Highlights

Computer-Aided Diagnosis of Skeletal Metastases in Multi-Parametric Whole-Body MRI

Jakub Ceranka, Joris Wuts, Ophélye Chiabai, Frédéric Lecouvet, Jef Vandemeulebroucke

- An automated CAD system is proposed, detecting skeletal metastases in whole-body MRI.
- An ablation study on the impact of whole-body MR image preprocessing was performed.
- Deep learning approach outperformed the state-of-the-art methodologies.
- Obtained detection F_2-score of 0.50 and segmentation lesion Dice coefficient of 0.53.
- CAD system could assist radiologists in the quantification of the bone tumour volume.

Computer-Aided Diagnosis of Skeletal Metastases in Multi-Parametric Whole-Body MRI

Jakub Ceranka[a,b,*,1], Joris Wuts[a,b,c,*,1], Ophélye Chiabai[c], Frédéric Lecouvet[c,2] and Jef Vandemeulebroucke[a,b,d,2]

[a] Vrije Universiteit Brussel, Department of Electronics and Informatics, Pleinlaan 2, Brussels, 1050, Belgium
[b] imec, Kapeldreef 75, Leuven, B-3001, Belgium
[c] Cliniques universitaires Saint Luc, Institut de Recherche Expérimentale et Clinique (IREC), Université catholique de Louvain (UCLouvain), Avenue Hippocrate 10, Brussels, 1200, Belgium
[d] Universitair Ziekenhuis Brussel, Department of Radiology, Laarbeeklaan 101, Brussels, 1090, Belgium

ARTICLE INFO

Keywords:
computer-aided diagnosis
metastatic bone disease
multi-parametric MRI
whole-body MRI

ABSTRACT

The confident detection of metastatic bone disease is essential to improve patients' comfort and increase life expectancy. Multi-parametric magnetic resonance imaging (MRI) has been successfully used for monitoring of metastatic bone disease, allowing for comprehensive and holistic evaluation of the total tumour volume and treatment response assessment. The major challenges of radiological reading of whole-body MRI come from the amount of data to be reviewed and the scattered distribution of metastases, often of complex shapes. This makes bone lesion detection and quantification demanding for a radiologist and prone to error. Additionally, whole-body MRI are often corrupted with multiple spatial and intensity distortions, which further degrade the performance of image reading and image processing algorithms. In this work we propose a fully automated computer-aided diagnosis system for the detection and segmentation of metastatic bone disease using whole-body multi-parametric MRI. The system consists of an extensive image preprocessing pipeline aiming at enhancing the image quality, followed by a deep learning framework for detection and segmentation of metastatic bone disease. The system outperformed state-of-the-art methodologies, achieving a detection sensitivity of 63% with a mean of 6.44 false positives per image, and an average lesion Dice coefficient of 0.53. A provided ablation study performed to investigate the relative importance of image preprocessing shows that introduction of region of interest mask and spatial registration have a significant impact on detection and segmentation performance in whole-body MRI. The proposed computer-aided diagnosis system allows for automatic quantification of disease infiltration and could provide a valuable tool during radiological examination of whole-body MRI.

1. Introduction

Detection of tumoural bone involvement is important for disease staging, therapeutic decision and evaluation of the response to treatment in patients with solid cancers with a preferential bone tropism in their metastatic distribution (prostate, breast or lung cancers for example) and in patients with primary bone malignancies (multiple myeloma) [1]. Bone is the third most common site for metastases after the liver and lungs. A postmortem examination showed that around 70% of patients with primary breast or prostate cancer had evidence of bone metastases [2]. According to another study, about 80% of patients with advanced prostate cancer had metastases to the bone [3].

Over recent years, magnetic resonance imaging (MRI) and its adaptation for whole-body coverage (WB-MRI) has demonstrated high diagnostic performance for the detection of bone involvement and is now integrated in several clinical guidelines for skeletal lesions detection, treatment decision and response evaluation in patients with multiple myeloma and metastases from solid cancers [4, 5, 6, 7].

Typically, WB-MRI examinations combine anatomic sequences for detection and characterization of lesions, and functional diffusion-weighted imaging (DWI) sequences which improve the sensitivity of the technique and allow the evaluation of tissue viability [8, 9]. DWI sequences add diagnostic value to anatomic sequences thanks to a high lesion to background contrast and extend cancer screening to lymph nodes and extraskeletal organs. A combination of anatomical and functional MRI is characterised by excellent sensitivity (86.7–93.7%) and specificity (93.6–96.8%) for the detection of metastatic bone lesions [10, 11].

Although WB-MRI has great potential for the detection of bone lesions in oncology, there have been however limitations to its widespread implementation in clinical routine. Firstly, WB-MRI suffers the already maximal occupancy of many MRI units and its relatively long acquisition time (25-50 minutes) [12].

Second, MRI provides non-quantitative information, i.e. intensities of signal cannot be compared from one acquisition to another, making it difficult to derive reproducible measurements. DWI offer a potential quantitative approach (apparent diffusion coefficient (ADC) measurements and maps), but are generally noisy, of limited spatial resolution and may suffer from spatial distortions [13]. Geometric distortions are mostly visible after the reconstruction of the whole-body image from separate stations, due to stitching

*Corresponding author
jceranka@etrovub.be (J. Ceranka); jwuts@etrovub.be (J. Wuts)
[1] Jakub Ceranka and Joris Wuts contributed equally to this work.
[2] Frédéric Lecouvet and Jef Vandemeulebroucke should be both considered as the last author.

artifacts at the station edges and misalignments between multi-parametric whole-body MRI data. Therefore, an application of an extensive image preprocessing pipeline could potentially increase image readability both for radiologist and artificial intelligence systems.

An additional challenge for integrating whole-body MRI in clinical routine comes from the large amount of data to be reviewed, making their quantification labor demanding for a radiologist but also prone to error. The sheer amount of data of whole-body, multi-parametric MRI will require dedicated software aids, currently unavailable, to allow efficient integration in the clinical workflow. Computer-aided diagnosis (CAD) could enable simultaneous consideration of the complementary image information provided by different MR approaches, while maximally exploiting the beneficial properties of all. Such an approach could automate the detection of bone metastases, leading to improved response assessment, facilitating the uniform processing of large patient studies and enabling prospective and retrospective big data studies on disease prognostic parameters.

A number of studies for various imaging modalities have been previously published, aiming at automated metastatic bone disease classification, detection, and volumetric quantification. To date, however, only a few approaches proposing automated or semi-automated CAD systems for bone metastases segmentation in whole-body MRI have been presented.

Multiple works on the classification of whole-body bone scintigraphy scans using convolutional neural networks (CNNs) are available [14, 15, 16, 17, 18]. Using a deep learning classification network, bone scans were classified as containing bone metastases or healthy. The authors reported high image classification accuracies, ranging from 0.89 to 0.96. A similar approach was presented for thoracic SPECT scans [19].

Wels *et al.* [20] presented a fully-automated method for the detection of osteolytic bone lesions from CT data. The method used a multi-stage approach subsequently applying multiple random forests (RF) discriminative models. Each random forest was consecutively trained on a subset of different image features including: 3D Haar-like, objectness measure-based, self-aligning, spacial and symmetry encoding features.

A number of authors focused on the segmentation of bone lesions in CT and MRI.

Liu *et al.* [21] proposed a two-step approach that first segments the pelvic bone structures that is later used as a mask while segmenting metastatic lesions in the pelvis. The model is trained using 334 prostate cancer patients, and a lesion detection F_1-score of 0.87 is reported. The lesion segmentation model achieves a segmentation Dice of 0.79 and 0.80 when trained on DWI and T_1-weighted images, respectively. Although only applied to the pelvic region, the work shows U-Nets' feasibility for segmenting metastatic bone disease on MRI.

Chmelik *et al.* [22, 23] proposed a CNN-based voxel-wise segmentation and classification framework for lytic and sclerotic metastatic lesions using spinal CT. The method employed a pipeline consisting of CT image preprocessing, individual voxel-based classification CNN architecture and medial axis transform post-processing algorithm for shape simplification of segmented lesion candidates. It was applied to whole-spine CT and provided high detection sensitivity of 0.80/0.92 with 1.59/3.40 (lytic/sclerotic) FP detections per vertebrae, respectively.

Moreau *et al.* [24] used a two-stage nnU-Net segmentation approach on whole-body ^{18}FDG PET/CT, first segmenting the skeleton, followed by a segmentation of lesions using a multi-channel PET/CT input and a skeleton mask. The dataset size of 25 annotated patients allowed to achieve a mean Dice score of 0.61 ± 0.16, and a detection sensitivity of 0.67.

Blackledge *et al.* [25] proposed a semi-automated approach for the quantification of diffuse bone disease from whole-body high b-value DWI-MR. The method, however, required manual selection of the DWI image b-value, segmentation intensity threshold and additional fine-tuning of the result by reviewing and excluding false positive regions lying outside of the skeleton.

Fränzle *et al.* [26] proposed a concept of a detection algorithm of multiple myeloma lesions from manually segmented vertebra. The method used multi-parametric MRI (i.e. T_1 and T_2-weighted) to extract voxel intensity features and a random forest classifier.

Almeida *et al.* [27] proposed a semi-automated lesion segmentation algorithm for WB-DWI images in multiple myeloma patients. It uses an atlas-based segmentation method to extract the skeleton from whole-body MRI, which is later followed by intensity outlier threshold-based segmentation technique to segment bone lesions using high-resolution anatomical MRI.

