Automatic Segmentation of Femoral Tumors by nnU-Net

Oren Rachmil[†], Moran Artzi[⋆,♯], Moshe Iluz[§], Ido Druckmann[§], Zohar Yosibash[†], and Amir Sternheim[‡,♯]

[†]Computational Mechanics & Experimental Biomechanics Lab, School of Mechanical Engineering, The Iby and Aladar Fleischman Faculty of Engineering, Tel-Aviv University, Ramat Aviv, 69978, Israel
[⋆]Sagol Brain Institute, Tel-Aviv Sourasky Medical Center, Tel-Aviv, Israel
[♯]Faculty of Medicine, Tel Aviv University, Ramat Aviv, 69978, Israel
[‡]Dept. of Orthopedic Oncology, Tel-Aviv Sourasky Medical Center, Tel-Aviv, Israel
[§]Dept. of Radiology, Tel-Aviv Sourasky Medical Center, Tel-Aviv, Israel

April 3, 2024

Abstract contains 218 words,
Manuscript excluding abstract, tables, figure, bibliography and appendix contains: 3717 words.

Corresponding Author: Prof. Zohar Yosibash, Head - Computational Mechanics and Experimental Biomechanics lab, School of Mechanical Eng, Faculty of Engineering, Tel Aviv University, Israel, yosibash@tauex.tau.ac.il

Abstract

Background Metastatic femoral tumors may lead to pathological fractures during daily activities. A CT-based finite element analysis of a patient's femurs was shown to assist orthopedic surgeons in making informed decisions about the risk of fracture and the need for a prophylactic fixation. Improving the accuracy of such analyses requires an automatic and accurate segmentation of the tumors and their automatic inclusion in the finite element model. We present herein a deep learning algorithm (nnU-Net) to automatically segment lytic tumors within the femur.

Method: A dataset consisting of fifty CT scans of patients with manually annotated femoral tumors was created. Forty of them, chosen randomly, were used for training the nnU-Net, while the remaining ten CT scans were used for testing. The deep learning model's performance was compared to two experienced radiologists.

Findings: The proposed algorithm outperformed the current state-of-the-art solutions, achieving dice similarity scores of 0.67 and 0.68 on the test data when compared to two experienced radiologists, while the dice similarity score for inter-individual variability between the radiologists was 0.73.

Interpretation: The automatic algorithm may segment lytic femoral tumors in CT scans as accurately as experienced radiologists with similar dice similarity scores. The influence of the realistic tumors inclusion in an autonomous finite element algorithm is presented in [14].

1 Introduction

Approximately 30% to 50% of all cancer cases have the potential to spread and metastasize to the bones, resulting in metastatic bone disease (MBD) that compromises the structural integrity of the skeleton [5]. Such tumors may lead to pathological fractures caused by everyday activities or severe pain that requires medical intervention. Thanks to immuno-oncological and chemotherapy treatment advancements, patients diagnosed with MBD now have extended life expectancies. In the United States, there are an estimated 280,000 reported cases of long bone skeletal metastases annually, with the femur being particularly vulnerable to pathologic fractures due to its weight-bearing nature [16]. Prophylactic fixation is recommended to prevent impending femoral fractures as it presents lower risks and mortality rates compared to surgery performed after a traumatic fracture. The cost of prophylactic femoral fixation is approximately \$78,000 per patient[9], \$21,000 less expensive than treatment following fracture occurrence. At the same time, unnecessary prophylactic surgeries should be avoided. Therefore, ensuring effective management of MBD requires accurate patient-specific assessments to evaluate fracture risk [19].

Clinicians currently rely on the Mirels' criterion [12] or their clinical experience to assess fracture risk in patients with MBD. However, the Mirels' criterion lacks specificity [16], with a sensitivity of 91% and specificity of 35%, resulting in unnecessary internal fixation procedures for approximately two-thirds of patients [8]. In recent years, more accurate methods utilizing computed tomography (CT) have emerged to predict fracture risk, considering both the patient's specific anatomical characteristics and the spatial distribution of material properties in metastatic bones. One notable tool is *Simfini*[1], an autonomous finite element (AFE) software [21] , which provides orthopedic oncologists with an assessment of fracture risk in patients with femoral metastatic tumors. Simfini employs Autonomous Finite Element Analysis (AFE) for a patient-specific evaluation of bone strength. The process involves automatic segmentation of femurs from CT scans using a U-net architecture, automatic generation of meshes, application of boundary conditions based on anatomical landmarks, high-order FE analysis with numerical error control and ultimately gener-

[1]Simfini is a trademark of PerSimiO, Beer-Sheva, Israel

ates an automatic report that delivers a clear assessment of bone fracture risk. An illustration of the Simfini workflow is illustrated in Figure 1.

[Figure 1 about here.]

While this approach has proven successful in predicting the risk of femoral fractures and assisting orthopedic oncology surgeons in determining the necessity of prophylactic fixation, it sometimes encounters inaccuracies because the current algorithm does not identify and does not segment the tumors. Since Simfini is fully-autonomous and because manual segmentation is both time-consuming and tedious, the exact location and dimensions of tumors within the femur are unknown and practically the tumors are assigned a reduced Young modulus because of the low intensity (Hounsfield Unit - HU) value. To address this limitation, a nnU-Net algorithm is beening implement, presented herein, to automatically identify and segment the tumors within the femur.

