Fast and robust femur fracture segmentation from computed tomography images for patient-specific hip fracture risk screening

Pall Asgeir Bjornsson, Alexander Baker, Ingmar Fleps, Yves Pauchard, Halldor Palsson, Stephen J. Ferguson, Sigurdur Sigurdsson, Vilmundur Gudnason, Benedikt Helgason, and Lotta Maria Ellingsen

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 ageing societies are associated with a higher prevalence of chronic diseases. Policymakers are thus forced to reconsider the status quo of health care systems, moving away from face-to-face consultation-based care towards a decentralised 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 characterised 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. (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 grey level based material properties to the FEIs, 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) 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 diffuse 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. (2016). In short, this method separates the background class from the target 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. (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. (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.

Figure 1. An axial CT image slice showing the distinction that our model has to make in order to correctly segment the femoral head.
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 modelling. 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 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 viewed as a particularly attractive alternative for application in biomedical imaging. However, the u-net architecture proposed by 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) ages: 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) comprises 48 ‘gold standard’ manually delineated proximal femur segmentations from 24 CT images. The second subset within the AGES data set (Sample II) 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, 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:

$$DSC = \frac{2 \sum_{i,j} p_i \cdot q_i}{\sum_{i} p_i^2 + \sum_{j} q_j^2}.$$  \hspace{1cm} (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 $q_i \in G$-net. 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 ages
- Normalisation
  - On-the-fly data augmentation
  - Patch-based 3D u-net
- Training using the Dice loss function for a pre-defined number of epochs
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 side hip/upper leg with in-plane resolution $512 \times 256$ voxels and 98–148 slices. The CT images were normalised such that all intensity values were linearly shifted and scaled from HU to the range $[0, 1]$. Min-max normalisation has the advantage over z-score normalisation of preserving the scale of the data. Models using both normalisation 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 maximising 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 Image Segmentation with Convolutional Neural Networks (MIScnn) framework. 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,
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.

| Parameter Range | Brightness | Rotation (X, Y, Z) | Scaling Factor (\( \alpha \)) | Elastic Smoothing (\( \sigma \)) |
|-----------------|------------|-------------------|-------------------------------|-------------------------------|
|                 | (0.75, 1.25) | (-3°, 3°) | (0.95, 1.05) | (0, 100)          | \( \sigma = (9, 13) \) |

allowing the model to encounter a diverse set of images, thereby decreasing redundancy. For the proposed method, on-the-fly data augmentation, in concert with parameter sharing (LeCun et al. 1990) and batch normalisation (BN) (Ioffe and Szegedy 2015), rendered the use of explicit regularisation 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 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 localisation with the use of transposed convolutions. In the final layer of the network, a 1 × 1 × 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 regularising 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 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

![Figure 4](https://example.com/figure4.png)

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 \( \text{conv} \) and rectified linear unit as \( \text{ReLU} \).
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 × 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.

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 minimise 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 \(X\) are two non-empty subsets, the one-sided HD from \(X\) to \(X\) is defined by the following equation:

\[
\delta H(X, Y) = \max_{x \in X} \min_{y \in Y} x - y.  
\]

Similarly, going from \(Y\) to \(X\) yields

\[
\delta H(Y, X) = \max_{y \in Y} \min_{x \in X} x - y. 
\]

The bidirectional HD between these two sets is defined as

\[
\delta_H = \max(\delta H(X, Y), \delta H(Y, X)) 
\]

The function \(\delta H(X, Y)\) 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 95th 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 Figure 5(a,b), and Table 4, our method achieved a higher mean DSC score of 0.975 ± 0.006 and a lower HD95 of 1.04 ± 0.33mm than that of the graph-cut method (0.973 ± 0.005 and 1.06 ± 0.16mm, respectively). As displayed in the figures, there is one outlier in the CNN prediction.

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 labelling 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 ± 0.008 and the mean HD95 was 0.999 ± 0.331mm. Only two data points had a DSC < 0.97 and an 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 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 anteverision (in-toeing), a lenient adherence to imaging protocol, or a multitude of other reasons. The implementation of bone registration preprocessing step to enforce spatial normalisation could be a requisite tool in achieving acceptable segmentations for this phenomenon.