All aforementioned methods, require either a manual correction of a result, manual segmentation of the bone, were developed for a different imaging modality or are not suited for whole-body assessment. In our previous work [28], we presented a concept for an automated CAD system for the detection of focal bone metastases using multi-parametric WB-MRI. The method, based on voxel-based random forest classifier approach, achieved high detection sensitivity, however, was not fully suited for volumetric segmentation of the lesions.

In this work, we propose a fully-automated deep learning method for bone metastases detection and segmentation adapted to multi-parametric WB-MRI. We compared our results to those obtained from alternative metastases segmentation approaches applied to whole-body MRI found in the literature. We additionally analyse the effect of image enhancement and preprocessing pipeline applied prior to the application of such algorithms.

2. Methods

2.1. Population & MRI

The data set consisted of 30 multi-parametric WB-MRI sets obtained from 27 advanced prostate cancer patients with skeletal metastases. This included 26 patients for whom only one image set is included taken prior to receiving any treatment (newly diagnosed disease). For the one patient, three additional image sets were included that were acquired at three-month intervals to evaluate treatment response. The study was approved by the Institutional Ethics committee and all data was anonymized prior to processing. Whole-body 3D-T_1 FSE [12] or more rapid 3D-T_1 Gradient Echo (GRE) mDIXON [29] anatomical images were used together with diffusion-weighted images [30]. The following parameters were used for the 3D-T1 FSE MRI sequence: echo time (TE) = 8 *ms*, repetition time (TR) = 382 *ms*, matrix size 480x480 with pixel spacing of 0.65 *mm* and slice thickness of 1.1 *mm*. The following parameters were used for the 3D-T1 GRE mDIXON sequence: TE = 1.15 *ms*, TR = 3.6 *ms*, matrix size 432x432, pixel spacing of 1.04 *mm* and slice thickness of 1.5 *mm*. Diffusion-weighted MRI with b-values equal to 0, 50, 150 and 1000 s/mm^2 and the following imaging parameters: TR = 8421 *ms*, TE = 66 *ms*, matrix size 192x192, pixel spacing and thickness equal to 2.3 *mm* and 6.1 *mm*, were acquired.

Images were scanned using a sequential acquisition of four image stations, covering the body from the vertex to the midthighs. Besides 3D-T_1 FSE or GRE and DWI-MR images, the majority of patients had Dixon T_2-weighted MRI images obtained of the whole-body to comply with MET-RADS-P recommendations [5]. These images, or low b-value DWI images (T_2 equivalent) in acquisitions with no available T_2 Dixon images, were available by the time of image reading before annotation to rule out false positive lesions, e.g. scar tissue caused by prior treatment cycles, inflammation, benign bony islands, osteoarthritis regions and fractures caused by disease [31].

For each patient, all bone metastases were manually delineated on high-resolution T1 image following metastases reporting recommendations MET-RADS-P and confirmed by complementary information from high b-value DWI and/or T_2. A total of 201 independent bone lesions were segmented using ITK-SNAP software [32]. The mean number of lesions per patient was 6.7 ± 5.65, with an average size of 1.30 *ml* ± 4.09 *ml*. The manual delineations were done by medical image processing specialist with 7 years of experience in WB-MRI and confirmed during a consensus session with an expert radiologist, specialized in oncology imaging and bone metastases evaluation in multi-parametric WB-MRI. The image reading was additionally supported by the prospective reading of the radiologist in charge of the examination by the time of its acquisition (in routine clinical practice).

2.2. Image Prepossessing

Prior to the application of segmentation algorithms for bone metastases, all images underwent extensive preprocessing. Figure 2 and Table 1 presents an overview of image preprocessing techniques applied to each data set. The following steps were used to generate five data sets, each with a increasing number of consecutive preprocessing steps:

- **Step 1: Calculation of additional image modalities.**
  The whole-body DWI scanning protocol allowed for the acquisition of 4 b-value images (b = 0, 50, 150 and 1000 s/mm^2). Using DWI images of different b-value weighting, we additionally computed whole-body apparent diffusion coefficient maps, according to

$$f_{ADC}(x) = \frac{1}{(b_1 - b_0)} \ln \frac{f_{b_0}(x)}{f_{b_1}(x)}. \quad (1)$$

  In this equation, f is a continuous intensity map (for which we assumed an interpolation scheme), x is the vector spatial coordinate, f_{b_0} and f_{b_1} are the signal intensity maps obtained from diffusion-sensitized T_2 imaging with at least 2 values for the gradient factor b (s/mm^2). The linear regression was used to calculate the ADC value which essentially is the absolute line slope of exponential decrease of the natural logarithm of DWI signal intensity. The b-value equal to 0 s/mm^2 was not included in the computation of a mono-exponential ADC, as it can result in a measurable contribution from microperfusion which often represents microvascular flow effects [33]. The ADC map is combined with T_1-weighted anatomy and DWI high b-value image and presented in Figure 1.

- **Step 2: Atlas-based segmentation of the skeleton.**
  An automated segmentation of the skeleton based on T_1 whole-body image was performed using an atlas-based approach [34]. The segmented bones were those most relevant for metastatic bone disease involvement and included the clavicle, vertebra from the first cervical up to the sacrum, pelvis and femur bones. The segmentation was dilated using a radius equal to 10 voxels to compensate for possible under-segmentation in the obtained skeleton which might result in increased detection of false negatives number.

- **Step 3: Image noise, low-frequency bias field and inter-station intensity standardization (ISIS).**
  Anisotropic diffusion filtering [35] was applied on image stations to reduce the image noise while conserving the edge information. A potential bias

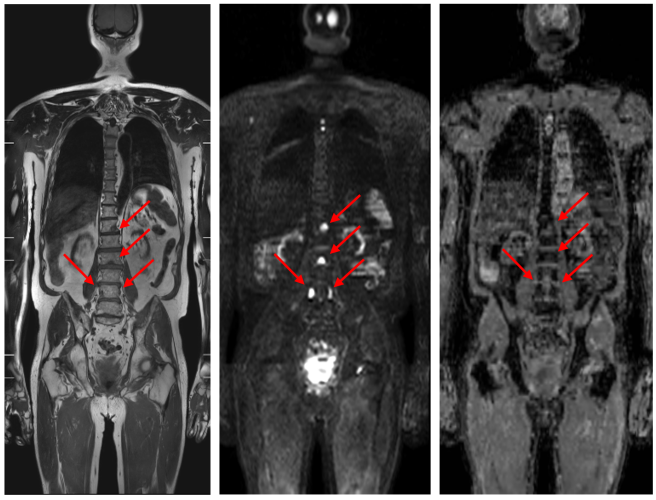


Figure 1: A coronal slice of whole-body MR modalities used in the CAD system. Bone metastases are marked with red arrows. Lesions represent different intensity profiles compared to the healthy bone, dependent on the used MR modality.

field within the station was reduced using the N4ITK non-parametric non-uniform intensity normalization algorithm [36] with default ITK parameters [37]. A linear intensity matching between the minimum intensity value and a mean of the station overlap region was sequentially applied to compensate for the differences in intensity profiles of the same modality image stations (legs to head) with a middle station (pelvis) used as a reference in order to reduce cumulative bias.

- **Step 4: Inter-station and inter-parametric spatial registration.**

  Due to patient movement during scanning and differences in applied frequency offsets [13], separate DWI image stations are often misaligned at the station edge, mostly along phase-encoding direction (anterior-posterior). Additionally, multi-parametric MR image stations (i.e. anatomy vs functional) do not correspond to each other spatially. In the first step, DWI mosaicking was performed by sequential, pairwise registration of neighbouring stations, taking the centrally located pelvis station as a reference image in order to minimize the cumulative registration error. Registration was performed rigidly, on the overlapping boundary of each image station. In a second step rigid and then deformable mapping of whole-body DWI image to T_1-weighted whole-body was performed [38].

- **Step 5: Whole-body image reconstruction and image resampling.**

  After geometrical and intensity intra-patient corrections, a whole-body volume is reconstructed by stitching image-stations into a whole-body image. In the regions of overlay between adjacent stations, linear interpolation along the cranio-caudal direction was applied providing a smooth transition between the stations. Due to differences in voxel spacing and slice thickness between different MR image modalities (i.e. T_1, DIXON, DWI), all images were resampled to the same voxel size equal to 1x1x1 *mm*, as this setting would be suited for applying all segmentation approaches described in Section 2.3. A cubic B-Spline interpolator was used for images with a floating point voxel type, and nearest neighbour interpolator for ground truth and skeleton binary masks. The choice of used imaging modalities was based on the expert opinion of the radiologists, taking into account the diagnostic capabilities of each modality and included T_1/DIXON anatomical sequence, high b-value diffusion-weighted image and an ADC map.

- **Step 6: Inter-patient intensity standardization.**
  The linear piecewise scaling algorithm first proposed by Nyúl *et al.* [39] was used to standardize the intensity profiles between different patients and MR images. First, the z-score intensity transform was applied independently to each whole-body image modality. Secondly, the average intensity histogram of the patient data distribution was acquired for each whole-body MR modality (excluding the quantitative ADC map). Finally, for a given MR modality, all patient image intensity profiles were aligned to match the standard histogram, according to six automatically selected intensity points of the histogram, corresponding to evenly spaced intensity percentiles equal to 0, 20, 40, 60, 80 and 95, image background excluded. No intensity corrections were performed on ADC images since they already represent quantitative voxel intensity values.