Various methodologies have been developed to automate tumor segmentation in bones, and a comparison of existing approaches illustrates the need for continued improvement. Yildiz et. al [6] utilized a deep learning network, Mask2Former [1], for automatic segmentation and classification of tumors in the femur. Analyzing 84 femoral CT scans, they achieved an average Dice similarity coefficient (DSC) of 0.56 ± 0.08. It is hoped that by employing a deep learning (DL) architecture based on an U-net one may obtain a higher DSC.

Claudio et. al [18], focusing on primary bone tumors, leveraged the Mask-RCNN-X101 architecture [10] on a dataset of radiographs from 934 patients, attaining a DSC of 0.6 ± 0.34. Though this outcome surpasses the aforementioned work, it relies on X-ray images, identifies only a 2D tumor by a modality incompatible with our focus which is a 3D AFE analysis. Moreau et. al [13] implemented a U-Net-based architecture using the nnU-net framework [7] for bone and bone metastatic lesions segmentation in PET/CT scans of breast cancer patients. By incorporating a bone mask during training, they realized an increased segmentation accuracy, obtaining a DSC of 0.61 ± 0.16. While their utilization of nnU-net architecture and inclusion of bone masks influenced our methodology, the low resolution of PET/CT modality and the because the segmentation is not concentrated on femurs, their outcome limit its applicability to our purposes.

Our study leverages the former studies: we aim at implementing the nnU-Net framework, tailored specifically to segmenting lytic femoral tumors in CT scans. By focusing on this specific area and modality, we target a higher DSC. In addition, we also compare the tumor segmentation accuracy of the DL algorithm to the manual segmentation of two independent experienced musculoskeletal radiologists. This comparative approach, including an evaluation of inter-individual differences between radiologists segmentations, serves to establish a robust benchmark for evaluating the DL performance.

We employ the U-Net [15] as it has emerged as a powerful tool for the accurate and efficient segmentation of medical images. The nnU-Net is an ensemble of the U-Net architecture with an automated pipeline comprising pre-processing, data augmentation and post-processing [7]. Leveraging its capabilities, the nnU-Net can automatically configure a U-Net-based segmentation pipeline tailored to the provided training cases, thus the nnU-Net is used in our study. We investigate its ability to determine an U-net architecture for the segmentation of femoral lytic tumors utilizing a dataset of 50 CT scans of patients that were manually annotated for femoral lytic tumors. The nnU-Net performance is assessed based on DSC on randomly selected 10 CT scans annotated by two experienced radiologists.

2 Methods

CT scans of oncologic patients identified with femoral lytic tumors were collected at Tel-Aviv Sourasky Medical Center (TASMC) after receiving approval from the institutional review boards. It adheres to national and international guidelines, following the Helsinki committee approval number TLV-17-0532.

2.1 Data Collection

A dataset consisting of 50 anonymized CT scans of lower limbs from patients with various types of cancer was considered (overall 100 femurs). These patients exhibited lytic tumors in at least one of their femurs. Manual segmentation of lesion area in the entire dataset, was performed by a trained segmentation specialist from the Computational Mechanics and Experimental Biomechanics Lab at Tel-Aviv University. The manual segmentations were closely supervised by the head of the Orthopedic Oncology Department at TASMC after training by an experienced musculoskeletal radiologists from TASMC. In addition, for the test data, manual segmentation was also performed by two independent experienced musculoskeletal radiologists (Radiologist 1 and 2). The objective of the annotations was to identify the regions containing lytic tumors within the femur. This involved a careful examination of each 2D DICOM slice of the entire CT scan. When a lytic tumor was visually apparent (manifesting as a darker and distorted area within the bone tissue), the corresponding 2D image was annotated by segmenting and masking the tumor region using the ITK-SNAP software [22]. Figure 2 provides an illustrative example of annotating a lytic tumor within the femur using ITK-SNAP.

The dataset was divided randomly in two: 80% for training and validation, and 20% for testing. The training and validation datasets were further subdivided into a stratified 5-fold cross-validation setup. In this setup, each fold consisted of 80% of the cases for training the algorithm, while the remaining 20% were set aside for validation. The utilization of stratified cross-validation ensured that the entire dataset was adequately represented in the evaluation process while maintaining a consistent proportion of different classes within each fold. This approach was chosen due to the relatively small size of the dataset and its imbalanced nature.

[Figure 2 about here.]

2.2 Pre-Processing

As a pre-processing step, the two femurs were segmented from the CT scan separately, resulting in two new CT scans, each containing only the voxels corresponding to one femur. Based on a preliminary study, training a deep learning model to segment lytic tumors in a femur alone yields significantly better results, excluding any irrelevant tissues from the CT scan.

The femur segmentation was based on Yosibash et. al [20], which employs a U-net architecture specifically designed to segment the femur from an abdominal CT scan. This method demonstrated outstanding performance with a DSC of 99.24%. Two masks for the two femurs in the CT scan were generated representing each patient's femur. In addition, two additional text files were generated for each femur containing a list of voxels coordinates and their corresponding Hounsfield units. Each of the two femurs masks was converted to a NIFTI format CT scan to adhere to the specific dataset requirements of the nnU-Net. The NIFTI files, along with their corresponding manually segmented tumors (also in NIFTI format) consisted the datasets for the nnU-net pipeline. Figure 3 provides a visual representation of the pre-processing procedure.

The nnU-Net framework applies additional pre-processing steps. These involve intensity normalization, where the femur voxels in each image are transformed through a process known as Z-score normalization. This normalization entails subtracting the mean and dividing by the standard deviation of the femur voxels. Meanwhile, the non-femur voxels remain unchanged at 0. Furthermore, all samples are cropped to the region encompassing non-zero values, effectively reducing their size and alleviating computational demands. Additionally, the samples are re-sampled to match the median voxel spacing of their respective dataset.