Table 4. The $R^2$, RMSE, MAE, and maximum difference for the FEA-derived femoral strength values between the automated and manual methods (left femurs and right femurs).

| Proposed Method | $R^2$  | RMSE [N] | MAE [%] | Max Difference [%] |
|-----------------|--------|----------|---------|-------------------|
| Left femur      | 0.986  | 212.2    | -2.14   | 25.3              |
| Right femur     | 0.988  | 177.0    | -1.86   | 30.1              |

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.

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 artefact 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).
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 “DICOM” 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).

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 femurs 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).
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-to-end 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; Beta CAE 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 modelling 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 grey 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 abduction and 0 degrees internal rotation) were modelled. A schematic of the FE model of the proximal human femur is shown in Figure 10. This femur modelling 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 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 4). The data used were the predicted segmentations and ground truths from Sample II.

Of the 611 subjects we segmented for the left femur, 593 simulation models were solved while 18 were exempted due to modelling 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 modelling errors. Of these 562 models, 553 corresponding models based on the semi-automated segmentation were readily available (see Table 5). As displayed in Table 4, 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-automated segmentations. These results demonstrate that our method’s segmentations are suitable for the FE pipeline and can be channelled through it in a robust manner.

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 ± 0.008 and a mean HD95 of 0.999 ± 0.331mm. Lastly, we demonstrated a $R^2$ 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.

Figure 10. An FE model of the proximal human femur (image reprinted from Fleps et al. 2022, with permission).
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 normalisation 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 $R^2$ 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 inherent 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.
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 ± 0.008 and mean HD95 was 0.999 ± 0.331mm 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.

Endnotes
1. The Hounsfield unit is a relative quantitative measurement of radio density used by radiologists in the interpretation of CT images.
2. This sample set was used by (Pauchard et al. 2016) to evaluate their model.
3. This sub-cohort from AGES-RS, including all fracture cases, was used by (Enns-Bray et al. 2019).
4. MIcnn is an open-source Python library and intuitive API for medical image segmentation pipelines (Müller and Kramer 2019).
5. On-the-fly data augmentation eliminates the need for excessive storage of augmented images by performing the augmentation prior to each optimisation iteration.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Icelandic Centre for Research [, The Strategic Focus Area “Personalized Health and Related Technologies” of the ETH Domain [018-430, 2018-325]

Notes on contributors

Pall Asgeir Björnsson received his M.Sc. degree in Electrical and Computer Engineering from the University of Iceland in 2020. He is currently working for the University in the field of image processing and deep neural networks, as well as the field of geothermal heat transfer. His primary research interests are the use of AI in image processing, financial markets, and decision making.

Alexander Baker is a doctoral student within the Laboratory for Orthopaedic Technologies at ETH Zurich. He possesses a B.Sc in Mechanical Engineering from UCLA and a M.Sc. in Mechanical Engineering from UC Berkeley. He has done previous work in the field of injury biomechanics, including assisting in the development of a finite element model of an American football helmet, and improving the methodology to construct injury risk curves. His doctoral research focuses on computational modeling of hip fracture using finite element analysis and applying those models to large population cohorts in order to better understand and predict hip fracture.

Dr. Ingmar Fleps is a mechanical engineer and material scientist with a Ph.D. and postgraduate training in biomechanics. His work focuses on medical image processing, experimental modelling of biomechanical systems, and computational modeling based on medical images to improve our understanding of the mechanical interaction and failure processes of human tissues, as well as tissues and implants. Dr. Fleps is a postdoctoral associate at Boston University where he currently works on mechanical modelling of the spine.

Dr. Yves Pauchard is an electrical engineer with graduate and post-graduate training in medical imaging and analysis. His research interests include software development, image analysis and computational modelling in musculoskeletal applications with a focus on bone. He is the project lead of MITK-GEM https://simtk.org/projects/mitk-gem, an image analysis application to generate finite element models from medical images. Dr. Pauchard holds a faculty position in Electrical and Software Engineering at the University of Calgary. In his role, he teaches courses in software development, data analysis and machine learning.