2.3. Segmentation Methodologies

Three segmentation strategies were compared in this study: masked thresholding applied to high b-value DWI, a voxel-based random forest segmentation and a deep learning U-Net segmentation method.

2.3.1. Masked thresholding

Inspired by the work of Blackledge *et al.* [25], we propose a modified intensity thresholding approach using dilated skeleton mask and a high b-value DWI image (Figure 3). The proposed implementation of the method did not use the applied computed DWI techniques [40], which visually maximize the contrast between diseased and normal tissues by manually tuning the computed b-value. This step was not included, as it would introduce a manual step into the proposed automatic algorithm implementation.

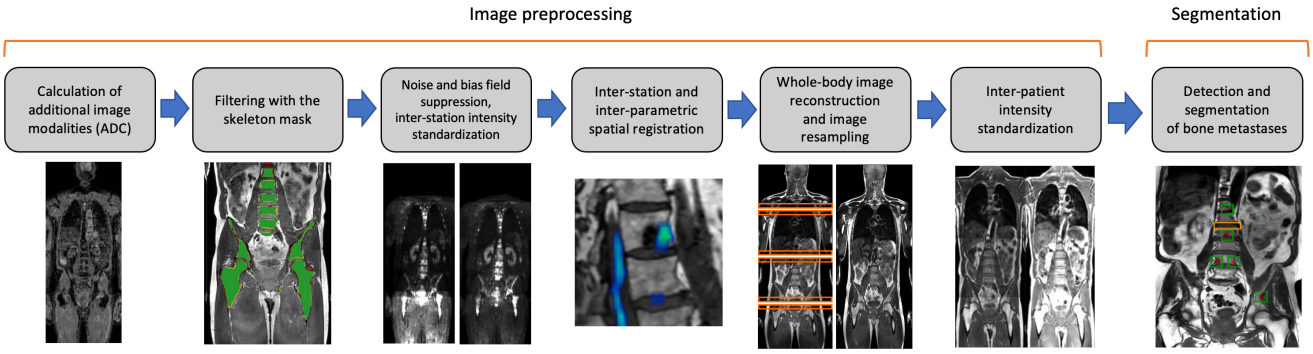


Figure 2: A schematic graphical representation of preprocessing steps applied to whole-body MRI prior to segmentation.

Table 1
An overview of preprocessing steps applied to each method.

Preprocessing step \ Method	Masked thresholding Data set 5	Random forest Data set 5	U-Net Data set 1	U-Net Data set 2	U-Net Data set 3	U-Net Data set 4	U-Net Data set 5
Calculation of additional image modalities	✓	✓	✓	✓	✓	✓	✓
Filtering with the skeleton mask	✓	✓		✓	✓	✓	✓
Noise and bias field suppression, inter-station intensity standardization	✓	✓			✓	✓	✓
Inter-station and inter-parametric spatial registration.	✓	✓				✓	✓
Whole-body image reconstruction and image resampling	✓	✓	✓	✓	✓	✓	✓
Inter-patient intensity standardization	✓	✓					✓

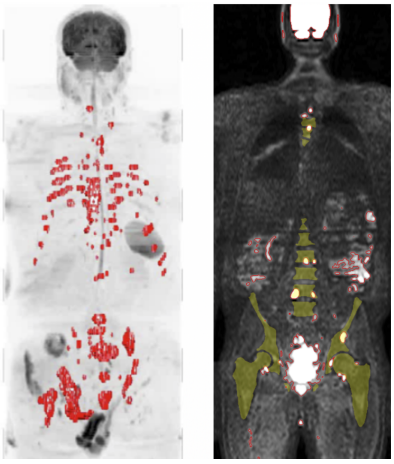


Figure 3: Example of the segmentation result obtained by Blackledge *et al.* [25] (left) and our proposed masked thresholding (right) using whole-body diffusion-weighted MRI of high b-value. The dilated skeleton mask was used to filter out false positives located outside of the skeleton region of interest.

2.3.2. Random forest

In this approach, lesion detection was treated as a voxel classification problem [26, 28] using a random forest classifier with 500 estimators, maximum depth of 100, and a Gini impurity loss. These parameters were empirically found to yield satisfying results during an initial testing [28]. A total of $n = 201$ manually delineated lesions, represented by 441782 voxels was used to extract intensity feature vectors for the training data set. The same number of healthy bone features was extracted by randomly sampling the healthy skeleton space. For each voxel, three independent intensity values were taken, each representing a different complementary MR whole-body image. Additionally, four intensity features per image were derived by filtering the image with a maximum, minimum, mean and median filter, with the kernel size equal to 10x10x10 voxels, which resulted in a total of 15 features per voxel.

2.3.3. U-Net

The implementation of the deep learning model, patch extraction and data augmentation was done using the Dynamic U-Net module available in the MONAI open-source framework [41]. This pipeline is an open-source implementation of the nnU-Net framework developed by Isensee *et al.* [42].

Patch extraction: Due to large WB-MRI dimensions (e.g. 1200x700x250 voxels), the image cannot be directly fed to a 3D U-Net. A common practice is to train the model on patches containing the structure of interest and its direct neighborhood. A patch of 128x128x128 voxel size was randomly extracted from lesion proximity or healthy bone

at every training epoch, with an equal probability. For each patch representing a lesion or healthy bone, a 4D tensor composed of four modalities (i.e. T_1, high b-value DWI, ADC and the skeleton mask) and a ground truth label map was constructed.

Data augmentation: Before images are fed to the U-Net, additional intensity normalization is applied to the whole-body MR images, per channel. In the U-Net trained on data set 1, the normalisation is applied considering all image content. For data sets 2-4, normalisation only takes into account the skeletal region.

U-Net, as most of deep learning models, has to be trained with large amount of data samples. Since medical data sets are often limited in size, an extensive data augmentation was performed in order for the network to learn invariance to deformations and obtain optimal performance. We applied several types of data augmentation techniques including: random cropping, affine transformations, random Gaussian smoothing, scaling, flip and noise addition. All augmentations were executed on the fly and with parameters randomly selected from predefined ranges to obtain uniquely augmented images each epoch. The details of the augmentation pipeline can be found in Table A.1. in Appendix A.

Network architecture: The schematic architecture of the used U-Net is presented in Figure 4. The chosen U-Net implementation is made using the Dynamic U-Net model available in MONAI which is an implementation of the nnU-Net architecture proposed by Isensee *et al.* [42], previously successfully applied on many medical segmentation tasks. The downsampling layers are made of strided convolutions. Additionally, a residual connection is added to each layer in the network.

Optimization of network hyper-parameter space: The hyper-parameters of the U-Net architecture are predefined and adopted from the implementation of nnU-Net [42]. The only hyper-parameter that was manually optimised on a validation set, was the initial learning rate. A search in the range of 0.1-0.00001 value combinations was performed to obtain an optimal value of 0.001.

Loss function: Dice-cross entropy loss with deep supervision was used to train the model [43]. It is expressed as:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c s_i^c + s_n}{\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^{c2} s_i^c + \sum_{i=1}^{N}\sum_{c=1}^{C} s_i^{c2} s_i^c + s_d} \quad (2)$$

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c log s_i^c \quad (3)$$

$$L_{Dice-CE} = \frac{1}{2}(L_{Dice} + L_{CE}) \quad (4)$$

$$L_{total} = \frac{1}{\sum_{d=1}^{D-1}\frac{1}{2^d}}\sum_{d=1}^{D-1}\frac{1}{2^d}L_{Dice-CE}^{d} \quad (5)$$

where s_n and s_d are constants set to $1e^{-5}$ to ensure computational stability. We denote image domain with N pixels and C classes where g_i^c is a binary indicator if class label c is the correct classification for pixel i, and s_i^c is the corresponding predicted probability. D is the total depth of the network, which was 5 for our model [44].

Optimiser and batch size: A SGD optimiser with Nesterov momentum ($\mu = 0.99$) was used. A polyniomial learning rate decay scheduler was used which is expressed by:

$$Lr_{epoch} = Lr_{initial} * (1 - \frac{epoch}{epoch_{max}})^{0.9} \quad (6)$$

All models where trained for 1000 epochs with a batch size of 6.

Segmentation reconstruction: The reconstruction of the segmentation image is achieved by a sliding window interference strategy. The whole-body MR image is divided into batches of 3D patches of the same size as the network input patch dimensions. The sliding window is propagated through each image allowing for a 50% patch overlap in each direction. In regions of overlap, multiple predictions made by the model are averaged with a Gaussian kernel giving more weight to voxels in the center of a patch. After the probability map of the image is reconstructed the image is thresholded to obtain a binary segmentation mask. At inference, every patch gets duplicated six times and undergoes augmentations by adding random noise and flipping over the sagittal plane. After flipping back to the original coordinates, a final prediction is computed as the mean of the six predictions and the original patch.