[Figure 3 about here.]

2.3 Network Architecture

The network architecture generated by nnU-Net is illustrated in Figure 4. It follows a similar pattern as the 3D U-Net [3], comprising an encoder and a decoder interconnected through skip connections. nnU-Net relies on standard convolutions for feature extraction, without incorporating additional architectural modifications. Downsampling is achieved using strided convolutions, while upsampling is performed using convolution transpose operations. The input patch size is set to $384 \times 64 \times 96$, with a batch size of 2, allowing the network to process multiple patches simultaneously. The network undergoes a total of five downsampling operations, progressively reducing the spatial dimensions of the feature maps and ultimately resulting in a bottleneck feature map size of $12 \times 4 \times 6$. The initial number of convolutional kernels is set to 32, doubling with each downsampling step until reaching a maximum of 320 kernels. The number of kernels in the decoder mirrors that of the encoder, ensuring symmetry in the network structure. Non-linear activation functions are applied using leaky ReLUs [11], introducing a small negative slope for negative input values to enhance learning capability. To normalize the feature maps and stabilize the learning process, instance normalization [17] is employed.

[Figure 4 about here.]

2.4 Training Details

The 3D U-net architecture, generated by the nnU-net framework, was trained in a five-fold cross-validation on the Tel-Aviv University servers with Docker virtualization services. It utilized an NVIDIA RTX A5000 GPU with 24GB memory, allowing for a large patch size. The dataset had a median image shape of $386 \times 75 \times 108$, and the initial patch size was set to $384 \times 64 \times 96$ while maintaining a batch size of 2. The nnU-net was trained for 1000 epochs, with each epoch consisting of 250 mini-batches. Stochastic gradient descent with Nesterov momentum ($\mu = 0.99$) [2] and an initial learning rate of 0.01 were used for weight learning. The learning rate followed a 'poly' learning rate policy with a decay factor of 0.9. The loss function combined cross-entropy and Dice loss [4] . Extensive data augmentation techniques, including elastic deformations, random scaling, random rotations, and Gamma augmentation, were employed to address the limited size of the database. The input data for the model consisted of femur masks extracted from the CT scan and converted to NIFTI format.

2.5 Inter-individual Difference Evaluation

To evaluate the variability and quality of manual segmentations, two expert musculoskeletal radiologists from the TASMC segmented separately lytic tumors in the ten CT scans from the test set. The segmented tumors area was compared between the segmentation specialist, Radiologists 1 and 2, and the DL automatic segmentation.

2.6 Evaluation Metric: Dice Similarity Coefficient (DSC)

Segmentation performance is commonly assessed by the Dice Similarity Coefficient (DSC), also known as the Dice score. The DSC quantifies the overlap between the tumor segmentation predicted by a model and the ground truth segmentation, with a higher DSC indicative of greater accuracy and precision in tumor segmentation:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

here, "TP" stands for true positives (tumor pixels correctly identified by the model), "FP" stands for false positives (non-tumor pixels that the model incorrectly labels as tumor), and "FN" stands for false negatives (tumor pixels that the model fails to detect). The DSC ranges from 0 to 1, with a score of 1 indicating perfect overlap between the predicted tumor segmentation and ground truth.

An important special case is when there are no tumors in the femur. This situation results in the DSC becoming undefined, often represented as a "NaN" (not a number) due to division by zero. A NaN DSC should not be considered a perfect score (DSC=1), nor should it be completely disregarded. This case is a "true negative" case, indicating a correct non-detection by the DL, which essentially means no segmentation task was required. For the average DSCs we exclude the cases resulting in NaN DSC.

Tumors segmentation similarity (DSC) was compared between the manual segmentation's of the specialist and Radiologists 1+2 (inter-individual difference), and between each manual segmentation and DL automatic segmentation.

3 Results

The details of the fifty CT scans, including the entire femurs, obtained from the TASMC are provided in Appendix A. All patients were diagnosed with different types of cancer and exhibited lytic tumors in at least one of their femurs. The patient population comprises 21 males and 29 females, spanning an age range from 26 to 84, with a median age of 64. In terms of data acquisition, all but two of the CT scans were acquired using a peak kilovoltage (kVp) of 120, with the remaining two scans being acquired at a kVp of 140. The milliampere-seconds (mAs) values for these scans range from 33 to 484. Most CT scans were acquired by Philips scanners except for three scans that were obtained by a GE scanner. The scans exhibited an average spacing of 0.85mm, with slice thickness varying from 0.9 to 3 mm. For training and testing, each femur was processed by the DL model (overall 100 femurs). Each femur contained an average of 377 axial slices for the right femur and 369 slices for the left femur. The regions of interest, identified as lytic tumors, represented approximately 19,057 voxels in each bone. The entire dataset was randomly divided into 80 femurs for training (40 patients) and 20 femurs (10 patients) for testing.

3.1 Inter-individual Difference

Inter-individual differences were evaluate based on the 20 femurs (10 CT scans), and included: differences between radiologists 1 and 2, and differences between the two radiologists and the specialist. These comparisons served to evaluate the labeling variability and training data quality, and to establish the ground-truth to assess the DL performance. In Table 1 we present the DSC for the 20 femurs (10 CT scans) of the two expert radiologists. The average DSC is 0.73 with a standard deviation of 0.08. The DSC of the specialist segmentations compared to experienced radiologists' segmentations is summarized in Table 2. The DSC between the specialist's and the radiologists' segmentations are 0.72 and 0.70. It is important to note that two entries in Table 2 are excluded. In these cases cysts are present which should be segmented for the subsequent AFE analyses addressed in Part II of this paper. However, these should not be identified as lytic tumors. We discuss these cases in the discussion section.

