusters where objects are grouped into a particular category; for this purpose, a self-organizing map (SOM) unsupervised artificial neural network (ANN) was trained. Self-organizing map (SOM) has been used for various applications and it is one of the most popular neural network methods for cluster analysis; its goal is to represent all input vectors in a high-dimensional space by prototypes in a low-dimensional space, such that the distance and topology are preserved as much as possible. Self-organizing maps are different from other artificial neural networks because they apply competitive learning instead of error correction learning (such as backpropagation with gradient descent), and in a sense, they use neighborhood functions to preserve the topological properties of the input space. When an input is proposed to the SOM network, the various neurons that make up the network enter “competition” between them, and the neuron that presents the most suitable characteristics is the winner to recognize that specific input. During this phase begin the learning process, i.e., there is a modification of the weights of the neuron itself which in the case of SOM networks also involves the neighborhood of the winning.

2.1.6 Genetic algorithm for feature selection and optimization of classifiers parameters

Genetic algorithm is an optimization algorithm inspired by natural selection and biology evolution; it can solve complex and non-complex optimization problems in a reasonable time and it can be solved by conventional linear search algorithm. It is based on the development of a set of initials solutions, they will randomly regroup and try again to find the best solution. Genetic operators are responsible for the recombination of solutions, and consequently evolving various generations. The mutation produces a change on one or more random bits of the single solution, complementing the value (from 0 to 1, or vice versa), in order to maintain a certain diversity of solutions and avoid that the individuals of the population become uniform. Each solution has a certain probability of change said probability of mutation (PM). The crossover instead takes place between two solutions, which are cut at one or more cut points, and the sub-strings come exchanged between them. Each solution has a certain probability of being recombined via crossover, called probability of crossover. Given the randomness of such algorithm, even starting with the same initial conditions, the final results may be different. For this purpose, the GA solutions were codified as binary vectors composed of 2 parts: the first part was used for selecting the most relevant features to be used for classification, and the second part was used to set the classifier parameters. After getting the numbers of true positives, false positives, true negatives, and false negatives, the sensitivity and specificity can be calculated.

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Figure 2. The four outcomes can be formulated in a 2×2 contingency table or confusion matrix, as well as derivations of several metrics

.

and a fitness function may be derived for the initial populations: fitness is a numerical value that is assigned to each solution through the calculation of a fitness function. This numerical value evaluates the "goodness" of the solution to the optimization problem, and according to the type of function must be minimized or maximized.

Feature selection is the process of selecting the most important and relevant features from a large number of features in a given data set. For a dataset with d input variables, the variable selection process will result in k variables such that k <d, where k is the smallest set of meaningful and relevant variables. In the Wrapper approach, the feature selection process is based on a specific machine learning algorithm that we are trying to fit a given data set. It follows a greedy search method, which evaluates all possible combinations of elements based on evaluation criteria. The evaluation criterion depends on the performance measurement regarding the type of the problem. For example, for the regression evaluation standard, it can be p-value, R-squared, and adjusted R-squared. For classification, the evaluation criteria can be accuracy, precision, recall rate, f1 score, etc. Finally, select the feature combination to get the best result of the specified machine learning algorithm.

2.2 Patients selection

A total of 5 patients affected with locally advanced squamous cell carcinoma of the anal canal was enrolled in the study as first step of the study (phase 1); the purpose was to understand the feasibility of identifying actBM from CT imaging, so we decided to implement a classifier for each patient. Then, we increase the sample to 25 patients extracting 5 patients for training set construction while the remaining 20 were employed to validate the method (validation set). Finally, a total of 50 patients were collected for the last phase of the study with 40 patients reserved to training set and the remaining ten for validation. In the training set we had 14 males (35%; average age 64±17 years) and 26 females (65%; average age 63±10 years) while for the validation set we had 2 males (20%; average age 74±8 years) and 8 females (80%; average age 63±13 years)

All the patients were treated with RT-CHT delivered with volumetric modulated arc therapy [3;4]. During the staging work-up, ^{18}FDG-PET-CT on a Philips Gemini PET/CT tomography was performed. Data acquisition started 90 min after intravenous injection of approximately 30 MBq/kg body weight of 18 F-glucose. After a full-body CT scan, PET scans were acquired for 2.5 min/bed position. A dedicated fusion workstation (Extended Brilliance Workspace 2.0) was used for PET clinical interpretation. A non-contrast-enhanced CT of the pelvic region was acquired in the supine position with both an indexed shaped knee rest and ankle support (CIVCO Medical Solutions, Kalona, IA, USA), which was used for radiotherapy planning on a Philips ‘‘BigBore’’ CT scanner (Philips Medical System, Eindhoven, NL). Pixel spacing was (0.93 mm, 0.93 mm, 3 mm) for CT. Using VELOCITY platform (Varian Medical Systems, Palo Alto, CA), planning CT was co-registered with 18 FDG-PET images on a rigid registration and the co-registered PET images were then resampled to match the

voxel dimensions of the planning CT to up-sample the lower digital resolution image to the higher resolution one [33].

2.3 Active bone marrow identification and specific configuration of parameters for Machine Learning

Bone tissue has two forms, both of which are found in every bone in the human body: compact bones and spongy bones. The main difference between the two forms is how the bone mineral is organized and how much empty space there is between the solidified extracellular matrix. The compact bones look very strong, and the spongy bones are composed of coagulated extracellular matrix arranged in a mesh or spongy shape. Most of most bone tissue is made of spongy bone where there is more uncompressed space. Bone tissue only accounts for between 10-70% of the usable volume, depending on how spongy it is. The rest is mainly composed of bone marrow (BM), although there are also blood vessels, lymphatic vessels and nerves passing through the gaps. In spongy bone, bone tissue is arranged into trabeculae, which are interconnected bone tissue columns, forming a spongy bone grid. Within a single trabecula, there are concentric slices, and the bone cells in the cavities are connected to each other by small tubes, similar to the tissue arrangement in the bone cells of compact bone. However, unlike bones, the trabecula has no central blood vessels or perforated tubes containing blood vessels, lymphatic vessels and nerves. The blood vessels and nerves of the cavernous bone pass through the space between the trabeculae and do not require a separate channel. Sponge bone is sometimes called cancellous bone or trabecular bone. The longs bones of the body, found in the arms, legs, hands, and feet of the body, have an additional feature unique to their long shape. In the diaphysis, or shaft, of each long bone, the is a central hollow cavity, called the medullary cavity. Having no heavy osseous tissue in the center of the long bones makes them lighter. The non-long bones just rely on having spongy bones in their interior to reduce their overall mass. The medullary cavity, live the spaces in spongy bone, is filled with bone marrow.

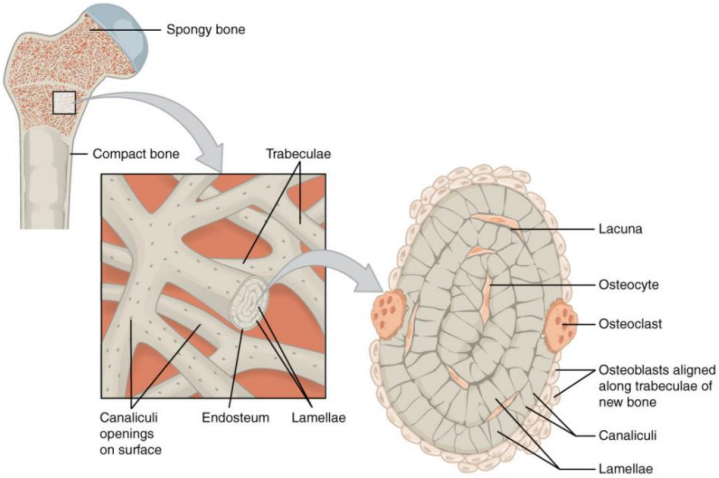


Figure 3. The structural features of spongy bone.

Pelvic bone marrow (PBM) was delineated on planning CT as a whole and then divided in 3 subregions: iliac bone marrow (IBM), lower pelvis bone marrow (LPBM) and lumbosacral bone marrow (LSBM). [34]. The ^{18}FDG-PET standardized uptake values (SUVs) within the PBM volume of all patients were corrected for body weight and considered [35]. ActBM was obtained by segmenting areas within PBM with SUV values higher than the mean SUV within the pelvic bones and then clustering them into 3 subregions: ActIBM, ActLPBM, ActLSBM [36]. The remaining PBM was defined as inactive BM (inactBM) and separated into the three subregions.

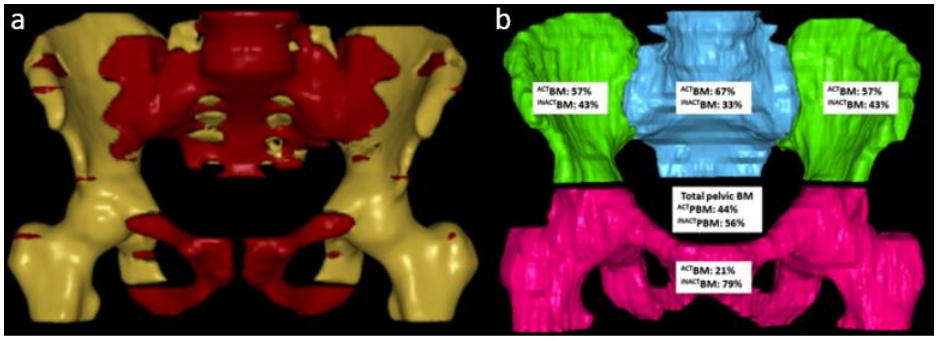


Figure 4. Three-dimensional view of the pelvic region with active (red) and inactive bone marrow (yellow) subregions (a). Three-dimensional view of the pelvic region with distribution of active and inactive bone marrow subdivided in 3 subregions: lumbar-sacral (light blue), iliac (green), and lower-pelvic (pink) bone marrow (b).

The learning process was conducted for each subregion (IBM, LPBM and LSBM) for different classifiers.

2.3.1 Extraction of Radiomic Features

In this work, radiomic features were extracted with home-made code in Matlab software version R2021a importing all RT object of the involved patients from the standard RT-Dicom format (images and structures). The first step was to use a k-means clustering algorithm which aims at labelling each voxel inside the bone as spongy or compact bone; then, the spongy bone was divided into 5-by-5 region of interest (ROI) centered in the selected voxel and labelled with the portion of PBM to which it belonged (ActBM or InactBM). Thirty-six features are then extracted: 4 first-order statistical features (mean, standard deviation, skewness and kurtosis) and 32 second-order statistical features. Twenty-two texture features were derived from the grey-level co-occurrence matrices(GLCM), that counts the number of occurrences for which a pixel with a gray level i occurs at specific distance and in a given direction from another pixel with grey level j. Starting from the GLCM, we calculated the following variables: autocorrelation, contrast, correlation1, correlation2, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity1, homogeneity2, maximum probability, sum of squares, variance, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation1, information measure of correlation2, inverse difference normalized, inverse difference moment normalized. Five texture features were extracted with the Gray Level Difference Method (GLDM), taking into account the difference of the intensities of two pixels at specific distance and in a given direction. From this analysis we extracted: contrast, angular second moment, entropy, mean and inverse difference moment. Finally, five texture features were computed from the Grey-Level Run Length Method (GLRLM), that represents the number of adjacent voxels with a certain grey level in a specific direction. In this case, the extracted variables were the following: extracted the last 5 variables: short run emphasis, long run emphasis, gray level distribution, run length distribution and run percentage. Since in this application we cannot identify a preferential texture direction, the GLCM, GLDM and GLRLM were evaluated for the four main directions (0°, 45°, 90°, 135°) with a distance of one pixel. The methodology for the extraction of radiomic features in this work was IBSI compliant (phase I indicated in the study) [37]. Then, the four matrices were averaged to extract the texture features [38] and the features related to the ROI will be assigned to the central pixel of these ROIs; then the ROI is moved, and the next pixel is characterized.