Dr. Hallóðr Pálsson finished a CS degree in Mechanical Engineering from the University of Iceland in 1993. He then finished a M.Sc. degree in the same subject from the Technical University of Denmark in 1995. This degree was followed by a Ph.D. degree at the Technical University of Denmark in the year 2000, where the subject was “Methods for planning and operating decentralized combined heat and power plants”. After that Hallóðr worked on a research project involving mathematical modeling of high energy electric arc FeSi production, until acquiring a position as an Associate Professor at the University of Iceland in 2005. Hallóðr has been a full professor at the University of Iceland since 2016, working on subjects regarding geothermal energy utilization, fluid mechanics, biomechanics, and mathematical modeling in general.

Dr. Stephen Ferguson is a Full Professor of Biomechanics at the Institute for Biomechanics of the ETH Zurich in Switzerland. He has a B.A.Sc. in Mechanical Engineering from the University of Toronto (91) and a M.Sc. (1994) and Ph.D. (2000) from Queen’s University. The focus of his research is the study of the mechanical and biological mechanisms of musculoskeletal disorders and injuries and the development and application of innovative technologies for their treatment. The challenge presented by an ageing population is of primary importance in his work, with research programs on bone fragility, joint arthroplasty, disc degeneration and postural decay. He and his group study new biomaterials, material fabrication technologies, implant concepts and diagnostic methods and develop the technical means for their application in the clinic. Dr. Ferguson is the author of over 250 scientific papers, 7 book chapters and 4 patent applications. He and his collaborators have received the European Spine Journal GRAMMER prize for best scientific work, the CTI Medtech Award for the translational project “BoneWelding”. He is on the editorial board of the journals Clinical Biomechanics, Journal of Biomechanics and European Cells & Materials and is a former president of the European Society of Biomechanics.

Sigurdur Sigurdsson is currently the CEO of the Icelandic Heart Association Radiology and a lecturer at the Faculty of Medicine, University of Iceland. Sigurdsson’s research involves radiology with focus on neuroimaging of the aging brain in cohort studies. Sigurdsson has published over 140 peer-reviewed papers in scientific journals. Sigurdsson received his MSc from South Bank University, London UK in clinical Magnetic Resonance Imaging in year 2001 and BSc in Radiography from the Technical University of Iceland in 1994.
**Dr. Vilmundur GudnaSON** is the director of the Icelandic Heart Association and a professor at the Faculty of Medicine, University of Iceland. He obtained his medical degree from the University of Iceland and a PhD in molecular genetics from University College London. He is the Principal Investigator for the Age Gene/Environment Susceptibility Reykjavik (AGES-Reykjavik) study, based on the 50-year-long Reykjavik study, and for the REFINe REykjavik study of younger generations. The AGES-Reykjavik study is a study on aging with extremely deep phenotyping, including using imaging with CT and MRI and detailed follow-up hospital data, such as myocardial infarction and fractures. Professor GudnaSON (https://orcid.org/0000-0001-5696-0084) has published widely in various high impact journals and has been on the list of Thompson Reuter/Clarivate Highly Cited Researchers 2015, 2016, 2017 and 2018. Vilmundur GudnaSON - Google Scholar.

**Dr. Benedikt Helgason** is a civil engineer with post-graduate training in biomechanics. His research focuses on the development of modeling techniques for simulating bone pathologies and treatments. He is the PI on several projects in the field of healthy aging, e.g. the Future Health Technologies program funded by the National Research Foundation of Singapore, and a project on in silico clinical trials on personalized preventive measures for managing hip fractures in the elderly, funded by the Swiss National Science Foundation. Dr. Helgason holds a position as a senior scientist at the ETH-Zurich, Zurich, Switzerland.

**Dr. Lotta Maria Ellingsen** is an Associate Professor and Vice Chair of the faculty of Electrical and Computer Engineering at the University of Iceland. She received her M.S.E. (2004) and Ph.D. (2008) degrees in Electrical and Computer Engineering from the Johns Hopkins University, Baltimore, USA. Her research interests are in the field of medical image processing and analysis with emphasis on 3D medical image registration and segmentation, with applications to bones, brains, and statistical atlases. She has particular expertise in deformable 3D-3D image registration of magnetic resonance images of the human brain for different application purposes. Her current research focuses on developing pipelines for automatic segmentation and labeling of brain anatomy for systematic analysis of brain morphology to better characterize neurodegenerative diseases for personalized medicine.

