Classification of Atypical Femur Fracture with Deep Neural Networks

Yupei Chen[1], Chunliang Wang[1], and Jörg Schilcher[2]

[1]Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Department of Orthopedics and Experimental and Clinical Medicine, Faculty of Health Science, Linkoping University, Linkoping, Sweden

May 2019

1 Introduction

Antiresorptive drugs such as bisphosphonates have been used successfully in the treatment and prophylaxis of osteoporosis for decades. Recently however, a strong association between bisphosphonate treatment and the occurrence of an insufficiency type of fracture in the femur has been shown. These fractures do not occur in the metaphyseal area of the femur, as most fragility type fractures do. Their location is in the diaphyseal or subtrochanteric area of the bone, and therefore has coined the name Atypical Femoral Fractures (AFF). The insufficiency type of fracture of AFF is associated with specific radiographic features, including a transverse or short oblique fracture configuration and focal cortical thickening. These features differ from common femur fractures (CFF) which shown oblique fracture lines and no signs of cortical thickening. In practice the diagnostic accuracy to identify AFF from radiology reports is poor. This study aims to test the ability of deep learning algorithms to classify atypical femoral fractures and common femur fracture on the first diagnostic X-ray examination.

2 Methods

X-ray images from 200 patients were extracted, of which 94 subjects with aff and 106 subjects with nff. Each subject has several x-ray images. Therefore, 796 imgaes (AFF, n=397; NFF, n=399) are used to train and test the proposed solution. The images were extracted from clinical PACS and anonymized. The original data were converted to JPEG from DICOM format. To eliminate the effect of varying image size, all images were converted to squares to avoid distortion during training and testing. All images were downsized to 256×256 pixel to reduce data size and computing time. The images intensity was normalized to the range of 0-1.0. The images were augmented through random rotation (±10 degrees), shifting ($< 10\%$) and zooming ($< 10\%$).

Two diagnostic pipelines were constructed using the convolutional neural networks (CNN) as the core classifier. One is a fully automatic pipeline, where the x-ray image is directly input into the network with size and intensity normalization steps as described above. Three popular CNN architectures, VGG19, InceptionV3 and ResNet50, were tested for classifying the images to either AFF or CFF. Transfer learning technique was used to pre-train these networks using images from ImageNet.

Another interactive pipeline requires more pre-processing for the dataset before the images are sent to the CNNs. The user to re-orientate the femur bones above the fractures to a vertical position and move the fracture line to the image center. The femur bones are in different positions and we expect the fracture in the center with vertical femur bones. This process was done by a medical engineering student. Additionally, the rotated images are cropped into the size of 256×256 pixels around the center of the fracture. In this way we can expect the networks to focus on the specific radiographic features. Manual screening was conducted after imputation to ensure the image quality.

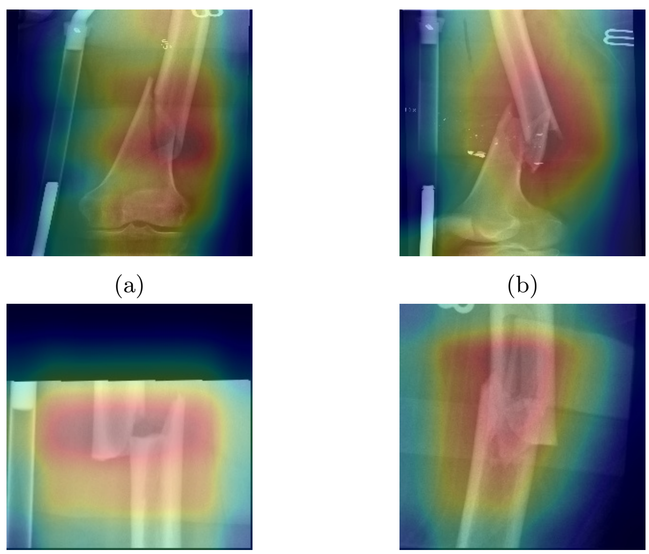
The diagnostic accuracy was evaluated using 5-fold cross-validation. The training and validation process was repeated several times and each time with different folds. Mean values and standard deviations were calculated to achieve higher accuracy and less bias with cross validation. Additionally, class activation mapping was applied to visualize the features that networks are learning.

3 Results

With the fully automated diagnostic pipeline, we achieved diagnostic accuracy of 73.6%, 88.4% and 93.9%, with VGG19, InceptionV3 and ResNet50, respectively. With the interactive diagnostic pipeline, the diagnostic accuracy was improved to 92.8%, 94.7% and 96.8%. The attention maps indicate discriminative image regions used by the neural networks to identify certain class (Figure 1). Based on these results we conclude that both the automatic and the interactive pipeline are learning with the fracture pattern, while the interactive pipelines show fewer off-target hot spots.

4 Conclusion

We tested two types of AI based diagnostic pipelines to classify AFF and CFF using plain x-ray images. The preliminary results show a promising high diagnostic accuracy with limited user intervention, by far exceeding that of clinical radiology reports.



(a) (b)

(c) (d)

Figure 1: Example results of Attention Maps