2.3.2 Training Set Construction

A training matrix was built: the rows are the ROIs and the columns are the 36 features, the three coordinates of each pixel, the label (ActBM or InactBM) plus the patient identification code (41 rows).

For **phase i)** we constructed a training set for each subject considering five slices equally spaced across the subject's CT sequence. For each slice we randomly extracted 1/5 of all valid elements (i.e., overlapping bone marrow ROIs.) Each element was characterized by the 36 features and a label indicating the class it belongs to: active or red marrow (RM) if the element overlaps an actBM region obtained from 18FDG-PET, inactive or yellow marrow (YM) otherwise. Then, the elements extracted from the five slices were pooled together to create the training set for each patient. Before constructing the classifier, the elements in each training set were discretized by means of the Chi2 algorithm [39], in order to reduce the effect of noise and improve the classification performance [40]. The Chi2 algorithm is a supervised and bottom-up discretization method based on the $\chi 2$ statistic. It tests, variable by variable, if the class of the elements in two adjacent intervals is independent from the values that the variable assumes for those elements. The algorithm stops when the $\chi 2$ value for all adjacent intervals is greater than the $\chi 2$ value for a give significance level and the number of degrees of freedom equal to the total number of classes minus 1. This process is repeated with a decreased significance level until the consistency rate of the discretized dataset is not lower than a desired value. In this study, the dependency degree [41], based on the Rough Set Theory, was used as consistency measure. It evaluates how much the class labels of the elements depend on a given set of variables. In this specific case, we run the discretization algorithm with significance levels from 0.5 to 0.001 and we stopped it when the dependency degree of the discretized dataset was lower than the one obtained with the continuous attributes. The discrete training set was used for constructing a Decision Tree (DT) for the specific patient. We employed the CART algorithm for the tree construction and the Gini Index [42] for the identification of the best splitting rule for each node.

In **phase ii)** three different training sets were created containing IBM, LPBM and LSBM voxels respectively. First, we use the same 5 patients of phase i). Then, for each subregion, 5 slices were randomly extracted from the CT sequence and, for each slice, 1/5 of the ROIs totally included in actBM areas were randomly inserted in the training set. The same number of ROIs completely overlapping inactBM areas were added, in order to obtain balance training sets. The final training sets were made up of 2400 ROIs for IBM, 2200 ROIs for LPBM, and 2500 ROIs for LSBM. Every ROI was characterized by the set of 36 radiomic features and labelled with the portion of PBM to which it belonged: active or red marrow (RM) if it overlaps an actBM area, inactive or yellow marrow (YM) otherwise. The results of 4 different classifiers (Decision Tree, K-Nearest neighbor, Neural Network and Support Vector Machine) were analyzed.

For **phase iii)** we introduce the clustering and, in the application, carried out in this case, a network was chosen following various tests SOM of size 12 and with neighborhood 1. Five slices were then selected from the 40 patients with the same conditions used to obtain the first training set. The ROIs belonging to these slices were then subjected to the SOM network, which through the mechanism competition separated the ROIs into 12 clusters. Subsequently, we consider a number of ROIs from each cluster, in proportion to the number of the cluster itself, to form the training set. Once determined how many ROIs to extract from each cluster, these were randomly selected, up to obtain

10000 ROI, 5000 representing the YM and 5000 representing the RM. The remaining were used for the composition of the test set. For the DT there are no parameters to be optimized, so the coding of the solutions to evolve with the GA is very simple: each solution is represented by a binary string vector, formed by 36 bits, one for each parameter.

In **phase iv)** we applied the same methodology of phase iii) using 40 patients for constructing training set matrix as input for three classifiers (Decision Tree, K-Nearest neighbor and Neural Network)

2.3.3 Genetic Algorithm and Classifier optimization

For each classifier, three different GAs, one for each subregion, were used to simultaneously select the input features and define the classifier parameters. [43,44]. An initial population was generated through initial solutions in binary vector form. The algorithm creates this population first with solutions considered all admissible. The initial population consists of 500 completely randomly generated solutions, with the only condition for the first 36 bits of at least two bits with a value of 1, or to have solutions that select at least two variables between possible descriptors. 500 solutions have been created for each type of classifiers and various repetitions of genetic algorithms always start from the same initial group. Each generation consists of 500 solutions. Of these solutions, 350 were selected and they will become the parents and on whom the genetic operators will apply. The method by which they were selected was the roulette wheel selection, in which each solution is assigned a probability of being chosen inversely proportional to the fitness value (the lower the fitness value, the greater the probability to be selected). This method of choice also allows to further explore the set of possible solutions, but minimally favors the maintenance during subsequent generations of solutions with less fitness. The GA runs for 100 iterations. However, an additional stop condition avoid that the GA necessarily executes 100 iterations at each repetition, saving computing time. It should be kept in mind, in fact, that the GA built in this way, where at each generation are created and evaluated the performance of 500 classifiers, the computational time is very high. Having an additional stop condition allows a considerable saving of time. As a stop condition, it was decided to interrupt the repetition if, for 30 consecutive iterations, there were no improvements of the solution with better fitness. So, every generation is taken into account of the best fit and its fitness until an individual solution with lower fitness for 30 consecutive times was found; then it stops

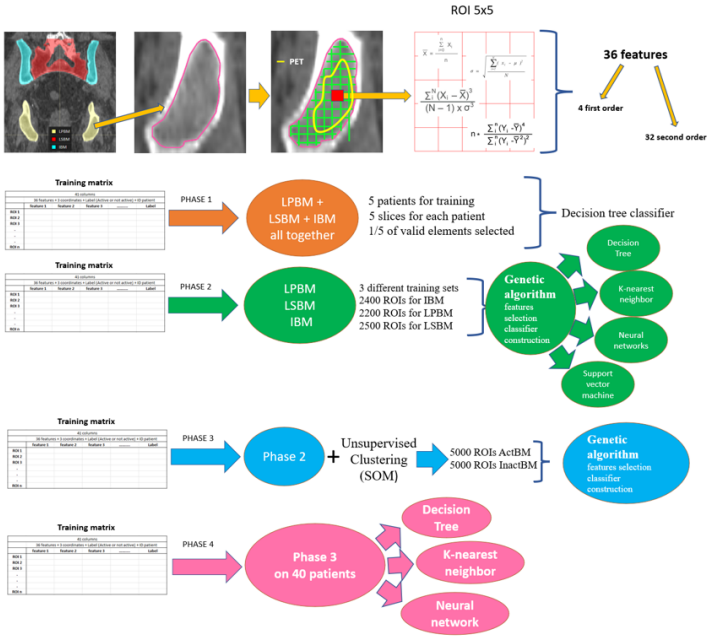


Figure 5. Visual representation of the 4 phases in which this work has been divided. Up to phase 3, 5 patients were used as training and up to 20 for validation (phase 3). For the last phase (phase 4), 40 patients were used for training and 10 for validation.

and keeps that solution as best. For the genetic operators the following values of PC=1 e PM=0.5 were chosen. These mutation and crossover probability values are very high. This choice is due to the very high number of possible solutions existing for each classifier, from 2^{36} to 2^{42} possibilities to examine. Given the huge set to be explored, high probability values for genetic operators allow to have a high variability in the solutions of the various generations, allowing to explore a high enough number of solutions and avoid that the GA converges too easily to solutions apparently optimized. The high variability avoids convergence, but not allows the solutions found to be highly optimized, since any good solutions was systematically changed. Basically, in this way the GA explores a high number of solutions, but not greatly optimizes the good solutions encountered. The fitness function has equation:

$$f = 1 - accuracy + 0{,}3 \cdot |sensitivity - specificity|$$

The lower the fitness value is better. Fitness is calculated using the values of accuracy, sensitivity and specificity that are obtained with the classifier from the selected features and the parameters encoded by the single solution. So, for each solution we constructed a classifier from the training set extracted from the 40 patients, so half of the

training set is used for the learning phase while the other half of the training set was used as a test set to test the performance of the classifier, for which we try to predict the class to which the elements of the test set belong. With the results of the predictions, it was possible to construct confusion matrix, from which the values of accuracy, sensitivity and specificity, are derived and used for fitness calculation. The fitness function is calculated in such a way as to evaluate more positively not only having solutions high accuracy, but, to a lesser extent, favors those solutions with specificity and sensitivity values similar. This additional evaluation ensures that the solutions found have balanced values and are able to correctly classify both true negatives and true positives. Solutions with low specificity e high sensitivity and vice versa, with the same accuracy, are penalized. In the segmentation of the active bone marrow, this prevents constructed classifiers from having a general tendency to under-segment or supra-segment.

2.3.4 Voxel Classification and Post-processing

Last step is to move to the bone marrow segmentation by classifying all the pixels belonging to the spongy bone slice by slice of the images relating to all 10 patients of the validation set. Before applying the individual classifiers for segmentation, an additional operation is required to improve its predictive capabilities. Each different classifier was optimized to have a different subset of features and different parameter values. Since the genetic algorithm has optimized the solutions in favor of segmentations by balancing sensitivity and specificity, combining different results, it was possible to improve overall accuracy by excluding any pixels that have been classified into a certain way from a single classifier, while pixels classified equally by both classifiers have a higher probability to be correctly classified and the combination of the results is called voting. In this work, it was decided to use a voting system with three classifiers optimized on three different criteria: i) classifier with lowest fitness value; ii) classifier with higher sensitivity value; iii) classifier with higher specificity value. For each individual slice of patient images 3 classifiers are built starting from the solutions found by the GA. Then the pixels of the slice belonging to the bone marrow are classified, in such a way as to be extracted 3 different masks of active bone marrow following the previous criteria. Finally voting is carried out with the 3 masks and the final mask is obtained. These operations are repeated for each slice until the segmentation of all the slices of the image, for all structures and classifiers. For each slice, a certain minimum number of pixels is classified as positive; however, if the number of pixels is less than 64, then those pixels are rejected and the slice is considered as not containing active bone marrow. We decided this threshold as a consequence of the nature of the PET image. In particular, since the difference in spatial resolution between PET and CT, a single pixel of the PET image corresponds to 64 pixels of CT image. When evaluating the performance of the segmentation a minimum threshold is set below such value and the contours are not considered to avoids any areas of smaller red marrow. To further improve the appearance and eliminate some artifacts often due to a bad classification of a few pixels, a morphological operation is performed on the active marrow mask by closing. In this way, the edges of the musk of the segmentation are more uniform and it may be considered as the automatic segmentation of the 3D images from CT.

2.4 Overlapping indices for the evaluation of results

Six parameters were used to evaluate the correspondence in terms of overlap, over-segmentation and under-segmentation, between the actBM regions identified by the PET scanner (PETactBM) and those detected on CT sequences (CTactBM). Three parameters were used to compare the binary masks obtained from PETactBM and CTactBM:

i) Dice index [45], which measures the total overlap between masks.