2.4. Train-Test Split

In order to perform a fair evaluation on the segmentation model performance, a train-test split strategy was applied. As the only hyper-parameter that needed optimisation was the learning rate, a small validation set was sufficient to see trends that lead to an optimal value. This allowed us to perform the statistical analysis on the test set with a larger amount of patients. First, out of a data set consisting of 27 patients, five patients were selected as a validation set. The split is performed on the patient level to ensure that images acquired from the same patient are kept together during training, testing and validation. Patients in the validation set were manually selected to provide representative validation data set, consisting of patients with varying involvement of metastases (low, medium and multiple lesions), intensity corruption (DWI image) and a type of anatomy sequence used (T_1, DIXON). Hyper-parameter optimisation was done solely on this small data set. Then a 5-fold stratified cross-validation strategy was applied to the remaining 25 images

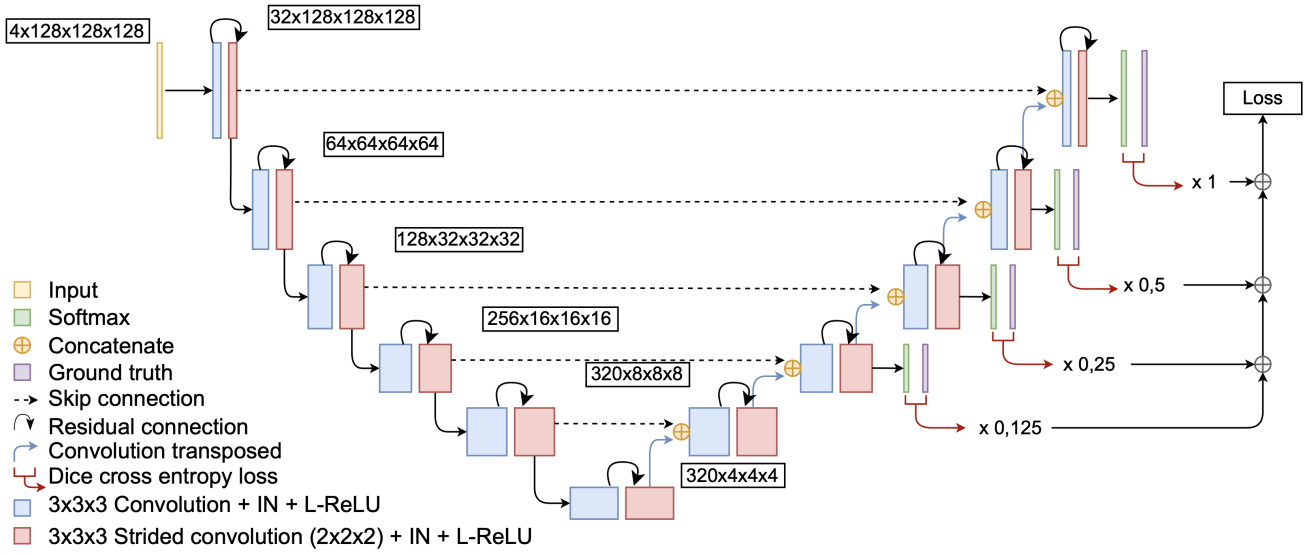


Figure 4: Schematic example of the used U-Net architecture. Each layer is composed of a strided (2x2x2) and normal convolution (blocks blue and red), each followed by an instance normalization (IN) and L-ReLU activation function with a negative slope of 0.01. Tensor addition is performed between the output of the layer and residual connection. The residual unit uses a convolution to change the input dimension to match the output dimension. To propagate spatial resolution through the model, skip connections (horizontal dashed arrows) are added between the up- and downsampling path. After the bottleneck layer, the dense representation gets upsampled to produce a binary segmentation image. The network loss is computed as a weighted sum of the last four upsampling layers.

at test time, ensuring splits have similar lesion distributions. Multiple acquisitions of the same patient are highly correlated, and therefore, all follow-up acquisitions of the same patient were kept together in a fold.

2.5. Image Postprocessing

The raw segmentation binary prediction masks often include some misclassified voxels, scattered within the skeleton. Postprocessing was applied to remove structures smaller than 30 voxels for all methods. Additionally, morphological binary closing with a radius of 2 voxels was applied to smooth the segmentation result.

2.6. Validation

The resulting segmentations were compared with the ground truth manual segmentation of bone metastases using a 5-fold stratified cross-validation. We have assessed proposed methodologies based on lesion detection performance metrics and lesion segmentation similarity metrics.

A cut-off probability threshold value was established for each method individually. This threshold value was determined on the independent validation data set and later applied to the test images in the cross-validation splits. The criteria for threshold selection for all methods involved the calculation of a F_2-score. The 45^{th} intensity percentile of the skeleton intensity distribution profile in DWI b=1000 s/mm^2 image was used as a threshold value for masked thresholding. A cut-off probability threshold value of 0.95 and $1-10e^{-5}$ was found to perform best for random forest and U-Net, respectively.

2.6.1. Detection

The detection was considered successful, if the Jaccard similarity between corresponding lesions in the segmented structure and the ground truth was larger than 0.03. A low value for the Jaccard index is chosen as the aim of detection is to purely localise the lesions. The validation criteria included for detection were: the number of true positives (TP), false positives (FP), false negatives (FN), and sensitivity of the detection (imagewise), represented by

$$sensitivity = \frac{TP}{FP+FN} \tag{7}$$

Additionally, an F_1-score and an F_2-score (emphasising on recall) represented by:

$$F_\beta = \frac{(1+\beta^2)\cdot TP}{(1+\beta^2)\cdot TP+\beta^2 FN+FP} \tag{8}$$

where, $\beta = 1$ weights equally recall and precision and $\beta > 1$ weights recall higher than precision; and FROC curve [45] are provided.

2.6.2. Segmentation

The segmentation accuracy was evaluated using quantitative metrics. The overlap measures included the global

Dice coefficient, defined as

$$DC(V_R, V_S) = \frac{2|V_R \cap V_S|}{|V_R| + |V_S|} \tag{9}$$

where V_R is the segmentation binary ground truth image and V_S is the obtained segmentation result image. As the Dice coefficient is heavily influenced by false negatives or over-segmentations, an average lesion Dice was introduced [46],

$$Dice_{Lesion}(V_R, V_S) = \frac{1}{TP} \sum_{L \varepsilon TP_{set}} \frac{2|V_R^L \cap V_S^L|}{|V_R^L| + |V_S^L|} \tag{10}$$

where V_R^L and V_S^L are segmentations of corresponding detected lesions in the ground truth and predictions respectively, and TP_{set} is a set of all found true positives.

The distance criteria included the Hausdorff distance averaged over all detected lesions (in mm). It was computed from the Euclidean distance map of the ground truth manual segmentation and the segmentation obtained by the studied method, according to the formula:

$$H_{Lesion}(V_R, V_S) = \frac{1}{TP} \sum_{L \varepsilon TP_{set}} H(V_R^L, V_S^L) \tag{11}$$

where

$$H(V_R^L, V_S^L) = \max(h(V_R^L, V_S^L), h(V_S^L, V_R^L)), \tag{12}$$

and

$$h(V_R^L, V_S^L) = \max_{a \in V_R^L} \min_{b \in V_S^L} ||a - b||. \tag{13}$$

.

2.7. Statistical Analysis

Since none of the metrics were normally distributed for all methods (Shapiro-Wilk normality test [47], $p > 0.05$), nor equality of variance was observed (F-test [48], $p > 0.05$), the Kruskal-Wallis [49] non-parametric test was used to compare differences in results between all methods. That approach was followed by a post-hoc Wilcoxon 2-tailed signed-rank test used to check for statistical significance ($p = 0.05$, with Bonferroni [50] correction) between pairs of evaluated methods.

3. Results

All proposed segmentation strategies were quantitatively validated. Results were divided into three groups: measures describing bone metastases detection performance, results describing bone metastases segmentation performance and the results of performed ablation study on the influence of whole-body MR image preprocessing on detection and segmentation accuracy of a U-Net. Results of the validation criteria for detection and segmentation accuracy of bone metastases from whole-body multi-parametric MRI, averaged over all splits of test set patients, for the different methodologies and preprocessing steps, respectively; are presented in Table 2 and in Table 3.

3.1. Detection Results

The masked thresholding detection algorithm achieved a sensitivity of 0.31 with a mean of 14.96 false positive findings per image. The voxel-based random forest segmentation method showed higher sensitivity (0.42) of the detection of bone metastases than the masked thresholding method. The mean false positives per images dropped to 11.16. The U-Net algorithm outperformed the sensitivity for both the masked thresholding and proposed RF method, scoring a detection sensitivity rate of 0.63 ($p < 0.01$, against masked thresholding) with a mean number of false positives of 6.44 ($p < 0.01$, against masked thresholding). The distribution of detection sensitivity and mean positive findings per method is represented in Figure 5 (a-b). The FROC curve, representing the detection performance in terms of sensitivity for all methods under all probability cut-off thresholds is presented in Figure 5 (c).

3.2. Segmentation Results

The masked thresholding method achieved a Dice coefficient of 0.12 with 10.07 *mm* Hausdorff similarity distance. Random forest and U-Net have further improved segmentation accuracies, scoring respectively, 0.24 and 0.33 for Dice coefficient, and 10.84 *mm* and 7.88 *mm* for Hausdorff distance. The U-Net significantly outperformed masked thresholding segmentation technique for mean global Dice ($p < 0.001$), and random forest for mean global Dice coefficient ($p < 0.01$). The average lesion Dice similarity, that reflects the segmentation quality of the detected lesions, is higher for the U-Net (0.53) compared to the thresholding and random forest method that have values of 0.38 and 0.41 respectively. The distribution of Dice similarity, average lesion Dice and Hausdorff distances per method is represented in Figure 5 (d-f). Qualitative results for three example patients are presented in Figure 6. It can be observed, that the masked thresholding method tends to undersegment the volume, random forest oversegment and U-Net provided the best qualitative result, which is the closest the boundary of the established ground truth.

