# COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGI-NEERING: IMAGING & VISUALIZATION

# Fast and Robust Femur Segmentation from Computed Tomography Images for Patient-Specific Hip Fracture Risk Screening

Pall Asgeir Bjornsson<sup>a</sup>, Alexander Baker<sup>b</sup>, Ingmar Fleps<sup>b</sup>, Yves Pauchard<sup>c</sup>, Halldor Palsson<sup>d</sup>, Stephen J. Ferguson<sup>b</sup>, Sigurdur Sigurdsson<sup>e</sup>, Vilmundur Gudnason<sup>e,f</sup>, Benedikt Helgason<sup>b</sup>, and Lotta Maria Ellingsen<sup>a,g</sup>

<sup>a</sup>The Dept. of Electrical and Computer Engineering, The University of Iceland, Reykjavik, Iceland; <sup>b</sup>The Institute for Biomechanics, ETH Zurich, Zurich, Switzerland; <sup>c</sup>The University of Calgary, McCaig Institute for Bone and Joint Health, Calgary, AB, Canada; <sup>d</sup>The Dept. of Industrial Engineering, Mechanical Engineering, and Computer Science, The University of Iceland, Reykjavik, Iceland; <sup>e</sup>The Icelandic Heart Association, Kopavogur, Iceland; <sup>f</sup>The Dept. of Medicine, The University of Iceland, Reykjavik, Iceland; <sup>g</sup>The Dept. of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218, USA

#### ARTICLE HISTORY

Compiled April 21, 2022

#### ABSTRACT

Osteoporosis is a common bone disease that increases the risk of bone fracture. Hip-fracture risk screening methods based on finite element analysis depend on segmented computed tomography (CT) images; however, current femur segmentation methods require manual delineations of large data sets. Here we propose a deep neural network for fully automated, accurate, and fast segmentation of the proximal femur from CT. Evaluation on a set of 1147 proximal femurs with ground truth segmentations demonstrates that our method is apt for hip-fracture risk screening, bringing us one step closer to a clinically viable option for screening at-risk patients for hip-fracture susceptibility.

#### **KEYWORDS**

Computed Tomography; Femur; Segmentation; Convolutional neural networks; Osteoporosis

# 1. Introduction

According to the United Nations, more than 40% of the population of some developed countries will be above the age of 60 by the year 2050 (UN 2015). This raises concerns about the burden placed on health care systems, since aging societies are associated with a higher prevalence of chronic diseases. Policy-makers are thus forced to reconsider the status quo of health care systems, moving away from face-to-face consultation-based care towards a decentralized community or home-based care, as well as transitioning from focusing on treatment to focusing on prevention.

One of the prevalent chronic diseases suffered by elderly populations is osteoporosis - a bone disease characterized by low bone mass and structural deterioration of bone tissue, leading to bone fragility and an increased risk of fracture. There is sufficient evidence that the majority of hip fractures are the result of a low trauma fall (Hayes et al. 1996; Parkkari et al. 1999). The fracture risk increases with age and, compared to other fracture types, hip fractures are associated with the most dire socioeconomic consequences. The most abysmal, and perhaps surprising, statistic is that 11-23% of individuals will be deceased six months after incurring the fracture, increasing to 22-29% after one year has passed since the incident (Haleem et al. 2008).

#### 1.1. Current Standard in Screening for Hip Fracture Risk

The present clinical "gold standard" to diagnose osteoporosis is the areal bone mineral density (aBMD) derived from dual-energy X-ray absorptiometry (DXA). However, a shortcoming of this method is that even though low aBMD scores are associated with population-based fracture risk, between 36-72% of incident fractures are sustained by individuals who do not have osteoporosis (Stone et al. 2003; Schuit et al. 2004; Wainwright et al. 2005). Moreover, the aBMD lacks specificity when stratifying risk considering the fact that the majority of subjects with osteoporosis do not incur hip fractures in their lifetime.

#### 1.2. Finite Element Analysis

In order to improve both the specificity and sensitivity of hip fracture screening, X-ray computed tomography (CT) image-based, subject-specific finite element (FE) models of the proximal femur have garnered significant attention and shown promise as a means to overcome the limitations of assessing hip fracture risk using aBMD. The motivation for this application is to incorporate it into a clinical screening tool that uses FE analysis for hip fracture risk prediction. The widespread use of such tools has the potential to dramatically reduce the economic toll of hip fractures on our healthcare systems, as well as mitigate the potentially devastating consequences for patients. Thus far, several hindrances have impeded clinical translation of FE analysis for hip fracture risk screening: the cost of hiring trained engineers to carry out simulations instead of the clinical staff, the ambiguous accuracy of these methods for fracture prediction, and the health risk of X-ray exposure caused by the CT scanner. In order to bring clinical applications to fruition, a robust and automated workflow for constructing the FE model and subsequent analysis is imperative. Fleps et al. (Fleps et al. 2021) demonstrated that femoral strength based on finite element analysis (FEA) can improve hip fracture risk assessment and we employed the same FE pipeline in this work. The workflow pipeline entails the segmentation of the CT image data, generating an FE mesh, applying heterogeneous gray level based material properties to the FEs, applying boundary conditions, solving FE equations and processing the results (Pauchard et al. 2016). The aim of this work is to develop a fully automated segmentation of the proximal femur from a CT image, without the need for any manual intervention during postprocessing.

# 1.3. Related Work

Bone segmentation of CT images is an elusive problem for several reasons. Firstly, there is an overlap of the Hounsfield units  $(HU)^1$  of the bones and surrounding tissue, rendering it impossible to segment solely based on intensity value. Moreover, bones themselves do not have uniform densities, nor do certain bone diseases affect all parts of the bone in the same manner. Adjacent bones pose an additional problem when the joint space approaches the resolution of clinical CT data, which is often the case for elderly subjects, and can result in poor segmentations. Hence, a method is needed to detect and connect thin and diffusive bone structure boundaries to obtain acceptable segmentations. Figure 1 displays an example of a hip joint from the data set at hand, which can prove challenging to segment if the boundary between the femoral head and acetabulum is unclear.