$DICE$ = (2*(PETactBM ⋂ CTactBM))/(PETactBM ⋃ CTactBM)

ii) Recall [45]:

$Recall$ = ((PETactBM ⋂ CTactBM))/PETactBM

iii) Precision [45]:

Precision = ((PETactBM ⋂ CTactBM))/CTactBM

These parameters range from 0, in case of no overlap between the two masks, to 1 in case of perfect match between them. Different studies have recommended a DSC of 0.7-0.8 to be considered a good overlap [46-48]. However, DSC can provide false impression of high agreement and, it over-penalizes small structures but is too permissive for large structures. The ratio between P and R may be used to put in evidence over-segmentation (P<R) or under-segmentation (P>R) tendency. So, we added the remaining three parameters that measure the average displacement (expressed in mm) of the voxels in CTactBM in order to perfectly match the PETactBM. The first is the Mean Distance to Conformity (MDC) index, also called Mean Distance to Agreement (MDA), which represents the average distance that all outlying points in the volume (CTactBM) must be moved in order to achieve perfect conformity with the reference volume (PETactBM). All parameters were calculated for each of the three subregions separately, considering only slices with at least one voxel of actBM in that region identified by PET.

Chapter 3

Results

3.1 Results – phase 1

Figure 6 shows an example of segmentation obtained using radiomics on CT (yellow dashed line) with the segmentation delineated from ^{18}FDG-PET (red line).

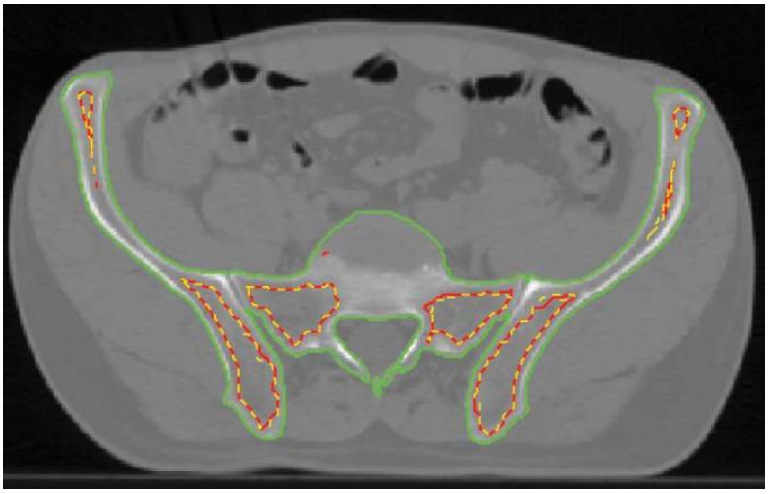


Fig. 6. Example of segmentation obtained using radiomics on CT. Green line: pelvic bone marrow segmentation. Red line: active bone marrow

For each of the 3 subregions we processed a number of slices ranging from 30 to 44 for every patient. The mean and standard deviation of the three indices is reported in Table I for the five patients and the three subregions separately. As it emerges from the table, the highest indices values are obtained for LSBM and IBM subregions, where we have an overlap between RS and CT above 75% in 4 out of 5 patients. The subregion with the lowest over-segmentation is the IBM, reaching precision above 80% for all patients. For the LPBM, not satisfying results are achieved, as the dice index is always below 0.5. However, the recall for this structure is higher than 0.7 in 4 out of 5 patients, meaning that at least the 70% of the RS segmentation is correctly recognized by the DT on CT images. These differences among the three subregions might be due to the different

amount of each of the 3 subregions we processed a number of slices ranging from 30 to 44 for every patient. The mean and standard deviation of the three indices are reported in Table 1 for the five patients and the three subregions separately. As it emerges from the table, the highest indices values are obtained for LSBM and IBM subregions, where we have an overlap between RS and CT above 75% in 4 out of 5 patients.

	IBM			LPBM			LSBM		
	Dice	***Precision***	***Recall***	***Dice***	***Precision***	***Recall***	***Dice***	***Precision***	***Recall***
Patient #1	0.88 ± 0.12	0.86 ± 0.17	0.93 ± 0.06	0.40 ± 0.19	0.27 ± 0.15	0.86 ± 0.19	0.87 ± 0.16	0.89 ± 0.12	0.87 ± 0.17
Patient #2	0.83 ± 0.13	0.89 ± 0.11	0.80 ± 0.17	0.36 ± 0.18	0.24 ± 0.15	0.79 ± 0.23	0.75 ± 0.31	0.75 ± 0.35	0.86 ± 0.19
Patient #3	0.57 ± 0.26	0.80 ± 0.22	0.48 ± 0.26	0.40 ± 0.20	0.37 ± 0.19	0.55 ± 0.27	0.40 ± 0.32	0.74 ± 0.29	0.37 ± 0.33
Patient #4	0.75 ± 0.18	0.90 ± 0.16	0.66 ± 0.21	0.43 ± 0.18	0.34 ± 0.19	0.72 ± 0.19	0.82 ± 0.22	0.85 ± 0.27	0.82 ± 0.13
Patient #5	0.93 ± 0.05	0.95 ± 0.07	0.91 ± 0.05	0.42 ± 0.10	0.28 ± 0.11	0.89 ± 0.11	0.78 ± 0.15	0.77 ± 0.21	0.82 ± 0.11

Table 1. Results of the comparison between active bone marrow identification from 18FDG-PET and CT.

These differences among the three subregions might be due to the different amount of actBM within them, with the LSBM containing almost entirely haematopoietically active bone marrow. Moreover, the actBM identification in the LPBM might be influenced by the presence of the femur. Possible improvements in this direction could be obtained by constructing a classifier specific for each subregion, so that the characteristics of the three structures can be captured more accurately.

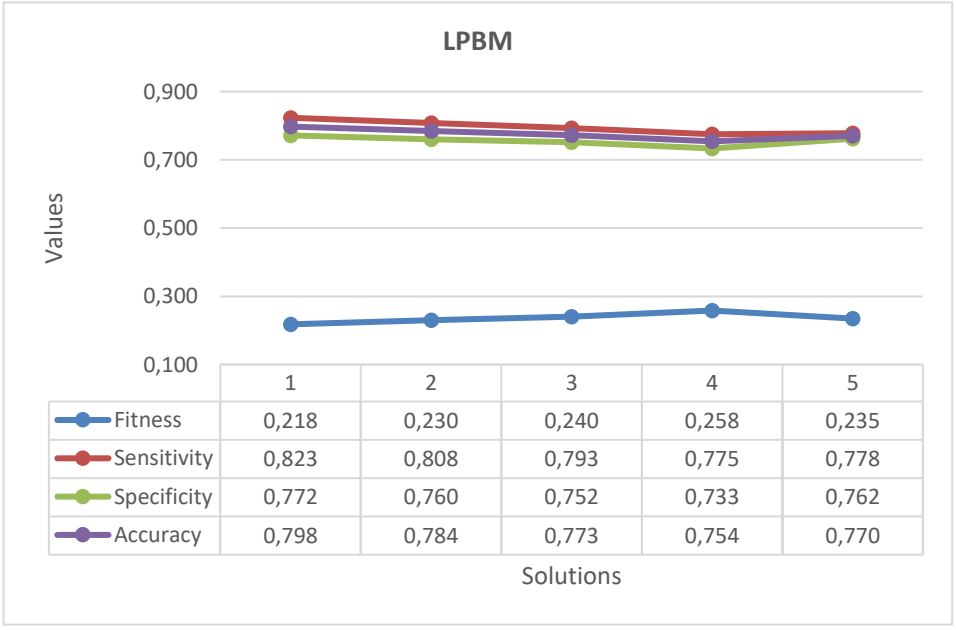
3.1.1 Discussion

This Phase 1 of this work proposes an exploratory study aiming to understand if radiomics is able to identify haematopoietically active BM from CT imaging. This was the first study in this direction, as the standard strategies for active BM detection are based on PET and MRI. These preliminary results are very promising, above all for lumbosacral and iliac structures where our approach is able to correctly identify more than 75% of actBM. With such an approach we expect possible inaccuracies due to the use of a specific classifier for each subject, since the number of elements included in the training set of each DT was substantially lower than the number of elements to be classified in all CT slices. The subsequent phases, with a larger population of patients, will be aimed precisely at better testing the generalization capability of our approach using the same feature extraction technique focused on the spatial distribution of the pixel intensities on the image; this allows for constructing classifiers that are robust to modifications in the acquisition parameters or protocol. Another aspect to be explored in the next phases will be certainly aimed to develop a specific classifier for each subregion, that could be applied for all subjects and to explore the ability of different classifiers.

3.2 Results – phase 2

3.2.1 Solution of the genetic algorithm

From the implementation of GA, for feature selection and classifier construction, 5 solutions of each classifier for each structure have been evaluated by the fitness value, sensitivity and specificity. Figure 9 shows an example for structure IBM for the SVM classifier derived from training set. We can observe how the values for the three parameters under analysis first of all do not differ much between them for various solutions and then the minimum of the fitness value also corresponds to the maximum of the value of sensitivity, specificity and accuracy; this is the case of solution n.1. This means that the choice of the fitness formulation can be considered well balanced to evaluate the best solutions from that proposed by the genetic algorithm. However, for identifying the best classifier for the three different structures, the minimum value of fitness and the maximum value of sensitivity, specificity and accuracy derived from the various solutions of the genetic algorithm applied respectively to the training and test are shown in the figure 8-9.



	1	2	3	4	5
Fitness	0,218	0,230	0,240	0,258	0,235
Sensitivity	0,823	0,808	0,793	0,775	0,778
Specificity	0,772	0,760	0,752	0,733	0,762
Accuracy	0,798	0,784	0,773	0,754	0,770

Figure 7. Different values of the fitness and maximum of the value of sensitivity, specificity and accuracy for different solutions of GA

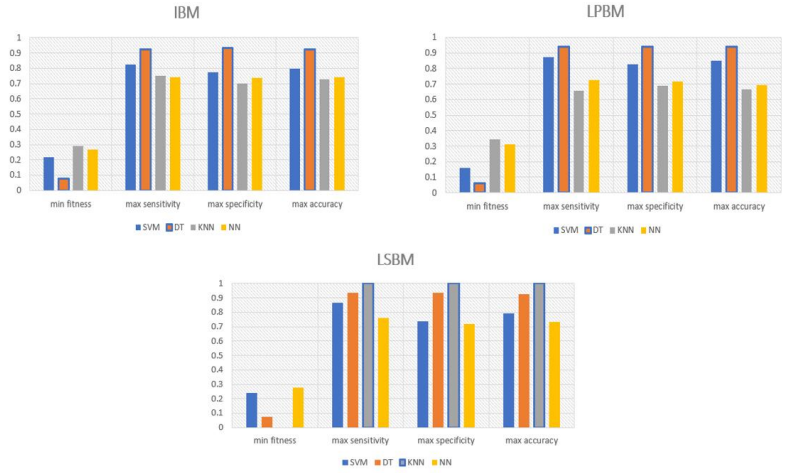


Figure 8. Results in terms of minimum fitness and maximum sensitivity, specificity and accuracy derived from training set

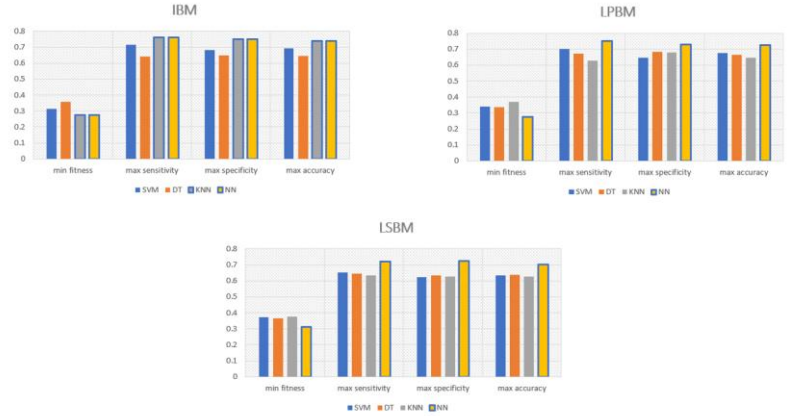


Figure 9. Results in terms of minimum fitness and maximum sensitivity, specificity and accuracy derived from validation set.