[Table 1 about here.]

[Table 2 about here.]

3.2 The Influence of the Number of Femurs in the Training Set on nn-U-Net's Performance

To investigate the impact of the training dataset size on the nn-U-Net accuracy we incrementally expanded the number of femurs included in the training set (24, 64 and 80 femurs). The training used a 5-fold cross-validation strategy, allowing to assess the model's performance in a robust manner.

In Table 3 we summarise the DSC obtained by 5-fold validation when using 24, 64 and 80 femurs. As expected when increasing the dataset from 24 to 64 femurs the DSC improves, but not so from 64 to 80 femurs. This may be attributed to the introduction of more diverse and complex cases in the expanded training set. Given the substantial variability in tumor characteristics, such as size, shape, and contrast, even a more comprehensive training dataset does not guarantee uniformly improved model performance. Our observations also reinforce the value of using cross-validation for robust performance evaluation, particularly in scenarios of a dataset for which complexity and variability are high.

[Table 3 about here.]

We use in the sequel the U-net trained on the 80 femurs dataset. This U-net has been exposed to additional diversity during training thus may be more robust to a wider range of complex tumor.

3.3 nnU-Net Performance

The performance of the trained nnU-net architecture was evaluated by comparison to the manual segmentation of lytic tumors performed by the segmentation specialist and the two expert radiologists based on the 20 femurs from the test group. The results are summarised in Table 4.

[Table 4 about here.]

An average DSC of 0.69 and a standard deviation of 0.23 was obtained when comparing the automatic DL segmentation to the segmentation specialist. It ranged from 0.00 (for ProspB10 left femur) to 0.88 (in Prosp5010 right femur). Of particular interest are the 'NaN' cases (Prosp1120, Prosp1140, Prosp5010), which suggested that neither the DL model nor the segmentation specialist identified tumors in these samples, a perfect scenario of true negatives.

Similar patterns were obtained when comparing the automatic segmentation to the radiologists: the average DSCs were 0.67 and 0.68. The DSC ranged between 0.15 (in ProspB10) to 0.87 (in Prosp1120) for radiologist 1, whereas for radiologist 2 it ranged from 0.16 (in ProspB10) to 0.86 (in Prosp7020). Similar perfect true negatives were observed ('NaN' cases). However, cases involving cysts were deliberately ignored by the radiologists and thus were excluded from this evaluation.

The left femur of ProspB10 consistently showed a low DSC in all tests due to a subtle, low-contrast tumor near its distal shaft, as identified by the radiologists (our adopted ground truth). The DL model's segmentation erroneously highlighted a substantial portion of the bone marrow as a tumor, likely influenced by its distinct brightness and low contrast in this case. Furthermore, the segmentation specialist's attempt to segment this region was unsuccessful, resulting in a DSC of 0 when compared to the radiologists, as reflected in Table 2. This underscores the challenging nature of certain cases in our model. Figure 5 depicts the tumor in ProspB10's left femur as outlined by the first radiologist, with the second radiologist arriving at a similar segmentation.

[Figure 5 about here.]

The DL model shows a similar DSC when compared to the specialist (average DSC of 0.69) and when compared to the radiologists (average DSCs of 0.67 and 0.68). The similarity in the DSC standard deviation indicates also a consistent level of variability. This DSC is very close to the inter-individual average DSC of 0.73. However, a larger standard deviation is observed in

DL-to-radiologist comparisons (0.23 compared to 0.08) indicating cases, such as ProspB10's left femur, where the DL model's performance falls short.

4 Discussion

The nnU-Net framework was utilized to generate a U-net architecture for lytic femoral tumor segmentation showing good agreement with human expert annotations. The DL performance marginally trailed the inter-individual agreement between two expert radiologists.

The annotations for the nnU-Net training were performed by the segmentation specialist. An inter-individual difference between the specialist and two experienced radiologists (DSC of 0.72,0.70) was comparable to the inter-individual difference between the expert radiologists themselves (0.73) when excluding the two cyst cases from the 20 test femurs. These isolated lesions located at the femur's head, shown in Figure 6, are essential to the finite element analysis and their segmentation is important.

Comparing the DSC scores achieved by the DL model with those from the segmentation specialist and the two radiologists, we found a generally consistent level of agreement across cases. Nevertheless, in some cases the DSC comparison shows some variability resulting from the intrinsic heterogeneity of lytic tumors, in terms of their size, shape, density, and location within the femur. While radiologists rely on years of experience and clinical knowledge, the nnU-net relies on learned patterns from the training data, which may not always capture the nuanced judgment of human experts and can lead to subtle differences in the delineation of tumor boundaries. These differences are particularly pronounced in challenging cases, such as the subtle, low-contrast tumor in ProspB10's left femur. The presence of several 'NaN' scores in Table 4 indicates true negative scenarios where the absence of tumor detection by the DL model was in complete agreement with the human annotators.

[Figure 6 about here.]

The marginal decrease in mean DSC and a slight increase in variability for the training set of 80 femurs compared to the 64 femurs suggest that increasing the training dataset may not always guarantee enhanced performance. This, coupled with the inter-individual difference between expert radiologists, suggests that femoral tumor segmentation may not achieve a high DSC.