Promising methods for segmenting the proximal femur from CT images that have gained traction as of late are statistical shape models (SSM) (Chang et al. 2019; Younes et al. 2019), multi-atlas segmentation (Wang and Yushkevich 2013; Chengwen et al. 2015) and graph-cut segmentation (Pauchard et al. 2016). The two former methods, however, require a database of gold standard segmentations, while the latter method does not necessitate such prior knowledge. Lastly, the implementation of 3-dimensional (3D) convolutional neural networks (CNNs) to address the problem of femur segmentation is one of the most recent developments (Chen et al. 2019; Zhao et al. 2020) and has become the method of choice in biomedical image analysis.

One of the most successful previous methods for segmenting the proximal femur is the aforementioned graph-cut method, carried out by Pauchard et al. (Pauchard et al. 2016). In short, this method separates the background class from the target

 $<sup>^{1}</sup>$ The Hounsfield unit is a relative quantitative measurement of radio density used by radiologists in the interpretation of CT images.

object by finding the global minimum of a cost function. If differences in intensity are large (i.e., at object boundaries) with respect to  $\sigma^2$  (variance of homogenous regions in the image), then the cost of cutting an edge is low. On the other hand, if differences are small in comparison to  $\sigma^2$ , then the cost is high. They reported a mean Dice Similarity Coefficient (DSC) (Dice 1945) of  $0.973 \pm 0.005$ , while the mean Hausdorff distance (HD) between manual segmentations and interactive graph-cut segmentations was  $3.75 \pm 1.26$ mm. This method, however, suffers from its only partially autonomous nature: when producing the segmentation predictions, manual input is required by the user to initiate the graph-cut segmentation process, which is a key limitation of this method.

Zhao et al. (Zhao et al. 2020) proposed an automated, patch-based 3D v-net architecture (Milletari et al. 2016) (employing the Dice loss function) on a cohort that comprised 397 quantitative computed tomography (QCT) scans, of which only 10% was used to evaluate the model. This reliance on such a large training/validation set, which is not always available, is a key limitation of this model. The method struggled to segment the femur around the most dynamic sections (i.e., the femoral head), resulting in some unacceptable segmentations. Nevertheless, the authors reported a mean DSC of  $0.9815 \pm 0.0009$  and, for a subject with 60 QCT slices, a segmentation time of 15s.

Another automated 3D CNN method conducted by Chen et al. (Chen et al. 2019) is based on the u-net architecture (Ronneberger et al. 2015) and employs both the Dice loss and the Jaccard loss functions to segment the entire femur. An edge detection task was embedded into a fully convolutional network (FCN) to address the problems of diffusive joint spaces and weak femur boundaries. The method, however, shares the same limiting factor as the previously discussed method (Zhao et al. 2020) in that it requires a large training set (120 samples) which, for many biomedical segmentation tasks, is not a viable option. The authors of this study reported a mean DSC of  $0.9688 \pm 0.0095$  on an evaluation set of 30 CT images. Table 1 compares the methods mentioned in this section.

The key limitation to previous deep learning methods has been the reliance on a vast training set, which requires an equally large set of ground truth segmentations. The proposed method necessitates far fewer ground truth segmentations than these methods. A preliminary version of our model was reported in conference form (Bjornsson et al. 2021); here the method has been validated on a significantly larger test set and compared with a state-of-the-art segmentation method. Moreover, we demonstrate its use in our patient-specific screening method, where the FEA-derived femoral strength based on our method was compared to that based on ground truth segmentations, further validating the method on the end-product.

# 1.4. Rationale for Deep Learning Approach

Current femur segmentation methods mostly require a "user-in-the-loop" paradigm in order to manually correct segmentations and produce acceptable masks for FE modeling. This lack of robustness is costly in terms of time and the need for highly trained specialists to manually correct the segmentation predictions. Consequently, these methods cannot process larger cohorts to the same degree as a fully automated one, rendering them impractical for clinical application. The justification for using a DNN is almost entirely a byproduct of the u-net. Since large data sets containing CT images from a particular scanner are hard to come by, neural networks were not Table 1.: A comparison between different femur segmentation methods. Here,  $\sigma$  denotes standard deviation. The time required to segment a femur volume should be interpreted with caution considering the varying workstations, resolution, and number of slices per scan.

Authors	Model	Training set (CNNs only)	Test set	Slice thick- ness [mm]	$DSC \pm \sigma$	Time p. fe- mur (s)
Pauchard et al. Pauchard et al. (2016)	Graph-cut	N/A	48 CT	1	$0.973 \pm 0.005$	120-300
Chu et al. Chengwen et al. (2015)	Multi-atlas	N/A	30 CT	1	$0.979 \pm 0.029$	275
Younes et al. Younes et al. (2019)	SSM	N/A	8 CT	-	$0.89 \pm 0.01$	180
Chang et al. Chang et al. (2019)	Conditional random field	N/A	60 CT	0.45-1.2	$0.973 \pm 0.0095$	-
Chen et al. Chen et al. (2019)	3D CNN	150 CT	30 CT	1.32-1.85	$\begin{array}{rrr} 0.9688 & \pm \\ 0.0095 & \end{array}$	56
Žhao et al. Zhao et al. (2020)	3D CNN	357 QCT	40 QCT	3	$\begin{array}{rrr} 0.9815 & \pm \\ 0.0009 & \end{array}$	15

viewed as a particularly attractive alternative for application in biomedical imaging. However, the u-net architecture proposed by Ronneberger et al. (Ronneberger et al. 2015) demonstrated fast and precise segmentation without the need for a large data set. DNNs have, as a result, become the state-of-the-art method for segmentation in biomedical image classification (Shao et al. 2019; Huo et al. 2019). Instead of requiring minutes to generate a segmentation prediction, CNNs can produce an output in the matter of seconds.