**ORCID**

Pall Asgeir Björnsson http://orcid.org/0000-0001-5702-4706
Ingrmar Fleps http://orcid.org/0000-0002-6254-8878
Yves Pauchard http://orcid.org/0000-0001-5477-0594
Hildur Palsson http://orcid.org/0000-0003-4112-6729
Vilmundur GudnaSON http://orcid.org/0000-0001-5696-0084
Benedikt Helgason http://orcid.org/0000-0001-8324-2651
Lotta Maria Ellingsen http://orcid.org/0000-0003-2139-0979

**References**

Björnsson 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 B Krol A, editors. Medical imaging 2021: biomedical applications in molecular, structural, and functional imaging; vol. 11600. International Society for Optics and Photonics; SPIE; pp. 324–330. doi: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. doi:10.1109/JBHI.2018.2834551.

Chen F, Liu J, Zhao Z, Zhu M, Liao H. 2019. 3D feature-enhanced network for automatic femur segmentation. IEEE J Biomed Health Inform. 23(1):243–252. doi:10.1109/JBHI.2017.2785389.

Chengwen C, Bai J, Wu X, Zheng G. 2015. Masgc: multi-atlas constrained graph method for accurate segmentation of hip CT images. Med Image Anal. 26(1):173–184. doi:10.1016/j.media.2015.08.011.

Dice LR. 1945. Measures of the amount of ecologic association between species. Ecology. 26(3):297–302. doi:10.2307/1932409.

Enns-Bray W, Bahaloo H, Fleps I, Ariza O, Gilchrist S, Widmer R, Guy P, Pålsson 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. J Mech Behav Biomed Mater. 78:196–205. doi:10.1016/j.jmbbm.2017.10.033.

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. doi:10.1016/j.bone.2018.09.014.

Fleps I, Enns-Bray W, Baker A, Bahaloo H, Sigurdsson S, Gudnason V, Ferguson S, Pålsson 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. J Bone Mineral Res. 34(10):1837–1850. doi:10.1002/jbmr.3804.

Haleem S, Lutchman I, 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. doi:10.1016/j.injury.2008.03.022.

Harris T, Launer L, Eriksdottr 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. Am J Epidemiol. 165:1076–1087. doi:10.1093/aje/kwk115.

Hayes W, Myers E, Robinovitch S, Koorenberg AVD, Courtney A, McMahon T. 1996. Etiology and prevention of age-related hip fractures. Bone. 18(1):775–865. doi:10.1016/8756-3282(95)00383-5.

Huo Y, Terry J, Wang J, Nair S, Lasko T, Freedman B, Cart J, Landman B. 2019. Fully automatic liver attenuation estimation combination CNN segmentation and morphological operations. Med Phys. 46(8):3508–3519. doi:10.1002/mp.13675.

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, Schick J, Klein A, Roß T, Wirkert S, et al. 2020. Batchgenerators - a Python framework for data augmentation. https://github.com/MIC-DKFZ/batchgenerators.

Kingma D, Ba J. 2014. Adam: a method for stochastic optimization. International Conference on Learning Representations. 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. Denver (CO): NIPS; pp. 396–404.

Milletari F, Navab N, Ahmadi SA. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 10:565–571.

Müller D, Kramer F. 2019. Mircnn: a framework for medical image segmentation with convolutional neural networks and deep learning. 01. Presented at KiTS19.

Park 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. doi:10.1007/s002239900679.

Pauchard Y, Fitzte T, Browarnik D, Eskandari A, Pauchard I, Enns-Bray W, Pålsson 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, Fischer P, Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical image computing and computer-assisted intervention (MICCAI); (LNCS; vol. 9351). Springer; pp. 234–241. arXiv:1505.04597 [cs.CV].
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. doi:10.1016/j.bone.2003.10.001.

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. 23:101871. doi:10.1016/j.neuroimage.2019.101871.

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. doi:10.1359/jbmr.2003.18.11.1947.

UN. 2015. World population ageing 2015. UN: Department of Economic and Social Affairs, Population Division. (ST/ESA/SER.A/368).

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. doi:10.1210/jc.2004-1568.

Wang H, Yushkevich PA. 2013. Multi-Atlas segmentation without registration: a supervoxel-based 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; pp. 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. doi:10.1007/s11548-013-0950-3.

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.