Figure 8 shows the classifier NN getting better values on all three considered structures in the validation set. However, the obtained values are not so high, setting just above 0.7 while higher values are reported for the training set in Figure 9.

3.2.2 Voting effect

To further improve the performance of the classifiers, we have included the so-called voting as the stochastic nature of genetic optimization can lead to slightly different results. To evaluate the effect of such process, we can analyze the first results in terms of Dice

Index for segmentation on a small sample of 10 patients comparing different strategies including voting for all classifiers on IBM, LPBM and LSBM structures. It can be observed how the features and parameters of the various classifiers may be better optimized through the voting process; however, for LPBM structure there is no advantage between voting and other methods with lower agreement in terms of Dice Index respect to the other two analyzed structures.

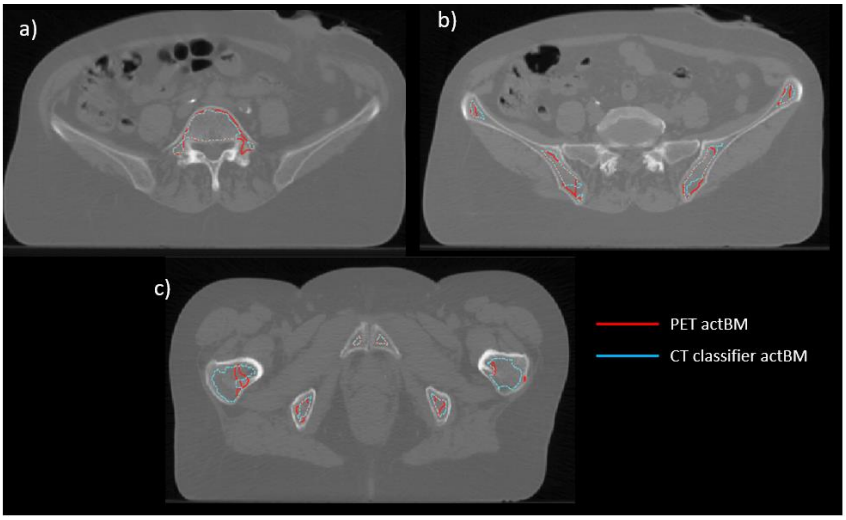


Figure 10. An example of a comparison between manual PET active bone marrow ROI (red) and the segmentation of classifier (blue) for lumbosacral bone marrow (a), iliac bone marrow (b) and lower pelvis bone marrow (c).

3.2.3 SUV threshold

So, the robustness of the method for the identification of active marrow through the SUV threshold was then evaluated by varying the previously described threshold by plus or minus 10%. In particular, just to see if there were any major changes on the LPBM structure. Unfortunately, it can be seen from Tables 2 and 3 that decreasing the threshold by 10% the results in terms of Dice agreement degrade and a general tendency to over-segment (P / R <1) increases even more. Conversely, by increasing the threshold by 10%, the Dice values improve significantly, but not enough to suggest changing the threshold. In particular, the P / R ratio also shows that the general tendency of the classifier to over-segment remains with average values of the P / R ratio that do not exceed the value of 0.66.

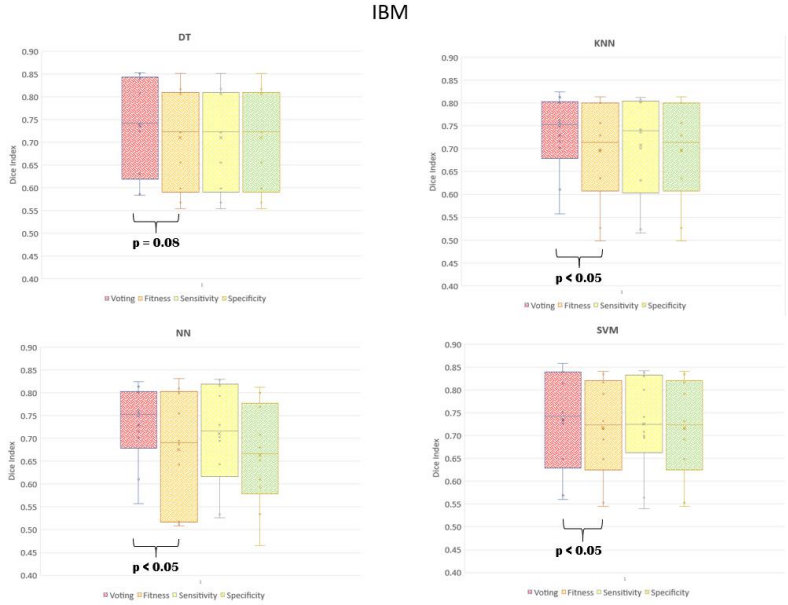


Figure 11. Box plot showing distribution of Dice index of actIBM between reference CT-Pet based contours and radiomic features with different strategies including voting for all classifiers

.

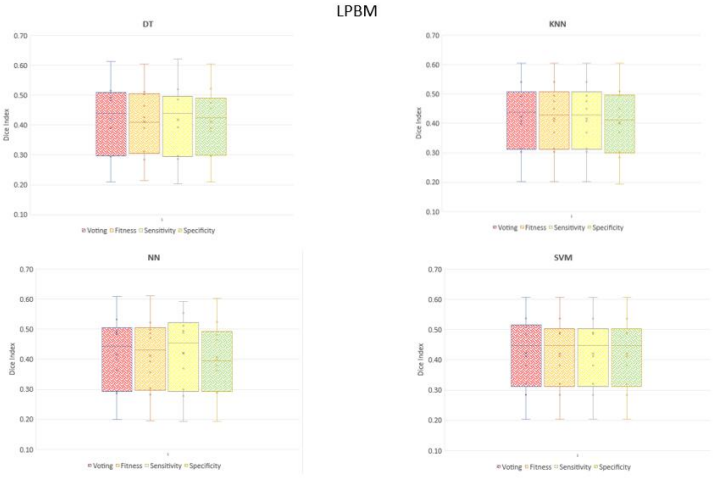


Figure 12. Box plot showing distribution of Dice index of actLPBM between reference CT-Pet based contours and radiomic features with different strategies including voting for all classifiers.

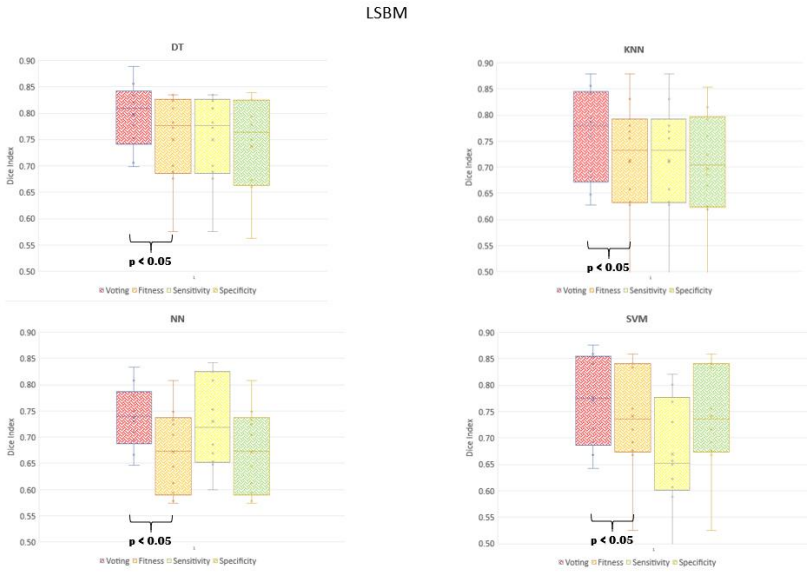


Figure 13. Box plot showing distribution of Dice index of actLSBM between reference CT-Pet based contours and radiomic features with different strategies including voting for all classifiers.

Dice Index LPBM

	DT			KNN			NN			SVM		
	SUV	**SUV+10%**	**SUV-10%**	**SUV**	**SUV+10%**	**SUV-10%**	**SUV**	**SUV+10%**	**SUV-10%**	**SUV**	**SUV+10%**	**SUV-10%**
	0.61	0.49	0.59	0.61	0.50	0.57	0.61	0.49	0.58	0.61	0.50	0.59
	0.39	0.28	0.46	0.40	0.28	0.46	0.36	0.25	0.42	0.38	0.27	0.47
	0.48	0.35	0.55	0.49	0.36	0.55	0.49	0.36	0.55	0.49	0.36	0.56
	0.51	0.34	0.59	0.46	0.32	0.55	0.48	0.33	0.56	0.49	0.32	0.58
	0.49	0.42	0.54	0.49	0.41	0.56	0.49	0.42	0.55	0.51	0.41	0.58
	0.52	0.32	0.66	0.54	0.35	0.67	0.53	0.34	0.66	0.54	0.33	0.68
	0.21	0.14	0.28	0.20	0.14	0.26	0.20	0.14	0.26	0.20	0.14	0.27
	0.29	0.07	0.24	0.30	0.07	0.24	0.29	0.07	0.23	0.29	0.07	0.22
	0.39	0.26	0.54	0.41	0.28	0.53	0.40	0.26	0.53	0.41	0.26	0.56
	0.30	0.21	0.39	0.31	0.22	0.39	0.29	0.21	0.38	0.32	0.23	0.42
mean	**0.42**	**0.29**	**0.48**	**0.42**	**0.29**	**0.48**	**0.42**	**0.29**	**0.47**	**0.42**	**0.29**	**0.49**
std	0.12	0.12	0.13	0.12	0.12	0.13	0.12	0.12	0.14	0.12	0.12	0.14
p		<0.05	<0.05		<0.05	<0.05		<0.05	<0.05		<0.05	<0.05

Table2. Different values for Dice Index for all classifiers related to SUV variation of 10% for LPBM structure.