3.3. Ablation study results

The distribution of sensitivity, mean false positives per image and the FROC curve for the different U-Nets are represented in Figure 7 (a-c). The U-Net trained on the non-preprocessed data achieved a detection mean sensitivity of 0.51 with a high FP rate of 31.36. Introducing the skeletal mask as a prepossessing step significantly reduced the false positive findings to 11.88 ($p < 0.01$) with a sensitivity of 0.47. Applying ISIS, noise and bias field correction further improved the detection sensitivity (0.58) at the cost of producing slightly more false positive findings (15.36). Spatial registration between the stations and modalities resulted in a significant reduction of false positive findings per image to 5.8 ($p < 0.01$), while maintaining an equal sensitivity of 0.58. Finally, the detection sensitivity is improved marginally when performing inter-patient intensity standardisation at the cost of producing slightly more false positives per image.

Table 2
Averaged bone metastases detection and segmentation results (± standard deviation) for three compared methods: masked thresholding, random forest and U-Net. The presented results were all computed on whole-body MR images with all preprocessing steps applied.

Method	Detection criteria				Segmentation criteria		
	Sensitivity	Mean FP per image	F_1-score	F_2-score	Global Dice	Lesion Dice	Hausdorff Distance (mm)
Masked Thresholding	0.31±0.32	14.96±11.91	0.14±0.19	0.19±0.19	0.12±0.13	0.38±0.15	10.07±4.37
Random Forest	0.42±0.31	11.16±9.52	0.33±0.26	0.35±0.23	0.24±0.17	0.41±0.16	10.84±6.94
U-Net	0.63±0.31	6.44±5.14	0.43±0.23	0.50±0.22	0.33±0.22	0.53±0.16	7.88±5.29

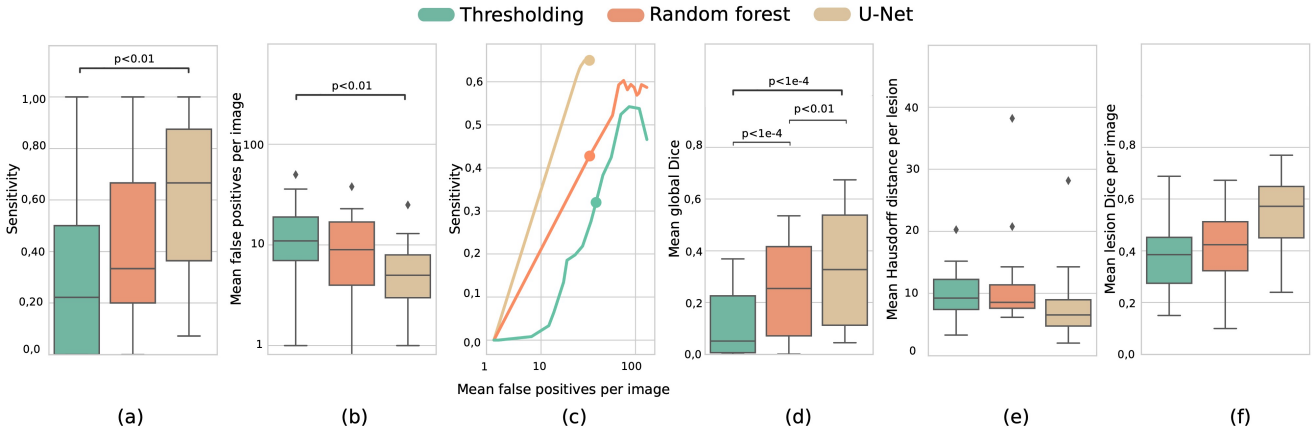


Figure 5: Distribution of the validation metrics for three different methods: masked thresholding, random forest and U-Net. Figures represent: **(a)** detection sensitivity, **(b)** false positive detections per image, **(c)** FROC plot of the lesion-based detection system, **(d)** segmentation global Dice similarity coefficient, **(e)** average Hausdorff distance per detected lesion, and **(f)** average detected lesions Dice. The final reported result for each method in a FROC is marked with a dot.

The obtained segmentation results show similar trends as the detection results. The distribution of Dice similarity, average lesion Dice and Hausdorff distances for the different U-Nets are represented in Figure 7 (d-f). All segmentation metrics improved incrementally with additional preprocessing steps. The largest improvement is achieved from introducing the skeletal mask to the U-Net. This step increased the Dice similarity from 0.05 to 0.26 ($p < 0.01$) together with a decrease in Hausdorff distance from 10.20 mm to 8.49 mm. For the remaining preprocessing steps, no statistical significance is found. Introducing ISIS and noise bias removal has the largest impact on the average lesion

Table 3
Averaged bone metastases detection and segmentation results (± standard deviation) for a U-Net with different input data sets obtained by gradually applying preprocessing steps.

Method	Detection criteria				Segmentation criteria		
	Sensitivity	Mean FP per image	F_1-score	F_2-score	Global Dice	Lesion Dice	Hausdorff Distance (mm)
U-Net - Data Set 1	0.51±0.26	69.36±31.44	0.09±0.07	0.16±0.11	0.05±0.04	0.33±0.15	10.20±4.64
U-Net - Data Set 2	0.47±0.27	11.88±6.04	0.27±0.16	0.34±0.27	0.26±0.23	0.45±0.18	8.49±4.64
U-Net - Data Set 3	0.58±0.32	15.36±6.31	0.25±0.15	0.35±0.18	0.28±0.20	0.52±0.17	6.65±2.08
U-Net - Data Set 4	0.58±0.33	5.80±4.22	0.42±0.26	0.48±0.26	0.33±0.24	0.54±0.18	7.38±5.47
U-Net - Data Set 5	0.63±0.31	6.44±5.14	0.43±0.23	0.50±0.22	0.33±0.22	0.53±0.16	7.88±5.29

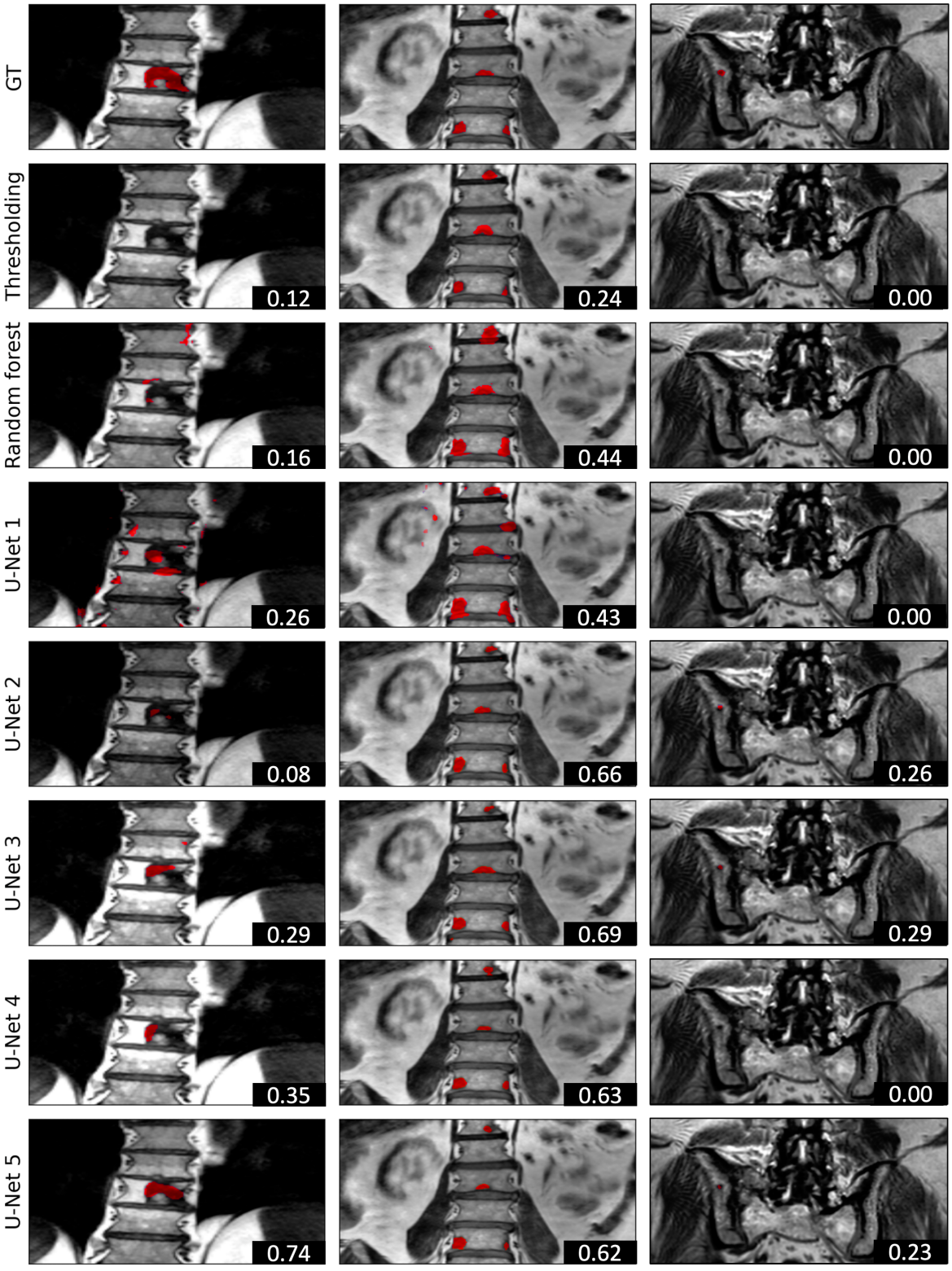


Figure 6: A coronal view extracted from anatomical whole-body T1-weighted MRI sequence for 3 patients (columns) in overlay with segmentation predictions for all compared methods and their ground truth (rows). Ground truth lesion segmentations and their predictions are represented in red. The selected patients were chosen to reflect different ranges of segmentation accuracies (high to low, global Dice coefficient reported for the presented volume of interest). Noteworthy is the first column where the spine of a patient is visualized with a large lesion located in the T8 thoracic vertebrae. Both the thresholding and the random forest methods missed the entire lesion while the final U-Net was able to detect it, with a high Dice similarity. The corresponding b1000 diffusion-weighted image of this patient is characterised by high intensity values in the skeletal region compared to the data set mean. This high b1000 value led the thresholding and random forest method to undersegment because there was insufficient difference in lesion intensity compared to skeletal intensity.