# 1.5. Contributions of our Work

In this paper, we propose a robust, fully automated, and fast segmentation of the proximal femur from CT images. The most salient contributions of our work to the field of biomedical image segmentation of the proximal femur are the following: First, our method takes a human-out-of-the-loop approach rendering the arduous and time-consuming task of making ad hoc corrections unnecessary; second, the model is highly robust, and hence, opens up the possibility of e.g. low-cost opportunistic screening for hip fracture risk based on existing CT data; and third, the processing time (in a matter of seconds) is well within reasonable bounds for clinical implementation.

### 2. Materials and Methods

### 2.1. The AGES-RS Cohort

The Icelandic Heart Association (IHA) provided us with CT scans from the Age, Gene/Environment Susceptibility-Reykjavik Study (AGES-RS) (Harris et al. 2007): a cohort that consists of both men and women born between 1907 and 1935, monitored in Iceland by the IHA since 1967. This unique database of high-quality CT images contains roughly 4800 density calibrated CT scans of the proximal femur at baseline and 3300 scans of the same individuals acquired at a five-year follow-up. The resolution of each scan is  $512 \times 512$  voxels with  $0.977 \times 0.977 \times 1 \text{ mm}^3$  voxel size and the number of slices ranges from 88 to 178. Our model was evaluated using two subsets of the AGES data set. The first subset (Sample I<sup>2</sup>) comprises 48 "gold standard" manually delineated proximal femur segmentations from 24 CT images. The second subset within the AGES data set (Sample II<sup>3</sup>) consists of 1207 manually delineated segmentations, generated with a semi-automated delineation protocol, that served as ground truth annotations. The proposed segmentation model was trained w.r.t. 60 of these ground truth segmentations and evaluated on the remaining 1147.

## 2.2. Validation and Loss Function

Since the femur only makes up a small part of each slice, there is a class imbalance problem that must be addressed to avoid the more prevalent class from dominating. The learning process tends to get trapped in a local minima of the loss function and yields a network whose segmentation predictions are heavily biased towards the background class. To combat this problem, the DSC (Dice 1945), which effectively renders the relative spatial areas of each class irrelevant, was implemented. The DSC measures spatial overlap between segmentations and is given by the following equation:

$$DSC = \frac{2\sum_{i}^{N} p_{i}g_{i}}{\sum_{i}^{N} p_{i}^{2} + \sum_{i}^{N} g_{i}^{2}}.$$
 (1)

In this equation, N is the number of voxels of the predicted binary segmentation volume  $p_i \in P$  and the ground truth binary volume  $g_i \in G$  (Milletari et al. 2016). The values of the DSC are restricted to the range [0, 1], where DSC = 0 indicates total misclassification and DSC = 1 indicates perfect classification. In order to formulate a loss function, the Dice loss function is defined as 1 - DSC.

# 2.3. The Proposed Segmentation Pipeline

The implemented fully automated proximal femur segmentation pipeline is illustrated in Figure 2 and consists of the following components:

- A training/validation set of 30 3D CT images (i.e., 60 proximal femurs) from Sample II of the AGES cohort (Harris et al. 2007)
- Normalization
- On-the-fly data augmentation

<sup>&</sup>lt;sup>2</sup>This sample set was used by Pauchard et al. (Pauchard et al. 2016) to evaluate their model.

 $<sup>^{3}\</sup>mathrm{This}$  sub-cohort from AGES-RS, including all fracture cases, was used by Enns-Bray et al. (Enns-Bray et al. 2019).



Figure 2.: A flowchart showing the workflow of our proposed method.

- Patch-based 3D u-net
- Training using the Dice loss function for a pre-defined number of epochs
- Prediction on data from the validation set to gauge the performance on unseen data and to aid in hyperparameter tuning

When the model is applied to evaluation data, a postprocessing step concatenates black (background class) voxels to the output masks so that their dimensions agree with the original CT scans.

# 2.4. Preprocessing

Each of the 30 CT images was cut in half, splitting the left and right proximal femurs into separate images. The resulting CT scans that included the left proximal femur were then mirrored to the right side. The training/validation set effectively became 60 images of the right hip/upper leg with in-plane resolution  $512 \times 256$  voxels and 98-148 slices. The CT images were normalized such that all intensity values were linearly shifted and scaled from HU to the range [0, 1]. Min-max normalization has the advantage over z-score normalization of preserving the scale of the data. Models using both normalization methods were implemented, however, there was no discernible difference between the two.

# 2.5. Data Augmentation and Regularization

Since obtaining manually segmented images is laborious and slow, the use of data augmentation is crucial for maximizing the efficiency of the training set. Data augmentation is used to teach the neural network invariance and robustness properties when a limited data set is available, thus artificially expanding the initial training set to avoid overfitting. These deformations can be simulated efficiently and aid the model in learning invariance between samples (Ronneberger et al. 2015). Here we applied both linear-spatial and intensity transformations (i.e., scaling, rotation and brightness) as well as elastic deformation (Figure 3) to simulate the variability between patients' scans using the Batchgenerators package (Isensee et al. 2020) within the Medical Im-

Table 2.: The data augmentation parameter ranges for the proposed model. Here,  $\alpha$  denotes the scaling factor (controls the deformation intensity) and  $\sigma$  denotes the smoothing factor (controls the displacement field smoothing) for the elastic deformation.