The automatic segmentation of femoral lytic tumors, which usually requires an experienced radiologist's insight, was shown to perform as well on average. To the best of our knowledge, the method proposed in this paper is the state-of-the-art in segmenting lytic femoral tumors in CT scans. The automatic DL model is integrated into an autonomous finite element pipeline, described in a follow-on paper, aimed at determining the risk of pathological femoral fractures and thus assisting orthopedic oncologists in their decision on the need of a prophylactic surgery.

4.1 Limitations

Several limitations in this study could be further investigated in a follow-up research:

- The training set consists of 80 femurs. Enlarging the training data set and the variety of CT scanner manufacturers may increase the accuracy of the DL model.
- Only two experienced radiologists annotated the test set and only 20 femurs were considered for the estimation of the inter-individual difference. Furthermore, the radiologists employed the ITK-SNAP software for their segmentation, which is not their routine segmentation tool in daily practice. A larger cohort of radiologists and a larger testing dataset is warranted.
- The correlation of the Dice score with the size of the tumor must be further investigated. Small tumors are challenging for automatic segmentation, so their detection becomes difficult for DL models. Therefore undetected small tumors can significantly reduce the DSC. Hence, the Dice score, despite its frequent usage, may not be a good measure of segmentation performance.

Acknowledgements

OR, ZY and AS acknowledge the support of this research by the Israel Ministry of Science and Technology under the Tenth Call of Israel-Italy Scientific collaboration.

References

[1] Cheng B, Misra I, Schwing AG, Kirillov A, and Girdhar R. Masked attention mask transformer for universal image segmentation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[2] Aleksandar Botev, Guy Lever, and David Barber. Nesterov's accelerated gradient and momentum as approximations to regularised update descent. *2017 International Joint Conference on Neural Networks (IJCNN)*, page 1899:1903, 2011.

[3] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432, 2016.

[4] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. *Deep Learning and Data Labeling for Medical Applications*, page 1899:1903, Springer, 2016.

[5] CSB Galasko. Monitoring of bone metastases. *Schweiz Med Wochenschr*, 111(49):1873–1875, 1981.

[6] Yildiz Potter I, Yeritsyan D, Mahar S, Wu J, Nazarian A, and Vaziri A. Vaziri A. Automated bone tumor segmentation and classification as benign or malignant using computed tomographic imaging. *J Digit Imaging*, 2023 Jan 10.

[7] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[8] MU Jawad and SP Scully. In brief: classifications in brief: Mirels' classification: metastatic disease in long bones and impending pathologic fracture. *Clin Orthop Relat Res*, 468(10):2825–2827, 2010.

[9] S Li, Y Peng, ED Weinhandl, et al. Estimated number of prevalent cases of metastatic bone disease in the us adult population. *Clin Epidemiol*, 4:87–93, 2012.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. Posted May 1, 2014. Accessed August 9, 2021.

[11] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. *Proceedings of the International Conference on Machine Learning (ICML)*, 30(1):3, 2013.

[12] H Mirels. Metastatic disease in long bones. a proposed scoring system for diagnosing impending pathologic fractures. *Clin Orthop Relat Res*, 249:256–264, 1989.

[13] N Moreau, C Rousseau, C Fourcade, G Santini, L Ferrer, M Lacombe, C Guillerminet, M Campone, M Colombie, M Rubeaux, and N Normand. Deep learning approaches for bone and bone lesion segmentation on 18 FDG PET/CT imaging in the context of metastatic breast cancer. *EMBC - Engineering in Medicine and Biology Conference*, 2020. DOI:10.1109/EMBC44109.2020.9175904.

[14] O Rachmil, K Myers, O Merose, A Sternheim, and Z Yosibash. The influence of femoral lytic tumors segmentation on autonomous finite element analysis. *Clinical Biomechanics*, 112:106192, 2024.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *MICCAI. Springer*, pages 234–241, 2015.

[16] A Sternheim, F Traub, N Trabelsi, S Dadia, Y Gortzak, N Snir, M Gorfine, and Z Yosibash. When and where do patients with bone metastases actually break their femurs? *Bone Joint J*, 102(5):638–645, May 2020.

[17] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[18] CE von Schacky, NJ Wilhelm, VS Schäfer, Y Leonhardt, FG Gassert, S Foreman, FT Gassert, M Jung, PM Jungmann, MF Russe, C Mogler, C Knebel, R von Eisenhart-Rothe, M R Makowski, K Woertler, R Burgkart, and AS Gersing. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology*, 301(2):398–406, 2021.

[19] JJ Willeumier, MAJ van de Sande, RJP van der Wal, and PDS Dijkstra. Trends in the surgical treatment of pathological fractures of the long bones: based on a questionnaire among members of the Dutch Orthopaedic Society and the European Musculo-Skeletal Oncology Society (EMSOS). *Bone Joint J*, 100(10):1392–1398, 2018.

[20] Z. Yosibash, Y. Katz, N. Trabelsi, and A. Sternheim. Femurs segmentation by machine learning from CT scans combined with autonomous finite elements in orthopedic and endocrinology applications. *Comp. Math. App.*, 2023. In Print.

[21] Z. Yosibash, K. Myers, N. Trabelsi, and A. Sternheim. Autonomous FEs (AFE) - A stride toward personalized medicine. *Comp. Math. App.*, 80:2417–2432, 2020.