Precision/Recall LPBM												
	DT			KNN			NN			SVM		
	SUV	SUV+10%	SUV-10%	SUV	SUV+10%	SUV-10%	SUV	SUV+10%	SUV-10%	SUV	SUV+10%	SUV-10%
	0.53	0.41	0.60	0.62	0.48	0.70	0.58	0.45	0.66	0.48	0.38	0.55
	0.46	0.33	0.64	0.56	0.40	0.77	0.51	0.37	0.71	0.42	0.30	0.58
	0.45	0.32	0.60	0.46	0.33	0.61	0.47	0.33	0.62	0.42	0.30	0.56
	0.61	0.38	0.84	0.73	0.45	1.00	0.68	0.42	0.93	0.57	0.35	0.78
	0.53	0.39	0.68	0.56	0.41	0.72	0.57	0.42	0.73	0.48	0.35	0.61
	0.59	0.31	0.93	0.62	0.33	0.98	0.63	0.33	0.99	0.52	0.27	0.82
	0.20	0.14	0.28	0.22	0.16	0.31	0.22	0.15	0.30	0.19	0.13	0.26
	0.23	0.04	0.15	0.28	0.04	0.18	0.25	0.04	0.17	0.22	0.03	0.14
	0.42	0.24	0.63	0.52	0.29	0.77	0.48	0.27	0.71	0.41	0.23	0.60
	0.32	0.23	0.47	0.38	0.27	0.55	0.37	0.26	0.54	0.30	0.21	0.43
media	**0.43**	**0.28**	**0.58**	**0.49**	**0.32**	**0.66**	**0.48**	**0.31**	**0.64**	**0.40**	**0.26**	**0.53**
std	0.14	0.11	0.22	0.15	0.13	0.25	0.15	0.12	0.24	0.12	0.10	0.20
p		<0.05	<0.05		<0.05	<0.05		<0.05	<0.05		<0.05	<0.05

Table3. Different values for the ratio of Precision and Recall for all classifiers related to SUV variation of 10% for LPBM structure.

3.2.4 Results on 20 patients

Finally, tables 3a and 3b report the values obtained on the external sample (validation set) of 20 patients for the 4 classifiers and for the 3 structures. The box plots of figures 14, 15 and 16 instead show the values of Dice, MDC and the Precision/Recall ratio, again for all the classifiers for the three structures.

		DT						KNN					
		Dice	Precision	Recall	MDC (mm)	UC (mm)	OC (mm)	Dice	Precision	Recall	MDC (mm)	UC (mm)	OC (mm)
IBM	Average	0.75	0.82	0.71	6.94	2.56	4.38	0.71	0.84	0.66	7.75	3.42	4.33
	Std	0.09	0.12	0.11	3.34	0.58	3.35	0.14	0.13	0.16	3.38	2.13	3.20
	Max	0.87	0.96	0.84	19.50	4.40	16.77	0.83	0.98	0.86	18.35	11.90	15.35
	Min	0.49	0.54	0.35	4.51	1.90	2.00	0.24	0.52	0.14	4.68	1.89	1.41
LPBM	Average	0.40	0.33	0.65	15.13	3.55	11.58	0.38	0.34	0.58	15.50	3.77	11.73
	Std	0.13	0.17	0.16	8.58	1.37	8.74	0.12	0.17	0.16	9.13	1.29	9.22
	Max	0.63	0.65	0.88	40.41	7.05	37.20	0.62	0.65	0.85	42.28	6.88	38.87
	Min	0.12	0.07	0.29	7.99	2.18	3.70	0.12	0.06	0.23	7.84	2.14	3.61
LSBM	Average	0.81	0.87	0.77	6.17	2.68	3.49	0.78	0.89	0.72	6.40	2.95	3.45
	Std	0.08	0.06	0.13	1.60	0.80	1.29	0.10	0.06	0.15	1.76	0.98	1.26
	Max	0.90	0.97	0.89	9.37	4.28	6.03	0.90	0.97	0.92	10.92	5.39	6.32
	Min	0.59	0.75	0.43	4.10	1.83	1.80	0.58	0.74	0.42	4.10	1.57	2.07

a)

		NN						SVM					
		Dice	Precision	Recall	MDC (mm)	UC (mm)	OC (mm)	Dice	Precision	Recall	MDC (mm)	UC (mm)	OC (mm)
IBM	Media	0.71	0.83	0.65	8.07	3.69	4.37	0.74	0.82	0.70	7.19	2.77	4.42
	Std	0.13	0.13	0.15	4.31	3.49	3.42	0.10	0.13	0.13	3.33	0.70	3.45
	Max	0.84	0.97	0.82	19.92	18.38	16.60	0.85	0.97	0.85	19.26	5.28	16.92
	Min	0.23	0.53	0.13	5.19	2.26	1.54	0.41	0.52	0.27	4.98	2.02	1.92
LPBM	Media	0.39	0.33	0.60	15.51	3.78	11.73	0.41	0.33	0.71	14.71	3.14	11.56
	Std	0.12	0.17	0.15	9.12	1.17	9.15	0.14	0.16	0.14	8.75	1.10	8.74
	Max	0.62	0.65	0.86	42.62	6.51	39.01	0.63	0.65	0.92	40.33	6.00	37.16
	Min	0.12	0.07	0.27	7.91	2.34	3.62	0.12	0.07	0.35	7.64	2.03	3.74
LSBM	Media	0.72	0.91	0.61	6.57	3.55	3.02	0.78	0.88	0.72	6.56	3.23	3.32
	Std	0.11	0.05	0.15	1.76	1.29	1.07	0.09	0.06	0.13	2.22	2.25	1.23
	Max	0.85	0.99	0.80	11.17	6.96	6.53	0.90	0.99	0.87	13.73	12.31	6.20
	Min	0.50	0.78	0.35	4.41	2.02	1.65	0.58	0.74	0.42	4.17	1.81	1.42

b)

Table 4. Values obtained on validation set of 20 patients with DT and KNN classifiers (a) and with NN and SVM classifiers (b) for the three considered structures.

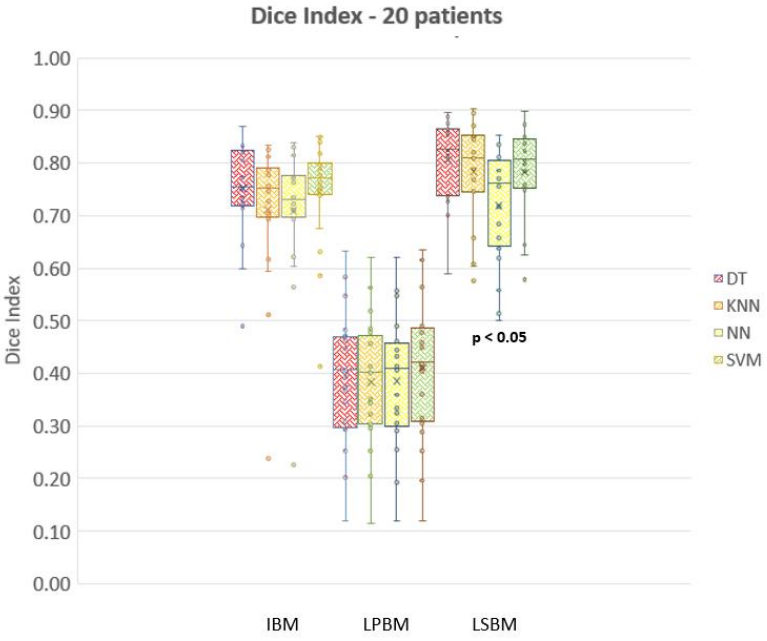


Figure 14. Box plot showing distribution of Dice index for IBM, LPBM and LSBM structures with all classifiers

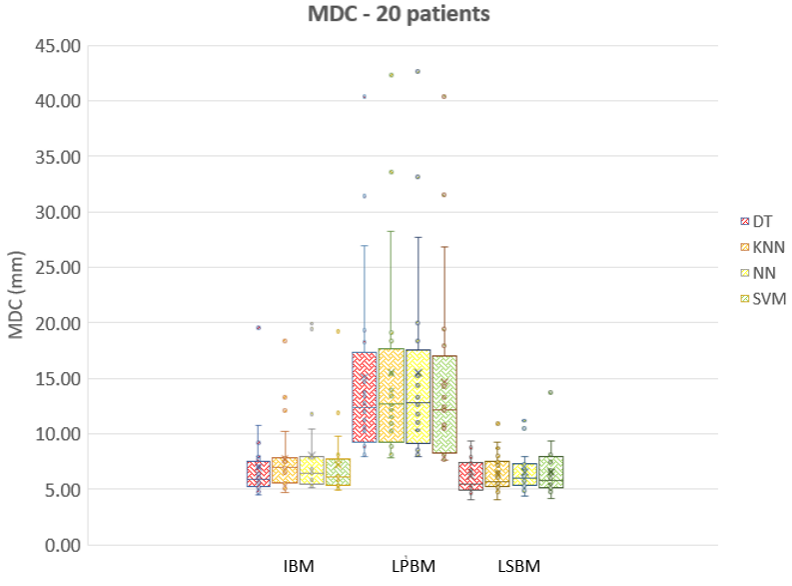


Figure 15. Box plot showing distribution of Dice index for IBM, LPBM and LSBM structures with all classifiers

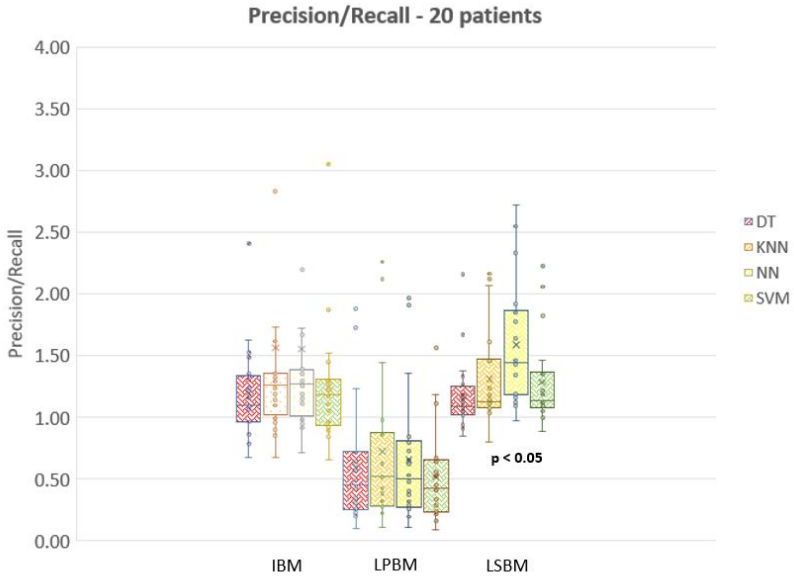


Figure 16. Box plot showing distribution of Dice index for IBM, LPBM and LSBM structures with all classifiers

These results stimulated us to investigate further and in particular we evaluate the percentage of Active BM (Red Marrow) compared to total BM. Box plot of the figure 17 illustrates very well how the percentage of ActBM within the LPBM structure is definitely lower than IBM and LPBM. Although this does not demonstrate anything particular at the moment, it will be resumed later, at the end of the presentation of the results of phase 4.

3.2.5 Discussion

With respect to the previous phase, in this study we included 20 more subjects in our population, that were used for validation. Here we compared four kinds of classifier and we proposed one classifier for each subregion, optimized using GA in terms of features subset and parameters. First of all, the results show that the performances of the 4 classifiers are similar: even analyzing results obtained on both training and validation set using different approaches to optimize GA (minimum fitness, maximum sensitivity, specificity and accuracy) there is no a classifier that performs over others within the three structures considered separately. The voting process helps to improve the results in terms of Dice index especially for IBM and LSBM, while its effect is similar to other approaches for LPBM. By changing the SUV threshold by ±10%, no changes are observed that would justify a redefinition of the threshold, therefore the identification methodology using the SUV threshold is substantially confirmed in terms of robustness.