	Brightness	Rotation (X,Y,Z)	Scaling	Elastic deformation
Parameter range	(0.75, 1.25)	$(-3^\circ, 3^\circ)$	(0.95, 1.05)	$\begin{aligned} \alpha &= (0, 100) \\ \sigma &= (9, 13) \end{aligned}$
0 -	•		0	
100 -	100		100 -	100 -
300 -	200		200 -	300 -
400	400		400	400
500 - 0	100 200 500	0 100 200	500 - <b>100</b> - 20	500 - <u>300 - 200</u>
	(a) Origin	nal	(b	) Deformed

Figure 3.: The effect of elastic deformation on a 2D CT image and its mask.

age Segmentation with Convolutional Neural Networks (MIScnn) framework<sup>4</sup>. The exact parameter ranges implemented for our proposed model are given in Table 2.

Data augmentation, with random transformation parameters from the pre-defined ranges was performed on-the-fly for each image before it was forwarded into the neural network. Each of the data augmentation transformations had a 35% likelihood of being applied to the image at hand, allowing the model to encounter a diverse set of images, thereby decreasing redundancy. For the proposed method, on-the-fly<sup>5</sup> data augmentation, in concert with parameter sharing (LeCun et al. 1990) and batch normalization (BN) (Ioffe and Szegedy 2015), rendered the use of explicit regularization techniques unnecessary and even counterproductive.

# 2.6. Model Architecture

DNNs have prevailed as the state-of-the-art learning models for biomedical image segmentation, most notably the renowned u-net (Ronneberger et al. 2015). This impactful and elegant network architecture, based on the FCN, addresses two main issues: namely, the ability to train a model from a very small data set and the ability to produce precise segmentations despite the former. A schematic of the proposed architecture is shown in Figure 4. The u-net derives its name from the u-shape of the model

 $<sup>^{4}</sup>$ MIScnn is an open-source Python library and intuitive API for medical image segmentation pipelines (Müller and Kramer 2019).

 $<sup>^5 {\</sup>rm On-the-fly}$  data augmentation eliminates the need for excessive storage of augmented images by performing the augmentation prior to each optimization iteration.



Figure 4.: A schematic of our proposed 3D u-net. The bold numbers at the corners represent the number of feature maps (channels) per layer. Here, convolution is abbreviated as conv and rectified linear unit as ReLU.

architecture, consisting of a contracting (downsampling) path and an expanding (upsampling) path. The contracting path is the encoder and captures the context in the CT image by way of stacked convolutional and max pooling layers. The expanding path, on the other hand, is the decoder and allows for precise localization with the use of transposed convolutions. In the final layer of the network, a  $1 \times 1 \times 1$  convolution is used to map the feature map to the number of classes. These outputs are of the same dimensions as the input volume and are converted to probabilistic segmentations of the foreground and background regions by applying a softmax layer voxel-wise. The voxels with a probability > 0.5 belong to the foreground class (proximal femur) and the rest to the background class. The proposed neural network model architecture was implemented using the flexible MIScnn framework (Müller and Kramer 2019) in Python.

#### 2.7. Hyperparameter Selection

A patch-based model, as opposed to analysis of the full image, was adopted in consideration of memory constraints and to exploit random cropping of patch volumes from the full images, further regularizing the model architecture. For the proposed u-net model, a patch volume of  $128 \times 128 \times 128$  voxels with an overlap of  $64 \times 64 \times 64$  voxels was forwarded to the network. This patch size is large enough to capture the entire femoral head, which is the most dynamic section of the proximal femur. Additionally, since the number 128 is readily divisible by two, we are left with integer values for patch dimensions after each use of max pooling.

A batch size of two, randomly cropped volumes of size  $128 \times 128 \times 128$  appeared to be the optimal combination w.r.t. memory constraints. This combination consistently outperformed stochastic models with the same size patch volumes or larger, as well as outperforming models with larger batch sizes, which necessitated smaller patch volumes to avoid memory overload. When implementing a model with a batch size of one, the loss function fluctuates heavily since it is only considering one sample at a

Model parameter	Proposed value		
Batch size	2		
Patch size	$128 \times 128 \times 128$		
Layers	4		
Feature map (highest resolution)	32		
Initial learning rate	$1 \cdot 10^{-4}$		
Epochs	300		
Iterations	80		
Training duration (hrs.)	12		
Adam optimizer (Kingma and Ba 2014)	$eta_1=0.9$ and $eta_2=0.999$		

Table 3.: Model parameters for the proposed neural network.

time. When the batch size was increased to four and a smaller patch size of  $64 \times 64 \times 64$  was used, the model performance slightly decreased because of the limited context in each patch. A variety of combinations were tested to arrive at these conclusions. Our proposed model was tuned to the parameter values displayed in Table 3.

## 2.8. Training

Our model was trained using a single Nvidia GeForce GTX 1080 Ti GPU for 300 epochs, which took roughly 12 hours. We randomly selected 30 CT images for training with corresponding manual segmentations of the left and right femur. Of these 60 proximal femurs, 54 were used for training and 6 were set aside to validate the performance of the model on unseen data. The number of slices was in the range of 98 to 148 slices. The ground truth annotations that comprised the training and validation sets are binary images identifying the voxels of the femur.

# 2.9. Postprocessing of Image Data

The postprocessing step of the masks was twofold: Firstly, each mask was padded with black voxels that were cropped out during preprocessing. The segmentation predictions were hence restored to the original resolution of the ground truth segmentations ( $512 \times 512$  voxels in-plane) and the same offset in the coordinate system. Secondly, the largest connected component had to be extracted to filter out noise in some of the segmentations outputted by our model.

#### 3. Experiments and Results

To evaluate our segmentation method we conducted three experiments: a comparison with a state-of-the art femur segmentation approach using Sample I, an evaluation on Sample II, and an FE analysis to assess the viability of using our model as part of our hip fracture screening tool.