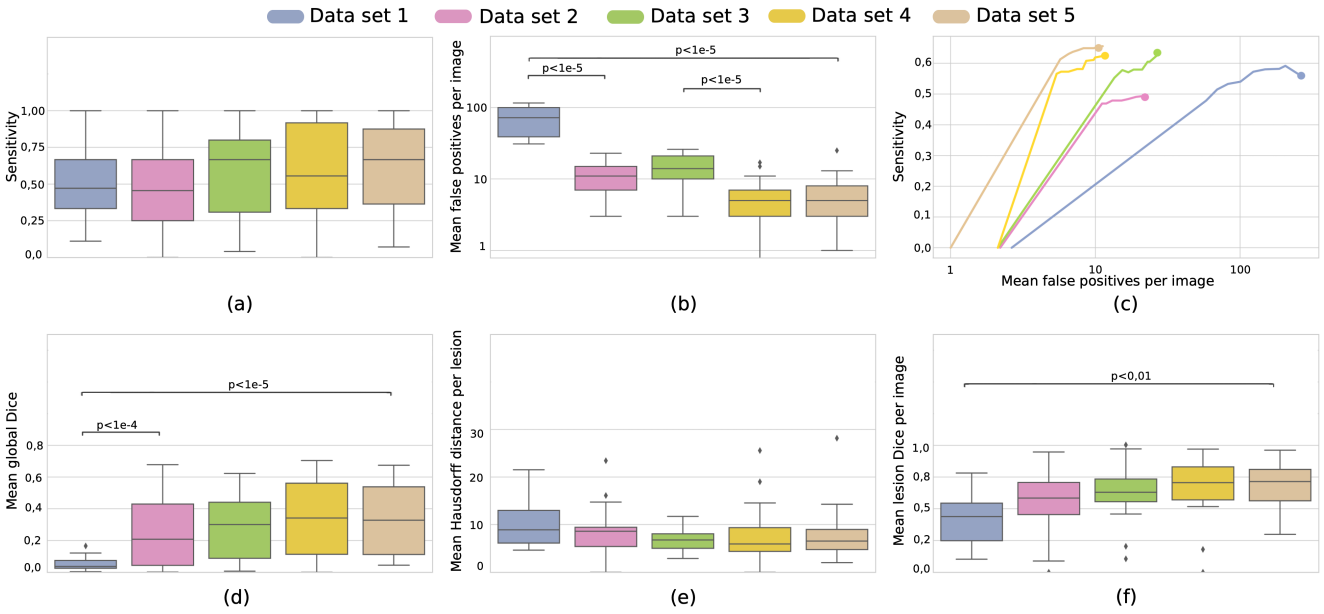


Figure 7: Distribution of the validation metrics for a U-Net trained on five different data sets with increased complexity of applied preprocessing. Figures represent: **(a)** detection sensitivity, **(b)** false positive detections per image, **(c)** FROC curve of the lesion-based detection system, **(d)** global Dice similarity coefficient, **(e)** average Hausdorff distance per detection lesion, and **(f)** average detected lesions Dice. The final reported result for each method in a FROC is marked with a dot.

Dice, increasing it from 0.45 to 0.52. The Dice similarity and mean Hausdorff distances for this data set are 0.28 and 6.65 *mm* respectively. Registration further improved the segmentation metrics, obtaining a mean Dice score of 0.33 and a Hausdorff distance of 7.38 *mm*. In the final step - the inter-patient intensity standardisation, a small decrease in segmentation metrics is seen. However, this decrease is combined with an increase in detection metrics.

3.4. Implementation Details

The networks were trained using the MONAI 0.7 library on an NVidia A100 GPU. Segmentation analysis routines were all implemented in Python 3.6, using the Numpy, Pandas, MONAI, Nibabel and SimpleITK libraries. Training a model for 1000 epochs took 8 hours and 30 minutes. The model and its source code are publically available on GitHub (https://github.com/jwutsetro/MBD_CAD).

4. Discussion and Conclusions

In this work, the concept of an automated CAD system for the detection and segmentation of focal bone metastases using multi-parametric MRI was proposed. The system relies on two main contributions: the preprocessing framework for WB-MRI and a deep learning segmentation algorithm for bone metastases. In the first part of the study, we compared the convolutional neural network approach with the state-of-the-art non-deep learning methodologies. Secondly, we propose an ablation study, where the influence of different contributions of the preprocessing pipeline are evaluated as a function of detection and segmentation algorithm output.

4.1. Detection and segmentation

We proposed a deep learning segmentation method and compared it to a masked thresholding and a random forest approaches, previously proposed in literature [25, 28]. The U-Net approach showed superior lesion detection performance over intensity thresholding of high b-value DWI image, showing the added value of non-linear classification and simultaneous evaluation of multiple complementary image modality information. Masked intensity thresholding, as a fixed value image segmentation operation, is not appropriate for the high b-value DWI images, since the intensity of bone metastases varies depending on the cellular density of the lesion. Additionally, it does not facilitate the use of anatomical MRI which are used by radiologists to establish the borders of the focal lesions. The random forest provided higher values for detection and segmentation than thresholding method, however, at the cost of over-segmentation. The U-Net outperforms both methods in detection accuracy, showing higher sensitivity (0.63, $p < 0.01$) while detecting less false positives. Additionally, the U-Net is less likely to miss all lesions present in a patient. The final U-Net finds at least one correct lesion in all patients while the thresholding and random forest method miss all lesions in 3 and 9 patients respectively. The results lead us to believe that voxel-based intensity classification, despite the inclusion of kernel derived features that take into account the immediate neighborhood, is suboptimal as it insufficiently captures the

surrounding image context and cannot fully grasp higher level features such as texture or granularity.

The U-Net outperforms the other two methods in terms of segmentation Dice similarity and average lesion Dice ($p < 0.01$ for both metrics). Especially a large difference is observed in the average lesion Dice suggesting that the deep learning method is more effective to recognize the complex shapes of the bone metastases.

An analysis on the correlation between the segmented and ground-truth volumes has been performed, where the random forest method and the proposed U-Net segmentation method showed a tendency for volume over-estimation compared to the ground-truth while we have observed that the thresholding method tends to undersegment the bone metastases. An over-segmentation is preferred over an under-segmentation for CAD tools that assess treatment response to cancer therapy. This is also reflected by the F_2-score that was used as the threshold selection criteria weighting recall over precision. With a F_2-score of 0.5, the U-Net is the favored technique for this purpose.

The segmentation accuracy was visually assessed in different skeletal regions. In Figure 6, the final U-Net with all preprocessing steps (U-Net 5) clearly outperformed the other models in terms of number of false positive findings. The qualitative results show good alignment between the segmentation and a ground truth mask for the U-Net.

We believe that the difference in quantitative validation metrics, in contrary to its visual inspection, is mainly caused by two factors. First, a large volume imbalance in the investigated segmentation problem where lesions occupy only a very small fraction of total skeleton volume and are represented by multiple small, irregular shapes. Dice coefficients of large, smooth objects with high volume-to-surface ratio (e.g. segmentation of the liver or lungs) will tend to have higher Dice coefficients despite disagreement at the edges. Small structures (lower volume-to-surface ratio) with comparable segmentation imperfections at the borders (e.g. bone metastases, blood vessels) will have lower Dice. This is even more pronounced when objects have irregular non-convex shapes. Secondly, the non-uniform intensity representation of some of the lesions in anatomical MR images, making it difficult to manually annotate the edge of the lesions in a reproducible way without introducing annotation variations across the ground-truth dataset. An example of such intensity behavior is presented in Figure 8.

4.2. Ablation study

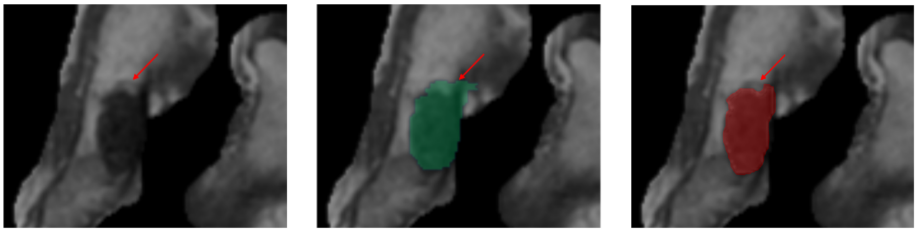
Whole-body MRI images are often corrupted with various intensity and spatial artifacts. The ablation study was performed to assess the relative impact for all performed image preprocessing steps. From the results, we can conclude that it is beneficial to include whole-body MR image preprocessing before the employment of deep learning segmentation techniques, here, a U-Net focusing on the segmentation of focal bone metastases. Applying additional preprocessing was steadily increasing the segmentation and detection performance metrics. Adding the preprocessing especially helps the model to detect more lesions while producing fewer false positives. As a result of the enhanced detection ability, the global Dice similarity did also improve.