So, the final results as in terms of Dice Index well as in terms of MDC are more or less as in phase 1 where, on average, they are satisfactory for IBM and LSBM subregions, but also in these cases there are patients with poor results. Conversely, inadequate results remain on the LPBM subregion with some outliers going towards acceptable values of Dice and MDC.

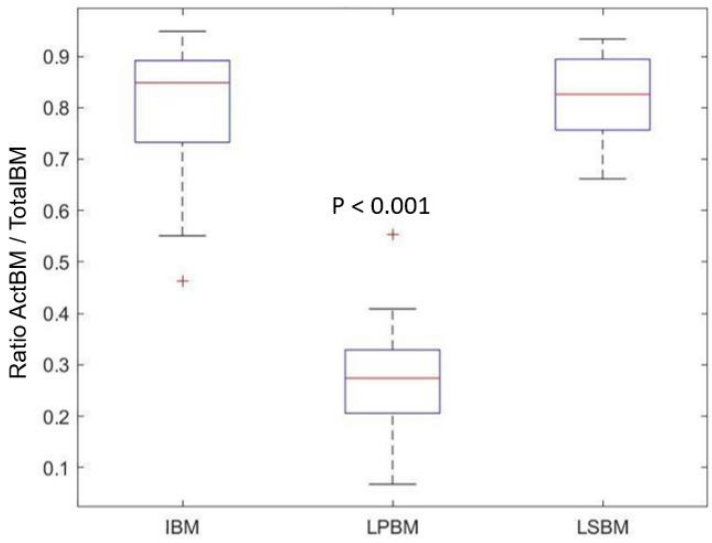


Figure 17. Box plot showing ratio ActBM/TotalBM for IBM, LPBM and LSBM structures.

There is a significant difference in the presence of percentage of active bone marrow between LPBM and the other two sub-regions, in particular in the LPBM there is a median value slightly higher than 25% while the other two sub-regions had median values of about 85%.

3.3. Results – Phase 3

Then, we proceeded with obtaining the ActBM segmentation masks following exactly the same voting operations of the 3 solutions and post-processing of the study carried out with the random training set, but with clustering. Table 5 represent the median, minimum e maximum value obtained on 20 patients for three classifiers on the three considered structures. We can observe as generally random extraction perform better in most of the tested parameters except for LPBM especially for KNN classifier.

.

	DT								
	LSBM			IBM			LPBM		
	Random	Clustering	p	Random	Clustering	p	Random	Clustering	p
Dice	**0.832 [0.59 0.92]**	0.796 [0.548 0.881]	**<0.05**	**0.773 [0.49 0.88]**	0.705 [0.542 0.832]	**<0.05**	0.447 [0.12 0.63]	0.363 [0.195 0.61]	0.44
Precision	**0.886 [0.74 0.97]**	0.831 [0.61 0.95]	**<0.05**	0.869 [0.53 0.96]	0.863 [0.52 0.96]	0.35	0.35 [0.06 0.65]	0.351 [0.15 0.68]	0.3
Recall	0.834 [0.43 0.91]	0.794 [0.5 0.90]	0.2	**0.752 [0.34 0.85]**	0.68 [0.46 0.81]	**<0.05**	0.704 [0.287 0.88]	0.576 [0.312 0.82]	0.08

	KNN								
	LSBM			IBM			LPBM		
	Random	Clustering	p	Random	Clustering	p	Random	Clustering	p
Dice	**0.82 [0.57 0.92]**	0.77 [0.53 0.87]	**<0.05**	**0.761 [0.24 0.85]**	0.695 [0.42 0.82]	**<0.05**	0.412 [0.11 0.62]	**0.61 [0.29 0.824]**	**<0.05**
Precision	**0.91 [0.73 0.97]**	0.83 [0.59 0.94]	**<0.05**	0.89 [0.52 0.98]	0.893 [0.52 0.99]	0.81	0.368 [0.06 0.65]	0.38 [0.10 0.72]	0.34
Recall	0.763 [0.42 0.92]	0.765 [0.43 0.91]	0.16	0.676 [0.14 0.86]	0.663 [0.285 0.834]	**<0.05**	**0.66 [0.23 0.85]**	0.39 [0.16 0.65]	**<0.05**

	NN								
	LSBM			IBM			LPBM		
	Random	Clustering	p	Random	Clustering	p	Random	Clustering	p
Dice	0.77 [0.501 0.89]	0.74 [0.53 0.85]	0.44	**0.752 [0.23 0.87]**	0.682 [0.36 0.81]	**<0.05**	0.415 [0.12 0.62]	0.361 [0.15 0.67]	0.5
Precision	**0.92 [0.78 0.99]**	0.83 [0.60 0.94]	**<0.05**	0.88 [0.53 0.97]	0.89 [0.54 0.98]	0.61	0.371 [0.06 0.64]	0.350 [0.10 0.69]	0.27
Recall	0.65 [0.35 0.78]	**0.69 [0.45 0.60]**	**<0.05**	0.686 [0.13 0.83]	0.634 [0.23 0.80]	0.07	0.671 [0.27 0.86]	0.504 [0.21 0.83]	0.14

Table 5. represent the median, minimum e maximum value obtained on 20 patients for three classifiers on the three considered structures. We can observe as generally random extraction perform better in most of the tested parameters except for LPBM especially for KNN classifier.

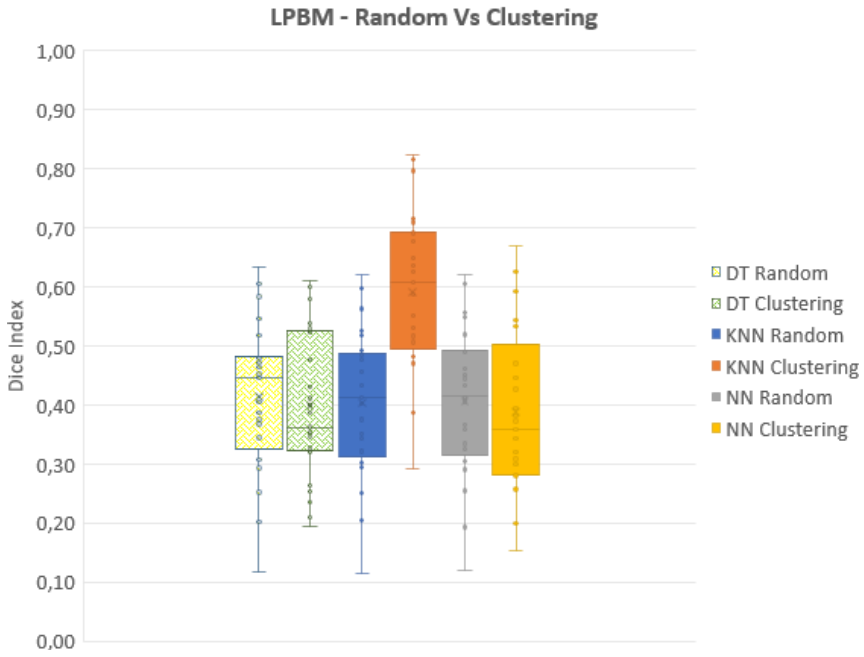


Figure 18. Difference in favor of clustering extraction from training set data.

3.3.1 Discussion

The goal of clustering is to determine the internal grouping in a set of unlabeled data, so we want to apply this concept to the extraction of ROI from training set trying to evaluate any differences with respect to the first approach which, as mentioned, involved a random extraction of the ROIs. Unfortunately, the results did not give precise indications and were rather difficult to interpret. While random appears to be significantly better in many cases, the difference with respect to clustering is still small. On the other hand, for the LPBM sub-region we have a significant increase in the results in the LPBM subregion for the KNN classifier, reason why it was decided to maintain this potentially more advantageous data extraction technique especially considering that in the last phase the training sample will be considerably increased from 20 to 40.

3.4 Results – phase 4

In this chapter we will present the results of the last phase of this work, i.e., those relating to a total of 50 patients dividing the sample between 40 patients used for training and 10 for validation; the selection of voxel for training the classifiers was made with

clustering through unsupervised learning algorithms and the results of 3 classifiers were analyzed.

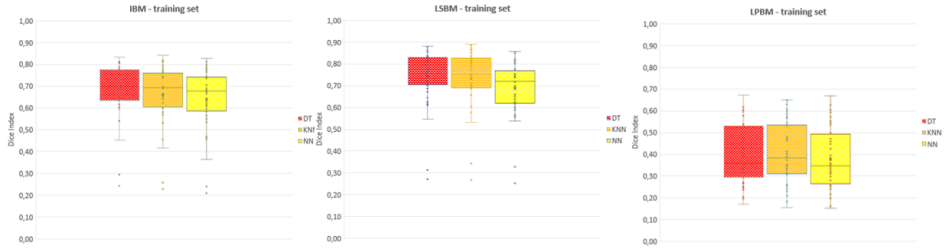


Figure 19. Box plot of the Dice Index for the three considered classifiers for the 40 patients used for training set

Figure 20. Box plot of the ratio Precision/Recall for the three considered classifiers for the 40 patients used for training set

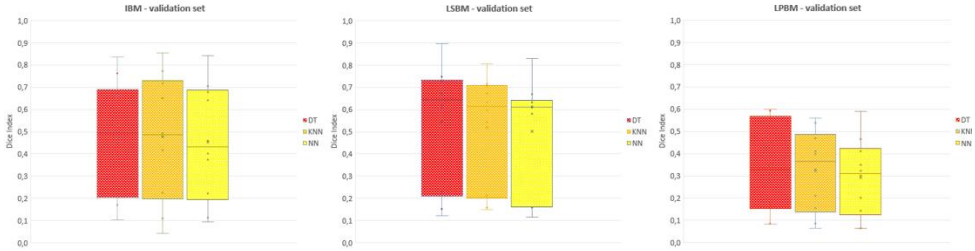


Figure 21. Box plot of the Dice Index for the three considered classifiers for the 10 patients used for validation set

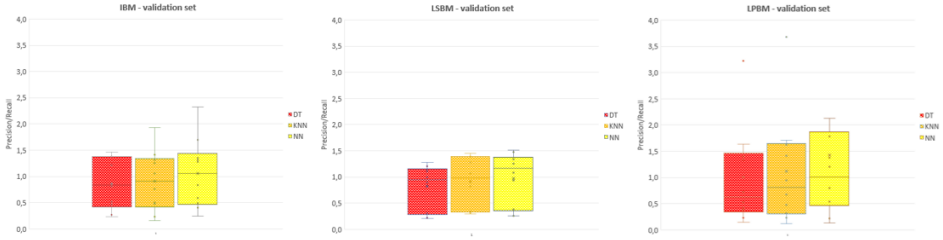


Figure 22. Box plot of the Dice Index for the three considered classifiers for the 10 patients used for validation set

The problems relating to the LPBM structure remain from figure 19 if we look at the data of the training set, the medians of the Dice Index for all three classifiers are below those obtained for the other two structures (IBM and LSBM). If we take as a reference the DT classifier which may be considered the overall best performing on all three structures, we can observe from the Precision / Recall ratio how the median value for the LPBM structure was about 0.5 (over segmentation), while for the others two structures the median of this ratio has values slightly higher than 1. Now if we move the same considerations on the Dice Index from the training set to the validation set (Figures 21 and 22), unfortunately the data worsens significantly, not only on LPBM, but also in the other two considered structures (IBM and LSBM); moreover, the P/R ratio does not show any trend of over or under contouring. Considering this unexpected worsening we started again from the results illustrated in figure 17 and we analyzed the same active marrow rate on the total marrow patient by patient and structure by structure. Figure 23 illustrates patient by patient the values of the Dice Index and the P/R rate with the values of the Active/Total BM rate for the three structures next to them. Figure 24 instead shows the correlation of these data structure by structure.