[22] Paul A. Yushkevich, Yang Gao, and Guido Gerig. ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3342–3345, 2016.

List of Figures

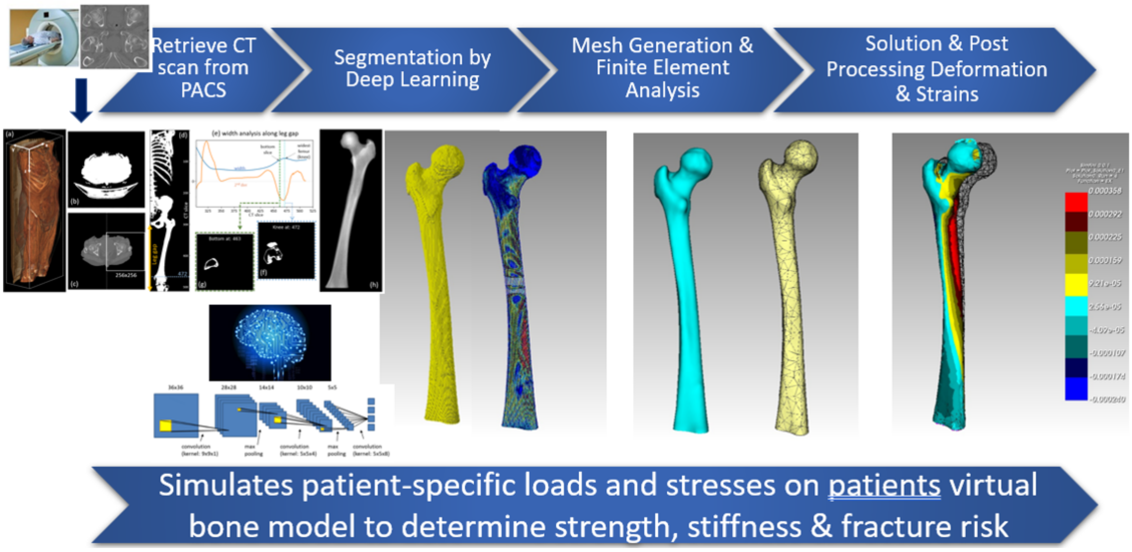

Figure 1 Simfini AFE pipeline, taken from [16]

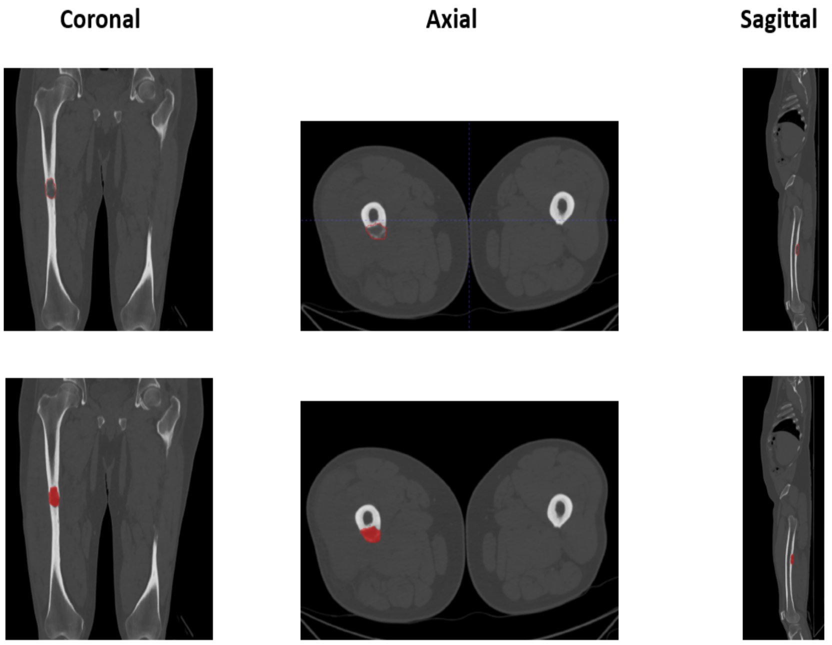

Figure 2 Segmentation of a lytic tumor within the femur viewed from multiple angles performed with ITK-SNAP software.

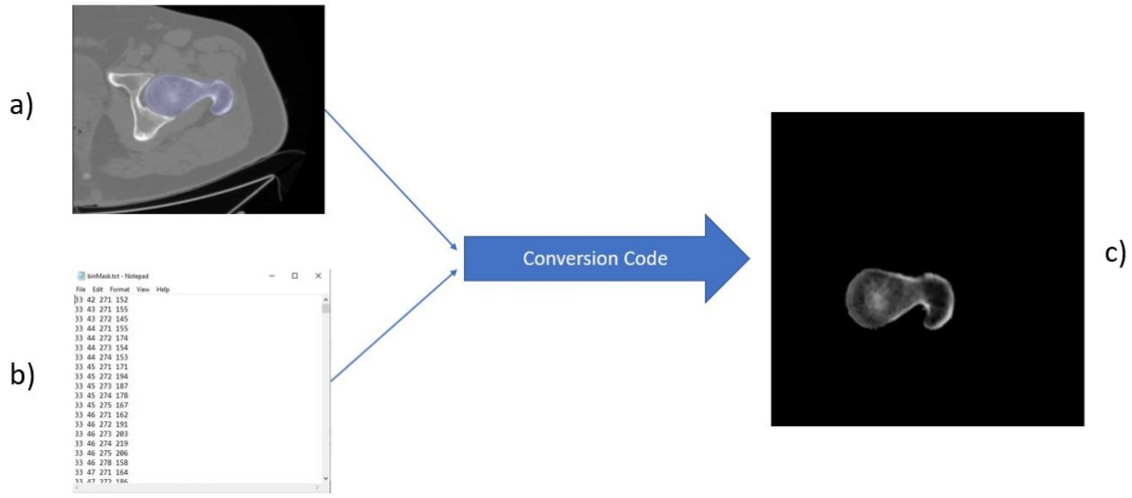

Figure 3 Segmented femur in NIFTI format viewed by ITK-SNAP. **a.** The original CT scan with the femur highlighted in blue, **b.** list of femur voxels coordinates saved in a text file, **c.** the segmented femur.