The worst results in the evaluated comparison came from the data set where no preprocessing has been introduced to the multi-parametric WB-MRI images, before feeding them to a U-Net. Obtained segmentation results not only suffered from low sensitivity, but also from extensive number of false positive findings spread around the whole image (note the log scale for the Figure 7(b). Quantitative results of the second evaluated data set (with the addition of a skeleton mask) provided substantial improvements ($p < 0.001$) reducing false positive findings as seen in Figure 7. Introducing the skeleton mask did not only allow to filter out lesions outside of the skeleton, but also helped the model to converge better in regions within the skeleton. This can be observed in the first column of Figure 6, where a U-Net trained with the skeletal mask (U-Net 2) produced fewer false positives inside and outside of the skeleton compared to one trained without the skeleton mask (U-Net 1). We believe that this improvement originates from a more efficient patch selection during training. When a skeleton mask is available, the networks are only trained on patches that contain skeletal regions allowing for the models to focus more on this specific region of interest.

An increase of sensitivity is observed with the addition of inter-station intensity standardisation, noise and bias removal. ISIS has a strong effect to compensate intensity differences between different stations within one patient, especially in DWI b1000 images, where the different stations of a patient often have large intensity variations. Not compensating such differences could lead to artifacts at the station edges, and impact other stations through erroneous inter-patient intensity standardization.

A second large drop in false positive findings is observed when introducing spatial registration, both on inter-station and inter-modality level. The registration that was included in the study is two-fold: station-to-station registration that improves station alignment and inter-modality registration that aligns the functional images with the anatomical images. We hypothesize that improvement is mainly caused by correct spatial alignment of bone metastases across different modalities (i.e. anatomical MRI, DWI and ADC) which allowed the U-Net to learn correct intensity correlations between MRI inter-modality representation of bone metastases. Finally, the last preprocessing step, involving performing inter-patient intensity standardisation, lowered the variance of almost all observed metrics although did not prove to be statistically significant.

The initial segmentation evaluation metrics (data set 1) showed low Dice coefficient and high Hausdorff distances. Adding the skeletal mask improved the results significantly by focusing the CAD on a smaller region of interest. The remaining four steps did not have a large impact individually, but after combining all of the added contributions, a noticeable and statistically significant impact was observed. Adding the last four preprocessing steps increased the



(a) (b) (c)

Figure 8: (a): An example of metastasis representation in in-phase DIXON MRI sequence located in a lower left part of the pelvis, **(b):** manual ground truth, **(c):** segmentation obtained by a U-Net. It can be observed that the top part of the lesion represented by higher intensities (red arrow) compared to the lesion core has been included as a part of the manual delineation. The result of the U-net however, does not include this region as a part of the lesion.

global Dice similarity coefficient from a mean of 0.26 to 0.33, while the Hausdorff distance dropped from 8.49 *mm* to 7.88 *mm*. To be noted is that the average lesion Dice of the five different data sets only improves marginally. This suggests that the segmentation quality of the lesions that are already found on the early stages of applied preprocessing (e.g. data set 1) only improves slightly. This is also visually confirmed on Figure 6 (middle column), where the accuracy on the border of the segmented spinal lesions is already accurate for U-Net 1. Adding the additional preprocessing steps (U-Nets 2-5) only gradually improves the segmentation accuracy but allows to drastically increase the detection rate of new lesions and reduce the false positives findings.

4.3. Limitations

A number of study limitations should be mentioned. First, the current model segments regions inside the skeleton mask. Multiple skeletal regions are however excluded in the current implementation of skeleton segmentation CAD algorithm, such as the ribs, the sternum and the humerus. These regions are excluded as the current multi-atlas segmentation method does not include these skeletal regions in the atlases. Although the metastatic bone disease is less likely to manifest itself in the peripheral skeleton, extending the skeletal mask to include such regions should be investigated in the future [51]. Additionally, as the studied patient cohort was a group of advanced prostate cancer patients with metastases to the bone, most studied lesions were of sclerotic (osteoblastic) type.

Secondly, the available data set was relatively small (30 multi-parametric images, from 27 patients covering 201 lesions). Nonetheless, the data was found sufficient to perform the comparison of different techniques. When more data is available, we expect the results to further improve, in particular for the U-Net.

This research focuses on patients with focal metastatic bone disease. Patients with diffuse disease were excluded during the annotation sessions. This is because the boundaries of metastatic disease are not well-defined for patients with diffuse disease. In the case of focal lesions, the total tumour volume is quantitative information that can be used as a biomarker that can be followed from one examination to the other. By definition, diffuse disease boundaries correspond to the boundaries of the (at least central) skeleton, and a tumor volume cannot be determined. Hence, in diffuse disease, other quantitative information is necessary. Adequate approaches are the quantification of changes in ADC values and fat fraction in the affected skeletal areas using ADC and fat fraction maps derived from DWI and Dixon sequences, respectively [52, 25].

The usability of the trained models is bounded to input images that have both an anatomical image sequence and functional sequences. An additional experiment has been conducted assessing the predictive power of multiple combinations when not all functional images are available. The results showed a decrease in the segmentation performance, resulting in a Dice coefficient drop of 65%, 29% and 13% for models trained solely on T_1, T_1-ADC and T_1-b1000 channels, respectively.

Finally, hyperparameter optimization for the U-Net was only performed for the fully preprocessed data set. However, according to Isensee *et al.* [42], different data set preprocessing steps should not influence the U-Net definition of hyper-parameters, which are mainly influenced by image size, spacing, segmentation task to perform and patch size. Those image properties remained unchanged during data set preprocessing.

4.4. Conclusion

In conclusion, this study is, to the best of our knowledge, the first to demonstrate the feasibility of automated detection and segmentation of bone metastases in WB-MRI examinations using a CAD system based on a U-Net architecture. We demonstrate that it is viable to automatically detect bone metastases from multi-parametric WB-MRI. Convolutional neural networks outperformed less complex methods and can be considered as the most promising tool, amongst those investigated to explore in further research.

Additionally, the presented ablation study showed the importance of preprocessing on the performance of U-Net deep learning segmentation algorithms. Due to the complex nature of WB-MRI, its large field of view and sensitivity to spatial and intensity artifacts, significant preprocessing

is necessary to optimize the result of the deep learning segmentation algorithm. Especially adding a rough mask that spatially limits the field of view, substantially increasing deep learning performance when little training data is available.

5. Acknowledgments

This work was funded by the Brussels Institute For Research and Innovation (Innoviris) - award number PFS-15. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

References

[1] Van Nieuwenhove, S., Van Damme, J., Padhani, A.R., Vandecaveye, V., Tombal, B., Wuts, J., et al. Whole-body magnetic resonance imaging for prostate cancer assessment: Current status and future directions. Journal of Magnetic Resonance Imaging 2022;55(3):653–680.

[2] Coleman, R.E.. Clinical features of metastatic bone disease and risk of skeletal morbidity. Clinical cancer research 2006;12(20):6243s–6249s.

[3] Tombal, B., Lecouvet, F.. Modern detection of prostate cancer's bone metastasis: is the bone scan era over? Advances in urology 2012;2012.

[4] Larbi, A., Dallaudière, B., Pasoglou, V., Padhani, A., Michoux, N., Vande Berg, B.C., et al. Whole body MRI (WB-MRI) assessment of metastatic spread in prostate cancer: therapeutic perspectives on targeted management of oligometastatic disease. The Prostate 2016;76(11):1024–1033.

[5] Padhani, A.R., Lecouvet, F.E., Tunariu, N., Koh, D.M., De Keyzer, F., Collins, D.J., et al. METastasis reporting and data system for prostate cancer: practical guidelines for acquisition, interpretation, and reporting of whole-body magnetic resonance imaging-based evaluations of multiorgan involvement in advanced prostate cancer. European urology 2017;71(1):81–92.

[6] Messiou, C., Hillengass, J., Delorme, S., Lecouvet, F.E., Moulopoulos, L.A., Collins, D.J., et al. Guidelines for acquisition, interpretation, and reporting of Whole-Body MRI in myeloma: Myeloma response assessment and diagnosis system (MY-RADS). Radiology 2019;291(1):5–13.

[7] Oprea-Lager, D.E., Cysouw, M.C., Boellaard, R., Deroose, C.M., de Geus-Oei, L.F., Lopci, E., et al. Bone metastases are measurable: the role of whole-body MRI and positron emission tomography. Frontiers in oncology 2021;11.

[8] Lecouvet, F.E.. Whole-body mr imaging: Musculoskeletal applications. Radiology 2016;279(2):345–365.

[9] Larbi, A., Omoumi, P., Pasoglou, V., Michoux, N., Triqueneaux, P., Tombal, B., et al. Whole-body MRI to assess bone involvement in prostate cancer and multiple myeloma: comparison of the diagnostic accuracies of the t1, short tau inversion recovery (STIR), and high b-values diffusion-weighted imaging (DWI) sequences. European radiology 2019;29(8):4503–4513.

[10] Hamaoka, T., Madewell, J.E., Podoloff, D.A., Hortobagyi, G.N., Ueno, N.T.. Bone imaging in metastatic breast cancer. Journal of Clinical Oncology 2004;22(14):2942–2953.

