Deep Learning Segmentation of Hip Radiographs for Guiding Opportunistic Management of Osteoporotic Bone Fractures

University of Pennsylvania Perelman School of Medicine, Philadelphia PA 19104
simimattikallli@gmail.com

Disclosures: None

INTRODUCTION: Osteoporosis, indicated by decreased bone mineral density (BMD), diminishes the density, structure, and quality of bones over time, thereby increasing the risk of bone fracture. It has a correlation with factors such as obesity, malnourishment, and aging, with its prevalence in American adults aged 50 and over being 12.6%. DXA scans, despite being the current clinical standard for osteoporosis diagnosis, BMD testing, and fracture risk assessment, are cost prohibitive, with each procedure costing between $160 and $175 without insurance. This issue highlights the lack of accessibility of DXA scans, particularly for the economically disadvantaged, for necessary orthopedic care. As a result, DXA scan screening rates are extremely low at 24.5%. Development of this machine learning segmentation model seeks to create an accessible method to generate accurate masks of femurs using pelvic x-rays, allowing for opportunistic screening, and thereby increasing the rates of osteoporosis screening. With this approach, osteoporosis can be diagnosed in the absence of DXA scans using pre-existing x-rays, which may have been taken for any reason in the past. Predicted segmentation can be used to determine risk for a variety of bone conditions, including osteoporosis, reducing time, cost, and resources of multiple exams.

METHODS: Pelvic x-rays from 3,185 patients (ages 45-79, BMI 16.9–48.7) were collected from the publicly available Osteoarthritis Initiative Dataset. The images were split in half, creating 6,370 (512x1024) total radiographs. The right side of the radiograph was flipped, so every x-ray was oriented identically. The radiographs were segmented manually, as shown below in Figure 1. 80% or 5,096 images were used as the training dataset for our model. The other 20% (1,274 images) were used to test the model. A convolutional neural network (CNN) was built in which upsampling operators replace pooling operators, thereby increasing the resolution of the output. To optimize accuracy, the model architecture was varied throughout the training process, consisting of 3-, 4-, and 5-encoding and decoding layers. Figure 1 and Table 1 demonstrate the architecture of the 4 layered variation CNN’s architecture.

RESULTS SECTION: 2 versions of a 3-, 4-, and 5-layer convolutional neural network (CNN) model were trained and validated on a dataset of n = 1,247 femur x-rays from 623 distinct patients. The resulting mean dice scores, HD95 distance, RAVID, and ASD are shown in Table 1. For the testing conditions, either the training and testing dataset was histogram equalized, or both were left unnoted. Notably, the 4- and 5-layer networks, regardless of the contrast of the training and testing dataset, performed equally well with DSC scores of .96 AU, with only a .01 AU difference in their standard deviations. However, the 3-layer models (runtime averaging 30 min) did not perform as well as the 4- (runtime averaging an hour) and 5-layer models (runtime averaging 4 hours), with average RAVID scores peaked at 22 mm, average HD95 scores peaked at 31.89 mm, and average ASD scores peaked at 150.61. The large difference in ASD scores for the 3-layer model indicates larger outliers in the performance compared to those of the 4 and 5 layered models. There seems to be a consistent pattern of whether equalizing the training and testing dataset improves the accuracy of the model.

DISCUSSION: The CNN developed accurately segments femurs on pelvic radiographs (as shown in Figure 2) across differing scales and graphic quality of images, demonstrating robust functionality under diverse conditions. The large range of both weight and height allows this model to be applicable and accurate for a wide variety of demographics, including the elderly and obese, who have a high risk of osteoporosis. Furthermore, out of the 5,096 training images, 55 had hardware in them, allowing the model to remain accurate when provided with x-rays with various metals present (e.g., screws). However, the possibility of human error in ground truth creation is increased due to the hardware impedes a clear view of the femur, consequently decreasing the model’s accuracy. In addition, the ground truth segmentations were created manually using ITK-Snap, which introduces the risk of human error in all masks that were used to train the model. Despite these potential limitations, the models had a mean DSC score of .91mm, which is classified as “excellent agreement.”

SIGNIFICANCE/CLINICAL RELEVANCE: The developed network allows for accurate and automatic femur segmentation with minimal labor, thereby increasing accessibility to preventative care through opportunistic screening of those at greater fracture risk. Through future developments, the model will allow for rapid and autonomous diagnosis of bone conditions such as osteoporosis.