Table 4.: A comparison between the graph-cut method (Pauchard et al. 2016) and our method w.r.t. the DSC and HD95 validation metrics. Let  $\sigma$  denote the standard deviation. We note that the results of the GC method were manually corrected for 47/48 proximal femures.

Method	Mean DSC $\pm\sigma$	Mean HD95 $\pm\sigma$ [mm]	Time [s]
Graph-cut method	$0.973 \pm 0.005$	$1.06\pm0.16$	120-300
Proposed method	$0.975\pm0.006$	$1.04\pm0.33$	9

#### 3.1. Evaluation Criteria

We used two evaluation metrics to evaluate the accuracy and robustness of our segmentation predictions, the DSC, (as discussed in Section 2.2 above) and the HD. While models seldom attempt to directly minimize the HD, this metric provides valuable insight into the performance of our model. This method quantifies the largest segmentation error by outputting the greatest distance from a point on the surface of the predicted segmentation mask to the closest point on the other surface of the ground truth segmentation mask. If X and Y are two non-empty subsets, the one-sided HD from X to Y is defined by the following equation:

$$\tilde{\delta}_H(X,Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|_2.$$
(2)

Similarly, going from Y to X yields

$$\tilde{\delta}_{H}(Y,X) = \max_{y \in Y} \min_{x \in X} \|x - y\|_{2}.$$
(3)

The bidirectional HD between these two sets is defined as:

$$\delta_H = \max(\tilde{\delta}_H(X, Y), \tilde{\delta}_H(Y, X)). \tag{4}$$

The function  $\tilde{\delta}_H(X, Y)$  finds the nearest point in Y to each point in X, and selects the largest distance. The bi-directional HD measures the degree of mismatch between the two subsets by taking the the maximum value between the one-sided HDs, as shown in (4). It is common practice in biomedical image segmentation to use the 95<sup>th</sup> percentile Hausdorff distance (HD95) in order to eliminate the influence of a small subset of outliers.

### 3.2. Comparison with the Graph-Cut Method

A direct comparison was carried out between the proposed method and the graph-cut method by (Pauchard et al. 2016) to demonstrate the effectiveness of our method on 24 unseen CT scans (Sample I). We computed the DSC and HD95 to quantitatively assess the accuracy of the two methods compared with ground truth manual segmentations (the current gold standard). As shown in Figures 5(a), 5(b), and Table 4, our method achieved a higher mean DSC score of  $0.975\pm0.006$  and a lower HD95 of  $1.04\pm0.33$ mm than that of the graph-cut method ( $0.973\pm0.005$  and  $1.06\pm0.16$ mm, respectively). As displayed in the figures, there is one outlier in the CNN prediction.



Figure 5.: A comparison between our method (CNN) and the graph-cut method (GC) on the same 24 CT image set (48 proximal femurs) of the left and right femurs (Sample I) validated on manual ground truth segmentations. Box plots (a) and (b) show the DSC and HD95, respectively, for the two methods.

The outlier in Figure 5(a) around DSC = 0.951 corresponds with the outlier in Figure 5(b) around HD95 = 3.25mm. Further investigation of the nature of the original CT scan revealed a possible cyst or the aftermath of intramedullary nailing to the femoral shaft (see Figure 6(a)). Since our method attempted to segment the structure inside of the bone, the DSC and HD95 metrics suffered moderately. This inadvertent labeling within the bone is, in part, a consequence of an absence of similar cases within the training set of the neural network. Data augmentation cannot be expected to simulate this type of variation if data of this kind are excluded from the training set. Figure 6(b) shows the results on a subject in which both methods performed well. We note that this comparison is not completely fair in the sense that a manual operator has corrected all but one (47/48) of the proximal femur segmentations outputted by the graph-cut algorithm. Nevertheless, it shows that similar results are achieved in a small fraction of the time (only 11s on average for the proposed CNN as opposed to 2-5 minutes for the graph-cut method) and in a completely automated manner.

# 3.3. Performance on Sample II of AGES

We demonstrate the performance of our model on the aforementioned Sample II subset by evaluating it on 1147 previously unseen proximal femurs that have been segmented semi-automatically. The box plots for both the DSC and HD95 scores are displayed in Figure 7. The mean DSC score was  $0.990 \pm 0.008$  and the mean HD95 was  $0.999 \pm 0.331$ mm. Only two data points had a DSC < 0.97 and a HD95 > 2.4mm (corresponding to the same two proximal femurs). The high average DSC, low HD95 value, and only two erroneous outliers out of a total of 1147 proximal femurs clearly reveal both the high accuracy and robustness of the proposed method. The time for each segmentation prediction averaged 11 seconds, which to our best knowledge is significantly faster than any current method, rendering our method viable for application to both large studies and clinical settings.

One of the segmentation predictions from our model that received a slightly lower DSC score and higher HD95 score (DSC = 0.930 and HD95 = 5.94mm) on the right



Figure 6.: A comparison between the original CT scan, the graph-cut method (GC), our method (CNN), and the manual ground truth segmentations (MAN) on five axial slices from two different patients in Sample I. Our method attempts to segment an artifact within the femoral shaft over the span of 16 slices (five shown) for a single case (a), however, it performs very well in all other cases, an example is shown in (b).

femur is shown in Figure 8. The region around the head of the proximal femur was heavily over-segmented, as shown in the 3D rendering of Figure 8 (right side). This comes as no surprise considering how unclear the separation is between the femoral head and the acetabulum in the axial view of the CT image (Figure 8, left). This is perhaps an indication of a birth defect or the result of a fractured bone that has since healed, however, any appraisal of the pathology without supplementary information on the subject is purely speculative. The other case that generated subpar results (DSC = 0.755 and HD95 = 10.6mm) is shown in Figure 9. The left femur appears to be tilted in the sagittal plane, causing our model to output a poor, and even fragmented segmentation for some axial slices. The tilt could be the result of femoral anteversion (in-toeing), a lenient adherence to imaging protocol, or a multitude of other reasons. The implementation of bone registration preprocessing step to enforce spatial normalization could be a requisite tool in achieving acceptable segmentations for this phenomenon.

