Journal of Digital Imaging (2020) 33:1209–1217
https://doi.org/10.1007/s10278-020-00364-8

ORIGINAL PAPER

Advanced Deep Learning Techniques Applied to Automated Femoral Neck Fracture Detection and Classification

Simukayi Mutasa[1] · Sowmya Varada[1] · Akshay Goel[1] · Tony T. Wong[1] · Michael J. Rasiej[1]

Published online: 24 June 2020

Abstract

To use deep learning with advanced data augmentation to accurately diagnose and classify femoral neck fractures. A retrospective study of patients with femoral neck fractures was performed. One thousand sixty-three AP hip radiographs were obtained from 550 patients. Ground truth labels of Garden fracture classification were applied as follows: (1) 127 Garden I and II fracture radiographs, (2) 610 Garden III and IV fracture radiographs, and (3) 326 normal hip radiographs. After localization by an initial network, a second CNN classified the images as Garden I/II fracture, Garden III/IV fracture, or no fracture. Advanced data augmentation techniques expanded the training set: (1) generative adversarial network (GAN); (2) digitally reconstructed radiographs (DRRs) from preoperative hip CT scans. In all, 9063 images, real and generated, were available for training and testing. A deep neural network was designed and tuned based on a 20% validation group. A holdout test dataset consisted of 105 real images, 35 in each class. Two class prediction of fracture versus no fracture (AUC 0.92): accuracy 92.3%, sensitivity 0.91, specificity 0.93, PPV 0.96, NPV 0.86. Three class prediction of Garden I/II, Garden III/IV, or normal (AUC 0.96): accuracy 86.0%, sensitivity 0.79, specificity 0.90, PPV 0.80, NPV 0.90. Without any advanced augmentation, the AUC for two-class prediction was 0.80. With DRR as the only advanced augmentation, AUC was 0.91 and with GAN only AUC was 0.87. GANs and DRRs can be used to improve the accuracy of a tool to diagnose and classify femoral neck fractures.

Keywords Deep learning · Artificial intelligence · AI · Femur · Fracture

Introduction

Proximal femur fractures, including osteoporotic femoral neck fractures, are common injuries and a major source of morbidity and mortality in the elderly population [1–3]. Accurate and timely diagnosis of a femoral neck fracture is essential for therapeutic decision-making and delay in surgical repair can lead to increased morbidity and mortality [4, 5]. The radiograph-based Garden classification system of femoral neck fractures is well-established and has been used for years to dictate management [6, 7]. While diagnosing a displaced Garden III or IV fracture is usually trivial, more subtle type I and II fractures can be challenging for a trainee, a clinician, or a radiologist without subspecialized experience in musculoskeletal imaging. An automated tool which would prioritize positive femoral neck fracture cases on a radiology worklist and provide a second opinion to the interpreting physician would be a valuable addition to the emergency radiology workflow.

Deep learning, a branch of machine learning, has emerged in recent years as a powerful statistical tool to address a range of real-life problems including biometric recognition, natural language processing, and autonomous driving. A specific deep learning construct commonly used for image recognition tasks is the convolutional neural network (CNN). This algorithmic technique allows a system to automatically extract features useful for a specific domain problem without explicit human instruction [8]. Insufficient quantity of training data is a limiting factor in the training of CNNs and novel data augmentation techniques which address this limitation continue to evolve.

With the rise of deep learning, there has been an understandable growing interest to harness its potential in healthcare, including in radiology, with medical image recognition as one of its many potential applications [9, 10]. The

✉ Simukayi Mutasa
stmutasa@gmail.com

[1] Columbia University Irving Medical Center, 622 West 168th Street, PB 01-301, New York, NY 10032, USA

ability of an algorithm to iteratively learn meaningful patterns from the input data gives it the potential to recognize features of images not apparent to human visual inspection. An appropriately trained network could, therefore, augment the work of the radiologist in rendering certain diagnoses, including subtle fractures. A fracture-recognition deep learning algorithm could also be used in the emergency room setting to "triage" positive femoral neck fracture cases for rapid definitive interpretation, thereby decreasing the time to intervention. While detection of proximal femur fractures, including intertrochanteric fractures, using deep learning had recently been described [11–13], to the best of our knowledge this study is the first to use deep learning with advanced data augmentation techniques to diagnose and classify femoral neck fractures.

Materials and Methods

Data Acquisition

An IRB-approved retrospective case-control study of patients with femoral neck fractures was performed. A search of 97,128 hip radiographs performed at our institution between February 2000 and February 2017 was undertaken. Emergency room adult hip radiographs that mentioned a femoral neck fracture within the "Impression" section of the radiology report yielded a potential 1444 hip radiographs from 1195 patients. The anteroposterior (AP) radiographs of the hip were extracted as DICOM files from the PACS system. Ground truth labels where applied by a board-certified, musculoskeletal fellowship-trained radiologist (MJR) who confirmed the presence of a femoral neck fracture, while excluding any images containing hip hardware, fractures not involving the femoral neck (including intertrochanteric fractures and subtrochanteric fractures), and equivocal fractures. Radiographs of various image quality were included, as long as they were considered diagnostic by the reviewing attending radiologist (MJR). A normal group was constructed using images of the contralateral non-fractured hip, when available. The final dataset contained 1063 unique anteroposterior radiographs of the right or left hip from 550 patients, classified into 3 groups: Garden I/II fracture ($n = 127$), Garden III/IV fracture ($n = 610$), and normal ($n = 326$). Contralateral hips were excluded if they contained orthopedic hardware. A class balanced holdout set consisting of 105 images from 105 unique patients, 35 images from each group, was set aside as the test dataset. Of the 550 patients who contributed to the final dataset, 352 were female (64%) and 198 male. Patient age range was 23–107 years old with a mean age of 75 and a standard deviation of 17 years. Patients with a fracture ($n = 217$) comprised 151 females and 66 males with a mean age of 79 and a standard deviation of 13 years.