[11] Yang, H.L., Liu, T., Wang, X.M., Xu, Y., Deng, S.M.. Diagnosis of bone metastases: a meta-analysis comparing 18 FDG PET, CT, MRI and bone scintigraphy. European radiology 2011;21(12):2604–2617.

[12] Pasoglou, V., Michoux, N., Peeters, F., Larbi, A., Tombal, B., Selleslagh, T., et al. Whole-body 3D T1-weighted MR imaging in patients with prostate cancer: feasibility and evaluation in screening for metastatic disease. Radiology 2014;275(1):155–166.

[13] Koh, D.M., Blackledge, M., Padhani, A.R., Takahara, T., Kwee, T.C., Leach, M.O., et al. Whole-body diffusion-weighted MRI: tips, tricks, and pitfalls. American Journal of Roentgenology 2012;199(2):252–262.

[14] Papandrianos, N., Papageorgiou, E., Anagnostis, A., Feleki, A.. A deep-learning approach for diagnosis of metastatic breast cancer in bones from whole-body scans. Applied Sciences 2020;10(3):997.

[15] Papandrianos, N., Papageorgiou, E., Anagnostis, A., Papageorgiou, K.. Bone metastasis classification using whole body images from prostate cancer patients based on convolutional neural networks application. PloS one 2020;15(8):e0237213.

[16] Cheng, D.C., Liu, C.C., Hsieh, T.C., Yen, K.Y., Kao, C.H.. Bone metastasis detection in the chest and pelvis from a whole-body bone scan using deep learning and a small dataset. Electronics 2021;10(10):1201.

[17] Hsieh, T.C., Liao, C.W., Lai, Y.C., Law, K.M., Chan, P.K., Kao, C.H.. Detection of bone metastases on bone scans through image classification with contrastive learning. Journal of Personalized Medicine 2021;11(12):1248.

[18] Han, S., Oh, J.S., Lee, J.J.. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. European Journal of Nuclear Medicine and Molecular Imaging 2022;:1–11.

[19] Lin, Q., Li, T., Cao, C., Cao, Y., Man, Z., Wang, H.. Deep learning based automated diagnosis of bone metastases with SPECT thoracic bone images. Scientific Reports 2021;11(1):4223.

[20] Wels, M., Kelm, B.M., Tsymbal, A., Hammon, M., Soza, G., Sühling, M., et al. Multi-stage osteolytic spinal bone lesion detection from CT data with internal sensitivity control. In: Medical Imaging 2012: Computer-Aided Diagnosis; vol. 8315. International Society for Optics and Photonics; 2012, p. 831513.

[21] Liu, X., Han, C., Cui, Y., Xie, T., Zhang, X., Wang, X.. Detection and segmentation of pelvic bones metastases in MRI images for patients with prostate cancer based on deep learning. Frontiers in Oncology 2021;11:773299.

[22] Chmelik, J., Jakubicek, R., Walek, P., Jan, J., Ourednicek, P., Lambert, L., et al. Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in 3D CT data. Medical image analysis 2018;49:76–88.

[23] Chmelik, J., Jakubicek, R., Jan, J., Ourednicek, P., Lambert, L., Amadori, E., et al. Fully automatic CAD system for segmentation and classification of spinal metastatic lesions in CT data. In: World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague, Czech Republic (Vol. 1). Springer; 2019, p. 155–158.

[24] Moreau, N., Rousseau, C., Fourcade, C., Santini, G., Ferrer, L., Lacombe, M., et al. Deep learning approaches for bone and bone lesion segmentation on 18FDG PET/CT imaging in the context of metastatic breast cancer. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE; 2020, p. 1532–1535.

[25] Blackledge, M.D., Collins, D.J., Tunariu, N., Orton, M.R., Padhani, A.R., Leach, M.O., et al. Assessment of treatment response by total tumor volume and global apparent diffusion coefficient using diffusion-weighted MRI in patients with metastatic bone disease: A feasibility study. PLoS ONE 2014;9(4):e91779. doi:10.1371/journal.pone.0091779.

[26] Fränzle, A., Hillengass, J., Bendl, R.. Spinal focal lesion detection in multiple myeloma using multimodal image features. In: Medical Imaging 2015: Computer-Aided Diagnosis; vol. 9414. International Society for Optics and Photonics; 2015, p. 94143B.

[27] Almeida, S.D., Santinha, J., Oliveira, F.P., Ip, J., Lisitskaya, M., Lourenço, J., et al. Quantification of tumor burden in multiple myeloma by atlas-based semi-automatic segmentation of WB-DWI. Cancer Imaging 2020;20(1):1–10.

[28] Ceranka, J., Lecouvet, F., De Mey, J., Vandemeulebroucke, J.. Computer-aided detection of focal bone metastases from whole-body multi-modal MRI. In: Medical Imaging 2020: Computer-Aided Diagnosis; vol. 11314. International Society for Optics and Photonics; 2020, p. 113140S.

[29] Lecouvet, F.E., Pasoglou, V., Van Nieuwenhove, S., Van Haver, T., de Broqueville, Q., Denolin, V., et al. Shortening the acquisition time of whole-body MRI: 3D T1 gradient echo dixon vs fast spin echo for metastatic screening in prostate cancer. European radiology 2020;30:3083–3093.

[30] Takahara, T., Imai, Y., Yamashita, T., Yasuda, S., Nasu, S., Van Cauteren, M.. Diffusion weighted whole body imaging with background body signal suppression (DWIBS): technical improvement using free breathing, STIR and high resolution 3D display. Matrix 2004;160(160):160.

[31] Chiabai, O., Van Nieuwenhove, S., Vekemans, M.C., Tombal, B., Peeters, F., Wuts, J., et al. Whole-body MRI in oncology: can a single anatomic t2 dixon sequence replace the combination of T1 and STIR sequences to detect skeletal metastasis and myeloma? European Radiology 2023;33(1):244–257.

[32] Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006;31(3):1116–1128.

[33] Padhani, A.R., Liu, G., Koh, D.M., Chenevert, T.L., Thoeny, H.C., Takahara, T., et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. Neoplasia 2009;11(2):102–125.

[34] Ceranka, J., Verga, S., Kvasnytsia, M., Lecouvet, F., Michoux, N., de Mey, J., et al. Multi-atlas segmentation of the skeleton from whole-body MRI - impact of iterative background masking. Magnetic resonance in medicine 2020;83(5):1851–1862.

[35] Perona, P., Shiota, T., Malik, J.. Anisotropic diffusion. In: Geometry-driven diffusion in computer vision. Springer; 1994, p. 73–92.

[36] Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., et al. N4ITK: improved N3 bias correction. IEEE transactions on medical imaging 2010;29(6):1310–1320.

[37] Yoo, T.S., Ackerman, M.J., Lorensen, W.E., Schroeder, W., Chalana, V., Aylward, S., et al. Engineering and algorithm design for an image processing API: a technical report on ITK-the insight toolkit. Studies in health technology and informatics 2002;:586–592.

[38] Ceranka, J., Polfliet, M., Lecouvet, F., Michoux, N., de Mey, J., Vandemeulebroucke, J.. Registration strategies for multi-modal whole-body MRI mosaicing. Magnetic resonance in medicine 2018;79(3):1684–1695.

[39] Nyúl, L.G., Udupa, J.K., Zhang, X.. New variants of a method of MRI scale standardization. Medical Imaging, IEEE Transactions on 2000;19(2):143–150.

[40] Blackledge, M.D., Leach, M.O., Collins, D.J., Koh, D.M.. Computed diffusion-weighted MR imaging may improve tumor detection. Radiology 2011;261(2):573–581.

[41] MONAI, . The MONAI consortium - project MONAI. 2020. URL: https://doi.org/10.5281/zenodo.4323059. doi:10.5281/zenodo.4323059.

[42] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 2021;18(2):203–211.

[43] Ma, J.. Segmentation loss odyssey. arXiv preprint arXiv:200513449 2020;.

[44] Wang, L., Lee, C.Y., Tu, Z., Lazebnik, S.. Training deeper convolutional networks with deep supervision. arXiv preprint arXiv:150502496 2015;.

[45] Miller, H.. The FROC curve: A representation of the observer's performance for the method of free response. The Journal of the Acoustical Society of America 1969;46(6B):1473–1476.

[46] Shirokikh, B., Shevtsov, A., Kurmukov, A., Dalechina, A., Krivov, E., Kostjuchenko, V., et al. Universal loss reweighting to balance lesion size inequality in 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2020, p. 523–532.

[47] Shapiro, S.S., Wilk, M.B.. An analysis of variance test for normality (complete samples). Biometrika 1965;52(3/4):591–611.

[48] Snedecor, G.W.C., William, G.. Statistical methods. 1989.

[49] Kruskal, W.H., Wallis, W.A.. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association 1952;47(260):583–621.

[50] Armstrong, R.A.. When to use the bonferroni correction. Ophthalmic Physiol Opt 2014;34(5):502–508.

[51] Larbi, A., Omoumi, P., Pasoglou, V., Michoux, N., Triqueneaux, P., Tombal, B., et al. Comparison of bone lesion distribution between prostate cancer and multiple myeloma with whole-body MRI. Diagnostic and interventional imaging 2019;100(5):295–302.

[52] Perez-Lopez, R., Rodrigues, D.N., Figueiredo, I., Mateo, J., Collins, D.J., Koh, D.M., et al. Multiparametric magnetic resonance imaging of prostate cancer bone disease: correlation with bone biopsy histological and molecular features. Investigative radiology 2018;53(2):96.