### 3.4. FE Pipeline Results

In our last experiment we wanted to assess if our automated femur segmentation method could replace the manual segmentation approach currently being used in our FE pipeline for hip fracture predictions, giving way to a fully automated, end-toend hip fracture screening process. The FE models were based on the automated femur segmentations from the proposed method using an automated pipeline based on in-house Python scripts and a commercial preprocessor (Ansa 20.0.0; Beta CAE



Figure 7.: The distribution of DSC and HD95 scores on Sample II.



Figure 8.: Three axial slices are displayed on the left hand side. The shorthand "DI-COM" refers to the original CT scan, "CNN" refers to our method's segmentation, and "MAN" refers to the manually delineated ground truth segmentation. On the right, a 3D rendering of the erroneous segmentation prediction from our model (red) is overlaid with the ground truth segmentation (white).

Systems, Switzerland). Models were solved using LS-Dyna (LS-Dyna v11.0, LS-Dyna, Livermore, CA, USA) and results postprocessed in Python. A detailed description of the modeling strategy was published in (Fleps et al. 2021) but is briefly described here for clarity and context. The proximal femurs were meshed with 10-node tetrahedral elements with an average mesh size of 3 mm. Heterogeneous literature-based non-linear material properties were assigned to the mesh based on CT gray scale values following a validated material mapping procedure (Enns-Bray et al. 2017; Fleps et al. 2019). A femur loading alignment and boundary conditions representative of an unprotected fall to the side (10 adduction and 0 degrees internal rotation) were modeled. A schematic of the FE model of the proximal human femur is shown in Figure 10. This femur modeling has shown improved hip fracture classification performance compared to aBMD in the AGES RS cohort (Fleps et al. 2021). Femurs were loaded until peak force was exceeded. Femoral strength was evaluated by recording the maximum force that the femur was able to withstand. The FEA-derived femoral strength, based on



Figure 9.: A single axial slice is displayed in the upper-left corner where the plane cuts the figure on the right. The left and right proximal femures are shown on the bottom-left to illustrate the slant in the sagittal (YZ) plane. On the right, the erroneous segmentation prediction from our model (red) is overlaid with the ground truth segmentation (white).

the semi-automated (manual segmentations from the graph-cut method (Pauchard et al. 2016)) and automated (our proposed method) approaches, was compared using the coefficient of determination  $(R^2)$ , root mean square error (RMSE), mean absolute difference (MAE), and the maximum difference (Table 5). The data used were the predicted segmentations and ground truths from Sample II.



Figure 10.: An FE Model of the proximal human femur (Image reprinted from Fleps et al. 2020, with permission).

Of the 611 subjects we segmented for the left femur, 593 simulation models were

Proposed Method	$\mathbb{R}^2$	RMSE [N]	MAE <b>[%]</b>	Max Difference[%]
Left femur	0.986	212.2	-2.14	25.3
Right femur	0.988	177.0	-1.86	30.1

Table 5.: The  $\mathbb{R}^2$ , RMSE, MAE, and maximum difference for the FEA-derived femoral strength values between the automated and manual methods (left femurs and right femurs).

solved while 18 were exempted due to modeling errors (e.g., femurs that were not part of the cohort, data processing errors on the FE side, self-intersecting meshes, or the presence of extraneous volumes). Of these 593 models, 583 corresponding models based on the semi-automated segmentation were available to us. The predicted femoral strength values derived from the two segmentation methods were highly correlated (Figure 11). Of the 576 subjects we segmented for the right femur, 562 simulation models were solved while 14 models were exempted due to modeling errors. Of these 562 models, 553 corresponding models based on the semi-automated segmentation were readily available (see Table 6). As displayed in Table 5, similar results were achieved for both the left and right femurs, showing a very strong linear relationship between FEA-derived femoral strength from our fully automatic segmentations and from the semi-automatic segmentations. These results demonstrate that our method's segmentations are suitable for the FE pipeline and can be channeled through it in a robust manner.



Figure 11.: FEA-derived femoral strength based on the proposed femur segmentation method compared to femoral strength based on manual femur segmentations for left (left figure) and right (right figure) femurs.

Table 6.: The number of segmented femures, simulations run, and the number of femures used to compare the automated method to the semi-automated method.

Femur set	Left femur	Right femur
Segmented femurs	613	591
Evaluated by pipeline	593	562
Used for the comparison to manual segmentation	583	553
Excluded femurs		
Corresponding left femur missing	-	12
Excluded due to FEA errors	2	2
Excluded due to segmentation or meshing errors	16	13
Not part of cohort	2	2

# 4. Discussion

The aim of this work was to develop a fully automated neural network for proximal femur segmentation from CT images for application to an existing FE pipeline for hip fracture risk prediction. We demonstrated that our model's performance in terms of the DSC and HD95 is comparable to that of one of the previous best methods (Pauchard et al. 2016), yet significantly faster and without a human interaction. We subsequently presented our model's evaluation performance on 1147 unseen proximal femurs from the AGES-RS Sample II cohort (Harris et al. 2007), achieving a mean DSC of  $0.990\pm0.008$  and a mean HD95 of  $0.999\pm0.331$ mm. Lastly, we demonstrated a R<sup>2</sup> value of approximately 0.987 between FEA-derived femoral strength values based on our method's segmentations and based on manual segmentations.