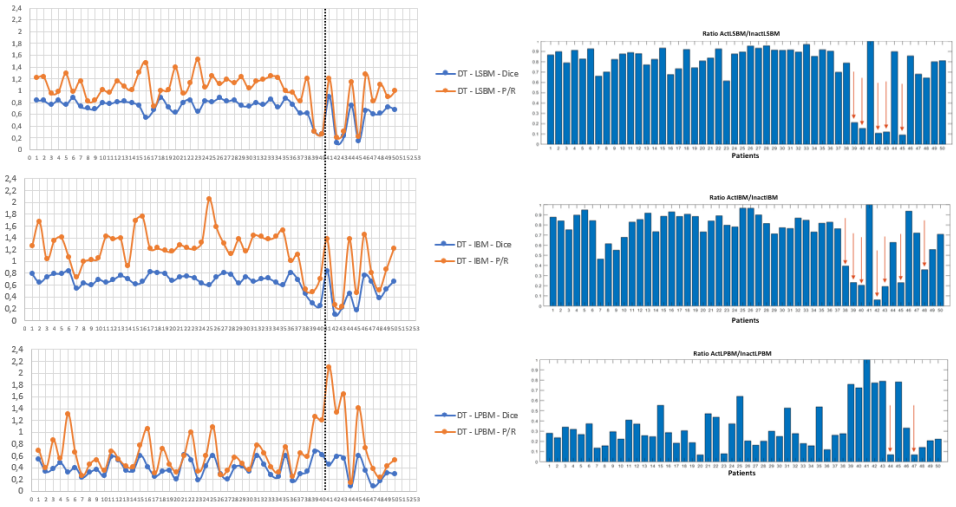


Figure 23. Dice Index and the P/R ratio reported patient by patient for the DT classifier on the left part, while the values of the values of the Active/Total BM rate for the three structures

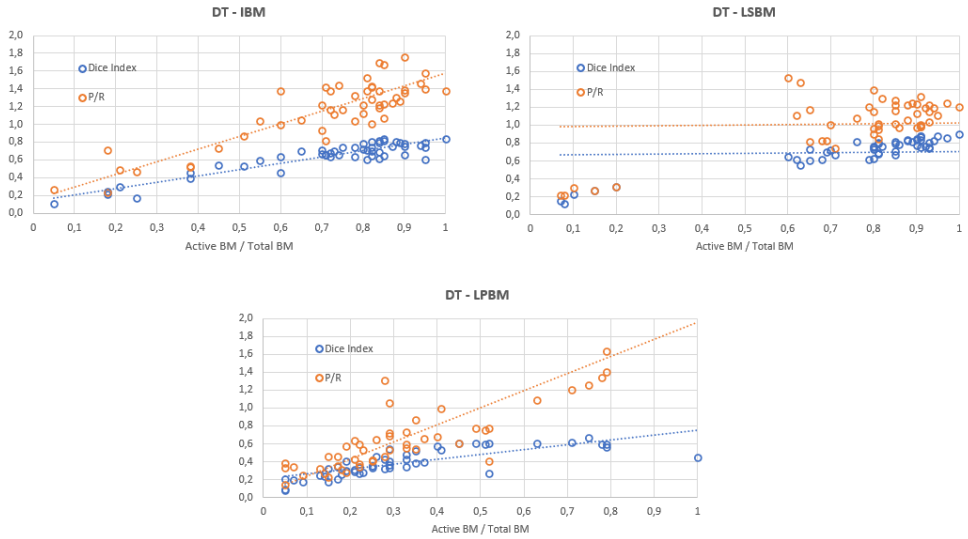


Figure 24. Correlation between the amount of Active bone marrow with both Dice Index and ratio P/R for DT classifier for the three structures.

From the data it emerges that there is a strong correlation between the quantity of active bone marrow and the correct segmentation by the various classifiers; in particular, some results from figure 16 are confirmed; the LPBM structure is the one that contains the lower part of active bone marrow (33% LPBM vs 71% IBM and 94% LSBM). Then it is interesting to note that there are some patients who present this lack also for the other two structures as shown in figure 22. Patient 39,40,42,43 and 45 present a % of active bone marrow for both LSBM and IBM below 20% and the results in terms of Dice index for those patients were well below a threshold that may be considered acceptable. Moreover, when the percentage of active LPBM was at least 50% (patients 15,25,39,40), also the results in terms of Dice index were about 0.6 that is considered as moderate correspondence [11]. From figure 23, if we look at the trend of the P/R ratio, we can clearly see how the classifier tends to over-segment (P/R <1) when the percentage of active bone marrow measured on the PET is less than 60% and this is quite independent of the considered structure.

	IBM			LSBM			LPBM		
	Dice Index	P/R	Active/Total	Dice Index	P/R	Active/Total	Dice Index	P/R	Active/Total
Average	0,64	1,16	0,71	0,70	1,02	0,94	0,39	0,69	0,33
Standard deviation	0,18	0,38	0,23	0,19	0,30	1,32	0,15	0,50	0,22
Spearman's coefficient of rank cor relation (rho)	0,74	0,29		0,82	0,42		0,86	0,85	
Significance level	P<0,0001	P=0,0403		P<0,0001	P=0,0026		P<0,0001	P<0,0001	

Table 6. Average and standard deviation values for all three structures of Dice Index, P/R and active and total marrow ratio. The Spearman's coefficient represents the correlation of each quantity with its corresponding ratio between active and total marrow

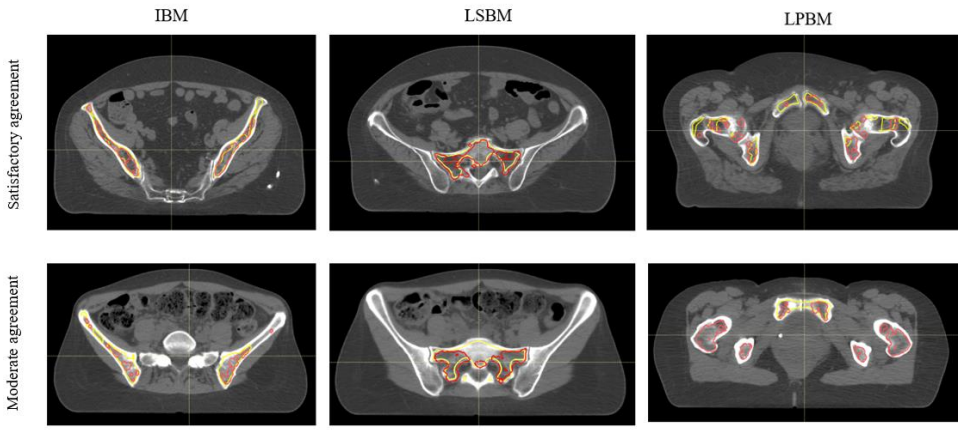


Figure 24. Visual representation of comparison between contours derived from radiomics (redline) and 18FDG-PET (yellow line) for two patients of validation dataset.

3.4.1 Discussion

Combining all the previous phases, we applied what for us could have been the best technique on a definitive sample of 50 patients using 40 patients as training and 10 as validation. In the general population, anal cancer is uncommon, with age-standardized incidence rates mostly between 1 and 2 per 100000 per year, so the extraction of 50 patients from a mono-institutional database was a great effort by the medical team that collaborated in this work. We can affirm that the size of the considered sample was appropriate with respect to the adopted methodology for classifiers, even if for some of the considered patients, the percentage of active bone marrow present in the three structures was found to be more dispersed among patients.

Both LSBM and IBM have a high proportion of active bone marrow comprised within the entire bone structure and in general, this favors the correspondence between the ActBM volumes between PET and the radiomic approach, which is shown by the mean

dice values around 0.7 for these 2 structures for values of percentage of active bone marrow greater than 70%. The likelihood of finding a geometric correspondence between volumes is generally increased by the absolute values of the same volumes, since the negative effects on the overlap performance due to slight misalignments caused by tumor delineation and position uncertainties can be smoothened.

The component of under-segmentation (P/R>1) observed for the radiomic approach when compared to PET, can be partially explained by the difference in voxel size between CT and PET scans. The average voxel size for our CT dataset was approximately 1 mm on axial images and 3 mm in the cranio-caudal direction, while the average voxel dimension for PET images was 4 mm. This difference may be responsible for the reduction in correspondence of the BM volumes as outlined by the 2 approaches, particularly for the edges of the volumes, in which PET-based delineation would consistently overestimate the edges of BM, relying on a larger-sized voxel, leading to under-segmentation coming from the CT-based radiomic delineation approach. This voxel size difference may also be responsible for the geometric uncertainties during the rigid co-registration process.

Conversely, when the percentage of active bone marrow was below 50%, we can observe from Figure 23 the ratio P/R<1 showing a tendency to over contouring from radiomics respect to ground truth. As previously mentioned, the classification of the CT sequences was achieved using classifiers capable of identifying ActBM and InactBM voxels for each specific subregion; each classifier was constructed using a dataset of 5000 voxels extracted from all of the voxels belonging to that specific subregion, using the 40 patients enrolled in the training set. If the voxels of active bone marrow are in smaller areas could be less used or not used at all for constructing the dataset, thus reducing the ability of the classifier to correctly identify a specific area in the image. This could be particularly cogent for patients that present low percentage of active bone marrow.

At the same time also the spatial resolution of PET may represent a limit when the area of active bone marrow is smaller, therefore only the addition of a possible further imaging technique such as MRI could give an important indication to see where it fits with respect to the consolidated technique of PET and the experimental one of CT described in this thesis.

Chapter 4

Conclusions

The dose received by BM comprised within pelvic bones is a predictive factor for HT occurrence during RT-CHT for squamous cell carcinoma of the anal canal; therefore, properly identifying and delineating these regions can be considered the starting point to

implement targeted approaches, during both planning and delivery to accurately spare these structures and consequently decrease the probability to develop HT.

To our knowledge, this is the first study attempting to identify and delineate active BM based on CT images only, particularly using a segmentation tool employing radiomic features extracted from CT images as input to Machine Learning classifiers to be compared to an established method based on ^{18}FDG-PET images as ground truth. Spatial cross- comparison between BM regions (IBM, LPBM and LSBM) segmented with the ^{18}FDG-PET / CT scan and our method revealed a substantial correspondence for two out of three selected regions since the beginning of our work. Mean values for Dice and MDC may be considered satisfactory as for IBM well as for LSBM even with some patient underperforming. As regard to LPBM, the agreement is globally less acceptable than other two regions with some patients having values close to an acceptability threshold.