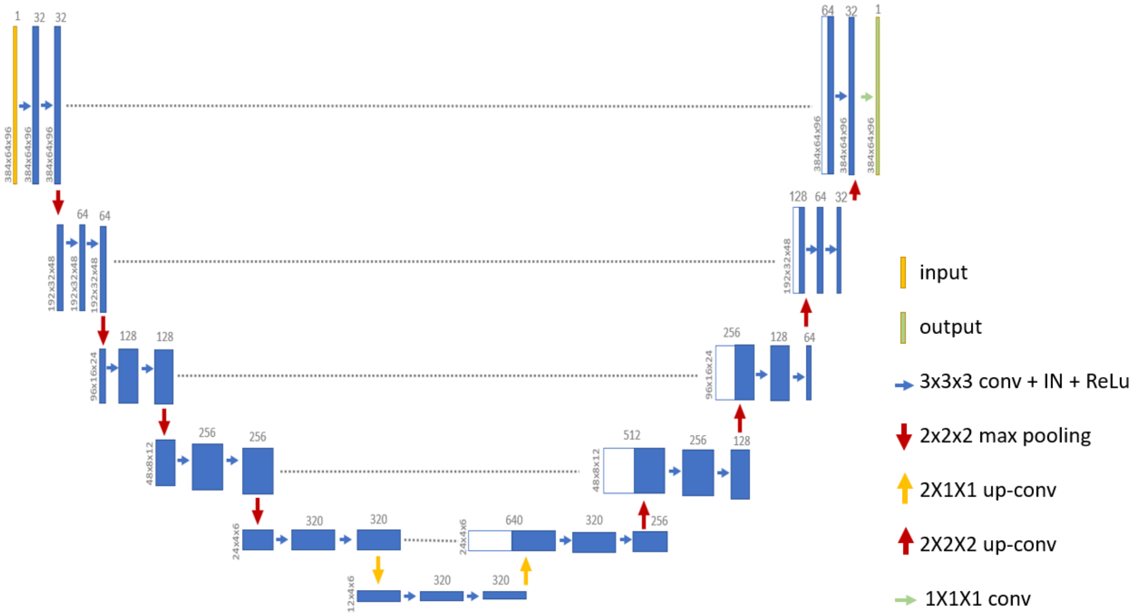

Figure 4 nnU-Net architecture for the training set. It follows a 3D U-Net pattern with an encoder, decoder, and skip connections. The input patch size is $384 \times 64 \times 96$, and the network includes five downsampling operations.

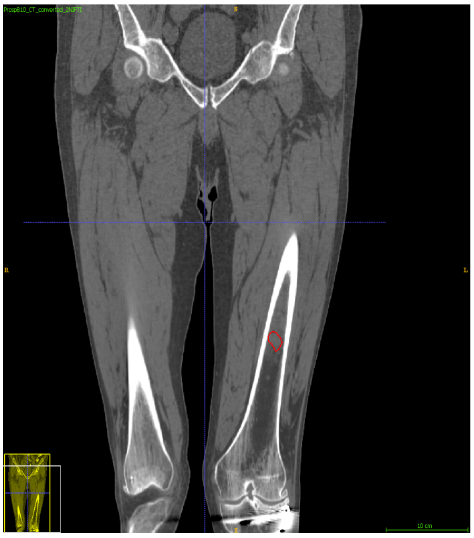

Figure 5 Abdominal CT scan of patient ProspB10. Highlighted in the left femur (viewed from the right) is a barely discernible tumor, encircled by the red polygon for clarity.