The comparison with the graph-cut method (Pauchard et al. 2016) demonstrates our model's superior nature in terms of accuracy and robustness, despite not having a trained human operator to correct unacceptable segmentations ad libitum. Not only does our method output segmentation predictions an order of magnitude faster than the graph-cut method, but additionally relieves our future end-users (e.g., health care practitioners) from the monetary cost of hiring a trained specialist to perform the corrections. This is one of the significant hurdles that prior methods have struggled to surmount.

The results from Sample II demonstrate our model's undeniable accuracy and robustness, allowing us to process even larger cohorts in the near future. With regard to the very few problematic cases encountered in Sample II, we speculate which measures are justifiable to take in order to further increase the robustness of our model. For the CT scans in which the proximal femur appears to slant, as in Figure 9, we can argue that the use of registration to a common coordinate system would improve our model's prediction. Registration will eliminate the need to capture the variability within the data set for some of the most extreme cases with data augmentation. The use of such aggressive augmentation parameters is a futile pursuit that severely hampers the overall performance of the model, considering its sensitivity to radical transformations. By spatially transforming a source image to align with a target image, representing the mean shape constructed from a statistical atlas of healthy patient CT images, we enforce spatial normalization to the source image. We must, however, ask ourselves whether these preprocessing measures are worth the added effort considering how infrequently we encounter such anomalies. It is reasonable to assume that in clinical practice, physicians would immediately flag any patient scans that deviate significantly from the mean and would not be good candidates for our hip fracture screening tool. If our evaluation set is any indicator of the prevalence of anomalies in the general population, then this would amount to a negligible number of patients who could not be screened with our method. We note that in order to apply our model to CT images from a different scanner, the model would likely have to be trained on a set of images from that CT scanner.

In the last part of the Results section, we fed our model's segmentation predictions to the FE pipeline. The strong  $\mathbb{R}^2$  values between FEA-derived femoral strength values based on our model's segmentations and manual ones, demonstrate our method's ability to reliably produce segmentations that can be processed by the FE pipeline with very similar predicted femoral strength.

The primary limitations of this research are twofold: firstly, we have yet to demonstrate our solution's ability to perform well on cross-cultural data, that is, on CT images beyond the Icelandic elderly as well as its performance on scans from different scanner manufacturers. If the performance of our model turns out to be unsatisfactory on other cohorts, then a possible solution would be to re-train the model with either a mix of scans from multiple populations or exclusively train on images from the cohort at hand. However, the desired outcome would be a segmentation tool that can be applied to all proximal femur scans independent of population and scanner manufacturer. The second limitation of our model is its performance on heavily deformed proximal femurs. There is an inherit trade-off between general segmentation performance and variability within the training set. That is, if we bias the training set with too many deformed bones, we will compromise the general performance on the validation set. As a result, we justify the exclusion of acutely deformed bones in the training set in order to improve performance on bones that do not deviate drastically from the mean.

In summary, our method has addressed the most pressing hindrances that have impeded prior methods; our method does not require a trained operator to make ad hoc corrections to unsatisfactory segmentations, the robustness of the model would justify the radiation exposure to the patient, and the processing time of each segmentation is well within reasonable bounds for clinical viability. More importantly, this demonstrates that with the proposed segmentation method, the hip fracture risk assessment can now be performed in a fully automated manner. The next step in the ongoing development of the hip fracture screening tool is applying the model to a much larger cohort of the AGES-RS.

# 5. Conclusion

Here we introduced a fully automated, accurate, robust, and fast segmentation method for segmenting the proximal femur from CT images. The mean DSC was  $0.990 \pm 0.008$ and mean HD95 was  $0.999 \pm 0.331$ mm when evaluated on 1147 manually segmented femurs. The proposed method is superior to preceding methods in terms of previously reported numbers of DSC and HD95 metrics and, most importantly, does not require any manual interaction. In addition, each segmentation prediction can be generated, on average, in 11 seconds instead of the many minutes it takes some other approaches. We will conduct a more extensive evaluation on a larger cohort and, in turn, integrate the method into our existing FE pipeline, bringing it one step closer to becoming a clinically viable option for screening at-risk patients for hip fracture susceptibility.

## 6. Acknowledgements

This research was supported in part by the RANNIS Icelandic student innovation fund, Iceland, and the Strategic Focus Area "Personalized Health and Related Technologies" of the ETH Domain, ETH Zurich, Switzerland [grant numbers 2018-430, 2018-325].