As for the clinical application, we can certainly say that further investigations are mandatory to better understand its added value, reliability, limits, advantages and disadvantages. Finally, for the application to other departments, it's necessary to test the robustness of this approach on several computed tomography scanners.

List of References

1 Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. Br J Radiol. 2019 Aug;92(1100):20190001. doi: 10.1259/bjr.20190001. Epub 2019 Jun 5. PMID: 31112393; PMCID: PMC6724618

2 Glynne-Jones R, Nilsson PJ, Aschele C, Goh V, Peiffert D, Cervantes A, Arnold D. Anal cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow up. Radiother Oncol, 111: 330-339, 2014.

3 Arcadipane F, Franco P, Ceccarelli M, Furfaro G, Rondi N, Trino E, et al. Image-guided IMRT with simultaneous integrated boost as per RTOG 0529 for the treatment of anal cancer. Asia Pac J Clin Oncol. 2018;14(3):217-23

4 Franco P, Arcadipane F, Ragona R, Mistrangelo M, Cassoni P, Munoz F, et al. Volumetric modulated arc therapy (VMAT) in the combined modality treatment of anal cancer patients. Br J Radiol. 2016; 89(1060): 2015832;

5 Ben-Josef E, Moughan J, Ajani JA, Flam M, Gunderson L, Pollock JD, et al. Impact of overall treatment time on survival and local control in patients with anal cancer: a pooled data analysis of Radiation Therapy Oncology Groups Trials 87-04 and 98.11. J Clin Oncol. 2010; 28(34): 5061-6

6 Julie DA, Oh JH, Apte AP, Deasy JO, Tom A, Wu AJ, et al. Predictors of acute toxicities during definitive chemoradiation using intensity-modulated radiotherapy for anal squamous cell carcinoma. Acta Oncol. 2016;55(2):208–16

7 Franco P, Arcadipane F, Ragona R, Mistrangelo M, Cassoni P, Racca P, et al. Hematologic toxicity in anal cancer patients during combined chemo-radiation: a radiation oncologist perspective. Expert Rev Anticancer Ther. 2017;17(4):335-45.

8 Franco P, Ragona R, Arcadipane F, Mistrangelo M, Cassoni P, Rondi N, et al. Dosimetric predictors of acute hematologic toxicity during concurrent intensity-modulated radiotherapy and chemotherapy for anal cancer. Clin Transl Oncol. 2017;19(1):67-75

9 Franco P, Fiandra C, Arcadipane F, Trino E, Giglioli FR, Ragona R, et al. Incorporating 18FDG-PET-defined pelvic active bone marrow in the automatic treatment planning process of anal cancer patients undergoing chemo-radiation. BMC Cancer 2017;17:710

10 Wyss JC, Carmona R, Karunamuni RA, Pritz J, Hoh CK, Mell LK. [(18)F]Fluoro-2-deoxy-2-d-glucose versus 3'-deoxy-3'-[(18)F]fluorothymidine for defining hematopoietically active pelvic bone marrow in gynecologic patients. Radiother Oncol. 2016;118(1):72

11 Andreychenko A, Kroon PS, Maspero M, Jurgenliemk-Shulz I, De Leeeuw AA, Lam MG, et al. The feasibility of semi-automatically generated red bone marrow segmentation based on MR-only for patients with gynecologic cancer. Radiother Oncol. 2017;123(1):164-8

12 Li XA, Ph D, Tai A, et al: Variability of target and normal structure delineation for breast- cancer radiotherapy: A RTOG multi-institutional and multi-observer study. Int J Radiat Oncol Biol Phys 73:944-951, 2009;
13 Eminowicz G, McCormack M: Variability of clinical target volume delineation for definitive radiotherapy in cervix cancer. Radiother Oncol 117:542-547, 2015
14 Voet P, Dirkx M,Teguh DN.Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiotherapy and Oncology. 98(3):373-377
15 Heimann T, Meinzer H-P: Statistical shape models for 3D medical image segmentation: A review. Med Image Anal 13:543-563, Aug. 2009
16 Maffei N, Fiorini L, Aluisio G, D'Angelo E, Ferrazza P, Vanoni V, Lohr F, Meduri B, Guidi G. Hierarchical clustering applied to automatic atlas based segmentation of 25 cardiac sub-structures. Phys Med. 2020 Jan; 69:70-80. doi: 10.1016/j.ejmp.2019.12.001. Epub 2019 Dec 10. PMID: 31835189
17 Geremia E, Clatz O, Menze BH, et al: Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage 57:378-390, Jul. 2011 57
18 Criminisi A, Shotton J, Konukoglu E: Decision forests: A Unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found Trends® Comput Graph Vis 7:81-227, 2011
19 Li W, Liao S, Feng Q, et al: Learning image context for segmentation of prostate in CT-guided radiotherapy. Med Image Comput Assist Interv 14:570-578, 2011
20 Bauer S, Nolte L-P, Reyes M: Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. In:Medical Image Computing And Computer-Assisted Intervention :MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 14; 2011. p. 354-361
21 Vaishnavee KB, Amshakala K: An automated MRI brain image segmentation and tumor detection using SOM-clustering and Proximal support vector machine classifier. In: 2015 IEEE International Conference on Engineering and Technology (ICETECH); 2015. p. 1-6; Lu J, Wang D, Shi L, et al: Automatic liver segmentation in CT images based on Support Vector Machine. In: Proceedings of 2012 IEEEEMBS International Conference on Biomedical and Health Informatics, 25; 2012. p. 333-336
22 Zhang X, Tian J, Xiang D, et al: Interactive liver tumor segmentation from CT scans using support vector classification with watershed. In:2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011; 2011. p. 6005-6008; Rendon-Gonzalez E, Ponomaryov V: Automatic lung nodule segmentation and classification in CT images based on SVM. In: 2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW); 2016. p. 1-4
23 Mahapatra D: Automatic cardiac segmentation using semantic information from random forests. J Digit Imaging 27:794-804, 2014; Pereira S, Pinto A, Oliveira J, et al: Automatic brain tissue segmentation in MR images using random forests and conditional random fields. J Neurosci Methods 270:111-123, 2016

24 Jin C, Shi F, Xiang D, et al: 3D fast automatic segmentation of kidney based on modified AAM and random forest. IEEE Trans Med Imaging 35:1395-1407, 2016;
25 Chang KW, Summers RM, Narayanan D, et al: Automated segmentation of the thyroid gland on thoracic CT scans by multiatlas label fusion and random forest classification random forest classification. Med Imaging 3: 2017;
26 Gao Y: Accurate segmentation of CT pelvic organs via incremental cascade learning and regression-based deformable models. ProQuest Diss Theses 35:153, 2016
27 Liu J, Hoffman J, Zhao J, et al: Mediastinal lymph node detection and station mapping on chest CT using spatial priors and random forest. Med Phys 43:4362-4374, 2016;
28 Serag A, Wilkinson AG, Telford EJ, et al: SEGMA: An automatic segmentation approach for human brain MRI using sliding window and random forests. Front Neuroinf 11:1-11, 2017
29 Shahin MA, Maier HR, Jaksa MB: Data division for developing neural networks applied to geotechnical engineering. J Comput Civil Eng 18:105-114, Apr. 2004
30 Brouwer CL, Steenbakkers RJ, van den Heuvel E, Duppen JC, Navran A, Bijl HP,et al. 3D Variation in delineation of head and neck organs at risk. Radiat Oncol 2012;7:32. https://doi.org/10.1186/1748-717X-7-32
31 van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. Radiother Oncol2019;137:915.https://doi.org/10.1016/j.radonc.2019.04.006
32 Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20: 273–97
33 Zhao C, Carass A, Jog A, Prince JL. Effects of Spatial Resolution on Image Registration. Proc SPIE Int Soc Opt Eng. 2016 Feb 27;9784
34 Mell LK, Kochanski JD, Roeske JC, Haslam JJ, Mehta N, Yamada SD, Hurteau JA, Collins YC, Lengyel E, Mundt AJ. Dosimetric predictors of acute hematologic toxicity in cervical cancer patients treated with concurrent cisplatin and intensity-modulated pelvic radiotherapy. Int J Radiat Oncol Biol Phys. 2006 Dec 1;66(5):1356-65
35 Rose BS, Liang Y, Lau SK, Jensen LG, Yashar CM, Hoh CK, Mell LK. Correlation between radiation dose to ^{18}F-FDG-PET defined active bone marrow subregions and acute hematologic toxicity in cervical cancer patients treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys. 2012 Jul 15;83(4):1185-91
36 Franco P, Arcadipane F, Ragona R, et al. Dose to specific subregions of pelvic bone marrow defined with FDG-PET as a predictor of hematologic nadirs during concomitant chemoradiation in anal cancer patients. Med Oncol. 2016;33(7):72
37 Zwanenburg A, Vallières M, Abdalah MA et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020 May;295(2):328-33.
38 Rosati S, Balestra G, Franco P, Fiandra C, Arcadipane F, Silvetti P, Ricardi U, Gallio E. Radiomics for identification of active bone marrow from ct: An exploratory study 2018 IEEE Life Sciences Conference, LSC 2018
39 H. L. H. Liu and R. Setiono, “Chi2: feature selection and discretization of numeric attributes,” Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, 1995

40 S. Rosati, G. Balestra, V. Giannini, S. Mazzetti, F. Russo, and D. Regge, "ChiMerge discretization method: Impact on a computer aided diagnosis system for prostate cancer in MRI," in 2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Proceedings, 2015, pp. 297–302.
41 R. Jensen and Q. Shen, Computational Intelligence and Feature Selection. Hoboken, NJ: Wiley-IEEE Press, 2008
42 J. Han, M. Kamber, and J. (Computer scientist) Pei, Data mining : concepts and techniques. Elsevier/Morgan Kaufmann, 2012.
43 S. Rosati, G. Balestra, M. Knaflitz, "Comparison of Different Sets of Features for Human Activity Recognition by Wearable Sensors," Sensors, vol. 18, no. 12,p. 4189, Nov. 2018
44 S. Rosati, C. M. Gianfreda, G. Balestra, V. Giannini, S. Mazzetti, and D. Regge, "Radiomics to Predict Response to Neoadjuvant Chemotherapy in Rectal Cancer: Influence of Simultaneous Feature Selection and Classifier Optimization," in 2018 IEEE Life Sciences Conference (LSC), 2018, pp. 65–68
45 Giannini V, Rosati S, Regge D, Balestra G. Specificity improvement of a CAD system for multiparametric MR prostate cancer using texture features and artificial neural networks. Health and Technology. Volume: 7 Issue: 1 Special Issue: SI Pages: 71-80
46 Anders LC, Stieler F, Siebenlist K, et al. Performance of an atlas-based auto segmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. Radiother Oncol 2012; 102: 68–73;
47 Mattiucci GC, Boldrini L, Chiloiro G, et al. Automatic delineation for replanning in nasopharynx radiotherapy; what is the agreement among experts to be considered as benchmark? Acta Oncologia 2013; 52: 1417–22;
48 Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. Strahlenther Onkol. 2012 Feb;188(2):160-7.
49 Jena R, Kirkby NF, Burton KE, Hoole AC, Tan LT, Burnet NG. A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes. Br J Radiol. 2010;83(985):44-51