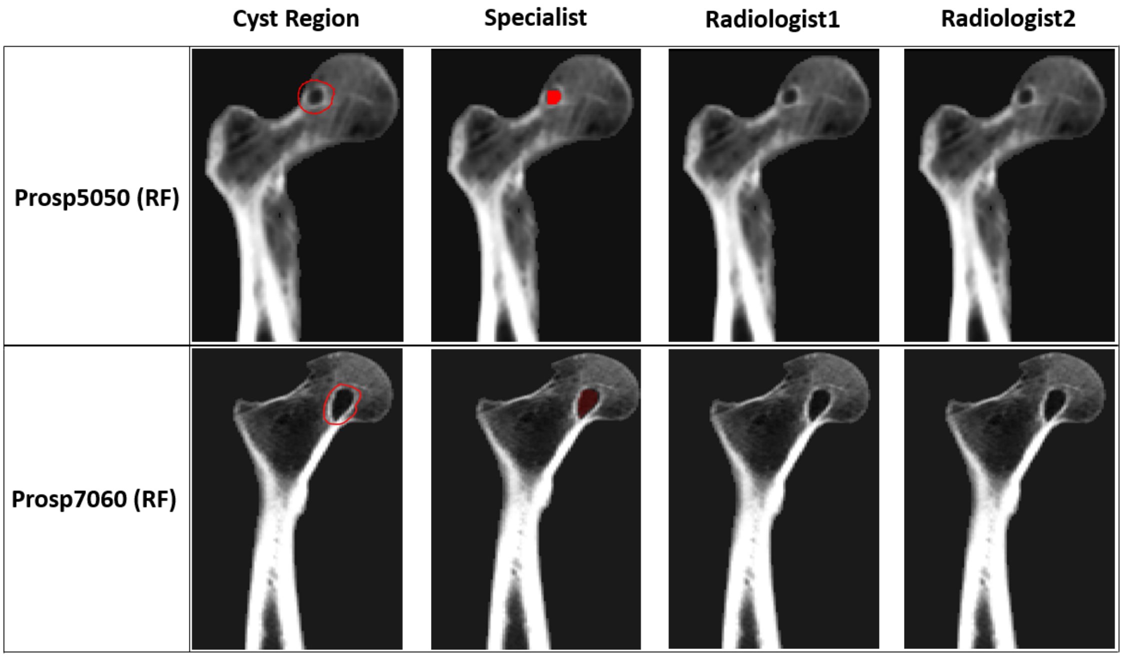

Figure 6 Cyst segmentation by the specialist and two expert radiologists in the two excluded test cases - Prosp5050 (right femur) and Prosp7060 (right femur).

List of Tables

Radiologist 1 vs Radiologist 2

#Case	Left Femur	Right Femur
Prosp1120	NaN	0.83
Prosp1140	NaN	0.64
Prosp1190	0.79	0.77
Prosp5010	NaN	0.68
Prosp5050	0.65	NaN
Prosp5060	0.80	NaN
Prosp7020	0.80	0.76
Prosp7060	0.77	NaN
ProspB10	0.70	0.70
ProspD100	NaN	0.55

Average DSC: 0.73
Standard Deviation: 0.08

Table 1 Comparison of lytic tumors segmentation similarity (DSC) between Radiologist 1 and Radiologist 2.

Specialist vs Radiologist 1

#Case	Left Femur	Right Femur
Prosp1120	NaN	0.87
Prosp1140	NaN	0.76
Prosp1190	0.83	0.80
Prosp5010	NaN	0.73
Prosp5050	0.53	**excluded**
Prosp5060	0.81	NaN
Prosp7020	0.80	0.80
Prosp7060	0.83	**excluded**
ProspB10	0.00	0.85
ProspD100	NaN	0.77

Average Dice Score: 0.72
Standard Deviation: 0.23

Specialist vs Radiologist 2

#Case	Left Femur	Right Femur
Prosp1120	NaN	0.80
Prosp1140	NaN	0.65
Prosp1190	0.86	0.79
Prosp5010	NaN	0.67
Prosp5050	0.72	**excluded**
Prosp5060	0.86	NaN
Prosp7020	0.83	0.86
Prosp7060	0.84	**excluded**
ProspB10	0.00	0.77
ProspD100	NaN	0.51

Average Dice Score: 0.70
Standard Deviation: 0.23

Table 2 Comparison of lytic tumors segmentation similarity (DSC) between the specialist and Radiologists 1 and 2 segmentations. Boldface numbers denote femurs with a cyst. The DSCs between the two experienced radiologists are similar to those obtained by the specialist.

Fold Number	24 Femurs	64 Femurs	80 Femurs
0	0.44	0.53	0.63
1	0.55	0.69	0.45
2	0.73	0.69	0.56
3	0.39	0.66	0.68
4	0.54	0.63	0.73
Mean	0.53	0.64	0.61
Std Dev	0.13	0.07	0.10

Table 3 Training DSC from a 5-fold cross-validation by the nn-Unet. Each column represents a different training set size. The 'Mean' and 'Std Dev' are the average and standard deviation of the DSC across the five folds.

Automatic vs Specialist

#Case	LF	RF
Prosp1120	NaN	0.85
Prosp1140	NaN	0.84
Prosp1190	0.84	0.67
Prosp5010	NaN	0.88
Prosp5050	0.71	0.83
Prosp5060	0.85	NaN
Prosp7020	0.65	0.75
Prosp7060	0.87	0.46
ProspB10	0.00	0.62
ProspD100	NaN	0.52

Average Dice Score: 0.69
Standard Deviation: 0.23

Automatic vs Radiologist 1

#Case	LF	RF
Prosp1120	NaN	0.87
Prosp1140	NaN	0.76
Prosp1190	0.77	0.80
Prosp5010	NaN	0.73
Prosp5050	0.59	excluded
Prosp5060	0.78	NaN
Prosp7020	0.55	0.80
Prosp7060	0.85	excluded
ProspB10	0.15	0.80
ProspD100	NaN	0.22

Average Dice Score: 0.67
Standard Deviation: 0.20

Automatic vs Radiologist 2

#Case	LF	RF
Prosp1120	NaN	0.80
Prosp1140	NaN	0.65
Prosp1190	0.81	0.79
Prosp5010	NaN	0.67
Prosp5050	0.72	excluded
Prosp5060	0.79	NaN
Prosp7020	0.60	0.86
Prosp7060	0.78	excluded
ProspB10	0.16	0.77
ProspD100	NaN	0.39

Average Dice Score: 0.68
Standard Deviation: 0.20

Table 4 DSC Comparison for the segmentation of lytic femoral tumors: Automatic vs Specialist and Radiologists 1 and 2.