## 7. Disclosure Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### References

- 2015. World population ageing 2015. UN: Department of Economic and Social Affairs, Population Division. (ST/ESA/SER.A/368).
- Bjornsson PA, Helgason B, Palsson H, Sigurdsson S, Gudnason V, Ellingsen LM. 2021. Automated femur segmentation from computed tomography images using a deep neural network. In: Gimi BS, Krol A, editors. Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging; vol. 11600. International Society for Optics and Photonics; SPIE. p. 324 – 330. Available from: https://doi.org/10.1117/12.2581100.
- Chang Y, Yuan Y, Guo C, Wang Y, Cheng Y, Tamura S. 2019. Accurate pelvis and femur segmentation in hip CT with a novel patch-based refinement. IEEE J Biomed Health Inform. 23(3):1192–1204.
- Chen F, Liu J, Zhao Z, Zhu M, Liao H. 2019. 3D feature-enhanced network for automatic femur segmentation. IEEE Journal of Biomed and Health Inform. 23(1):243–252.
- Chengwen C, Bai J, Wu X, Zheng G. 2015. Mascg: Multi-atlas segmentation constrained graph method for accurate segmentation of hip CT images. Medical image analysis. 26(1):173–184.
- Dice LR. 1945. Measures of the amount of ecologic association between species. Ecology. 26(3):297–302.
- Enns-Bray W, Bahaloo H, Fleps I, Ariza O, Gilchrist S, Widmer R, Guy P, Pálsson H, Ferguson S, Cripton P, et al. 2017. Material mapping strategy to improve the predicted response of the proximal femur to a sideways fall impact. Journal of the Mechanical Behavior of Biomedical Materials. 78:196–205.
- Enns-Bray W, Bahaloo H, Fleps I, Pauchard Y, Taghizadeh E, Sigurdsson S, Aspelund T, Büchler P, Harris T, Gudnason V, et al. 2019. Biofidelic finite element models for accurately classifying hip fracture in a retrospective clinical study of elderly women from the AGES Reykjavik cohort. Bone. 120:25–37.
- Fleps I, Enns-Bray W, Baker A, Bahaloo H, Sigurdsson S, Gudnason V, Ferguson S, Pálsson H, Helgason B. 2021. FEM-derived femoral strength is a better predictor of hip fracture risk than aBMD in the AGES Reykjavik study cohort. Under revision in Bone.
- Fleps I, Guy P, Ferguson S, Cripton P, Helgason B. 2019. Explicit finite element models accurately predict subject-specific and velocity-dependent kinetics of sideways fall impact. Journal of Bone and Mineral Research. 34(10):1837–1850.
- Haleem S, Lutchman L, Mayahi R, Grice J, Parker M. 2008. Mortality following hip fracture: Trends and geographical variations over the last 40 years. Injury. 39(10):1157–1163.
- Harris T, Launer L, Eiriksdottir G, Kjartansson O, Jonsson P, Sigurdsson G, Thorgeirsson G, Aspelund T, Garcia M, Cotch MF, et al. 2007. Age, Gene/Environment Susceptibility-Reykjavik Study: Multidisciplinary Applied Phenomics. American journal of epidemiology. 165:1076–87.
- Hayes W, Myers E, Robinovitch S, Kroonenberg AVD, Courtney A, McMahon T. 1996. Etiology and prevention of age-related hip fractures. Bone. 18(1):77S–86S.

- Huo Y, Terry J, Wang J, Nair S, Lasko T, Freedman B, Carr J, Landman B. 2019. Fully automatic liver attenuation estimation combing CNN segmentation and morphological operations. Medical Physics. 46(8):3508–3519.
- Ioffe S, Szegedy C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. ArXiv. abs/1502.03167.
- Isensee F, Jäger P, Wasserthal J, Zimmerer D, Petersen J, Kohl S, Schock J, Klein A, Roß T, Wirkert S, et al. 2020. Batchgenerators - a Python framework for data augmentation. Available from: https://github.com/MIC-DKFZ/batchgenerators.
- Kingma D, Ba J. 2014. Adam: A method for stochastic optimization. International Conference on Learning Representations. Available from: https://arxiv.org/pdf/1412.6980.pdf.
- LeCun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W, Jackel L. 1990. Handwritten digit recognition with a back-propagation network. In: NIPS; Denver, CO, USA. p. 396–404.
- Milletari F, Navab N, Ahmadi SA. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 10. p. 565–571.
- Müller D, Kramer F. 2019. MIScnn: A framework for medical image segmentation with convolutional neural networks and deep learning. 01. Presented at KiTS19.
- Parkkari J, Kannus P, Palvanen M, Natri A, Vainio J, Aho H, Vuori I, Järvinen M. 1999. Majority of hip fractures occur as a result of a fall and impact on the greater trochanter of the femur: A prospective controlled hip fracture study with 206 consecutive patients. Calcif Tissue Int. 65(3):183–187.
- Pauchard Y, Fitze T, Browarnik D, Eskandari A, Pauchard I, Enns-Bray W, Pálsson H, Sigurdsson S, Ferguson SJ, Harris TB, et al. 2016. Interactive graph-cut segmentation for fast creation of finite element models from clinical CT data for hip fracture prediction. Comput Methods Biomech Biomed Engin. 20(3):342.
- Ronneberger O, PFischer, Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MIC-CAI); (LNCS; vol. 9351). Springer. p. 234–241. (available on arXiv:1505.04597 [cs.CV]); Available from: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a.
- Schuit S, van der Klift M, Weel A, de Laet C, Burger H, Seeman E, Hofman A, Uitterlinden A, van Leeuwen J, Pols H. 2004. Fracture incidence and association with bone mineral density in elderly men and women: the rotterdam study. Bone. 34(1):195–202.
- Shao M, Han S, Carass A, Li X, Blitz A, Shin J, Prince J, Ellingsen L. 2019. Brain ventricle parcellation using a deep neural network: Application to patients with ventriculomegaly. NeuroImage: Clinical. 23:101871.
- Stone K, Seeley D, Lui L, Cauley J, Ensrud K, Browner W, Nevitt M, Cummings S. 2003. Bmd at multiple sites and risk of fracture of multiple types: long-term results from the study of osteoporotic fractures. J Bone Miner Res. 18(11):1947–1954.
- Wainwright S, Marshall L, Ensrud K, Cauley J, Black D, Hillier T, Hochberg M, Vogt M, Orwoll E, of Osteoporotic Fractures Research Group S. 2005. Hip fracture in women without osteoporosis. J Clin Endocrinol Metab. 90(5):2787–2793.
- Wang H, Yushkevich PA. 2013. Multi-atlas segmentation without registration: A supervoxelbased approach. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013; Berlin, Heidelberg. Springer Berlin Heidelberg. p. 535–542.
- Younes L, Nakajima Y, Saito T. 2019. Fully automatic segmentation of the femur from 3D-CT images using primitive shape recognition and statistical shape models. Int J Comput Assist Radiol Surg. 9(2):189–196.
- Zhao C, Keyak JH, Tang J, Kaneko TS, Khosla S, Amin S, Atkinson E, Zhao L, Serou M, Zhang C, et al. 2020. A deep learning-based method for automatic segmentation of proximal femur from quantitative computed tomography images. ArXiv. abs/2006.05513.