Radiographs used in this study were performed on digital radiography equipment from General Electric, Siemens, and Philips. Acquisition parameters were automatically adjusted based on patient exposure, as per routine.

Data Annotation and Model Development

Annotation of femoral neck regions was performed on each of the 1063 radiographs in the final dataset to facilitate training. This was accomplished by using an open-source software tool (DICOM Web Viewer version 0.25.2) to draw and record a circular region-of-interest (ROI) centered on the femoral neck on each deidentified radiograph.

Once data annotation was completed, two networks were implemented. The first network was trained to localize the femoral neck on an AP radiograph and generate 850 × 850 pixel crops from each raw image. The second network was trained to classify the femoral neck into a Garden I/II, Garden III/IV, or no fracture group based on downscaled 256 × 256 inputs from the localizer. For the localization network, 300 of the femoral neck location annotations were utilized for training with the remaining radiographs used to test and fine-tune performance. For the fracture classification network, data was split into training, validation, and test datasets. The training set is the set of images from which the network learns. The validation set is a shuffled subset of the training set used during the learning process to optimize the neural network parameters. After completion of training and fine-tuning, the network performance was measured on a test dataset containing sequestered images that the network had never encountered before.

Data Augmentation

To provide additional training data for the network, data augmentation techniques were employed. First, by exposing the network to multiple small variations of each radiograph, the network develops the ability to marginalize random noise and detect patterns important for diagnosis rather than focus on memorizing images. These classic data augmentation techniques, detailed in Appendix A, have the potential to expose the network to 3.2×10^8 possible variations of each input image. Second, 34 of the 550 patients underwent CT of the fractured hip prior to surgical repair. These CTs were used to generate 6000 images utilizing the digitally reconstructed radiograph technique (DRR) (Fig. 1). The generated images were used as extra input data to the final classification network. Finally, the deep convolutional generative adversarial network (GAN) used both real and digitally reconstructed radiographs to generate additional 2000 images that were saved and used as extra input data to the final classification network.

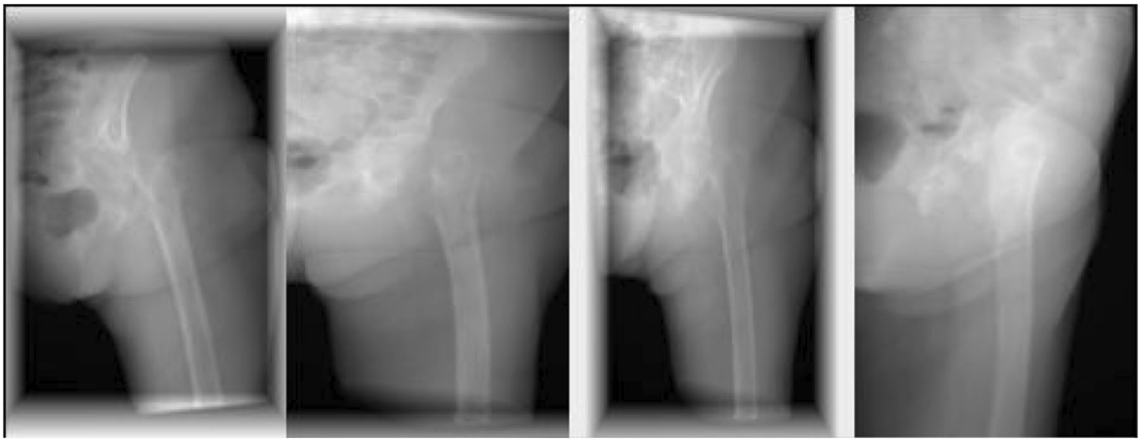

Fig. 1 Augmented DRRs. These radiographic projections were generated from a single CT scan with multiple small rigid warps and various simulated radiographic acquisitions applied

CNN Architecture and Testing

Classifier network inputs consisted of 256 × 256 pixel bounding boxes. The CNN was based on a novel 2D neural network utilizing a customized residual network based architecture (Appendix Table 2, Fig. 2). Runtime regularization techniques were used including L2 regularization to limit the square magnitude of weights, and dropout to limit unit co-adaptation. A softmax loss function with four classes (normal, Garden II/III, Garden III/IV, and GAN Fake) was utilized to generate final network outputs. Raw outputs of the network were four positive or negative numbers that were interpreted as the un-normalized logarithmic odds for each class. These numbers were normalized through the softmax function, which allowed us to interpret the outputs as class probabilities. For the test dataset, which did not contain any GAN or DRR inputs, softmax score indicating the highest class probability was chosen as the accepted class. In the case of a network prediction where the GAN Fake class had the highest probability, the second-highest predicted class was used as the network prediction. Network hyperparameters were fine-tuned based on performance on a validation set comprised of 20% of the training samples. After final hyperparameters were fine-tuned on the validation set, the network was run once on the sequestered test dataset and performance was recorded.

The primary performance metric for the network was the multi-class aggregated area under the receiver operating curve (AUC). In addition, accuracy, aggregate "one versus all" sensitivity and specificity, and aggregate "one versus all" positive predictive value and negative predictive value were calculated. Further network testing and implementation details are provided in Appendix B.

Learning Visualizations

While deep learning has long been considered a "black box" technique, several methods have been proposed to allow visualization of the inner workings of a network. We utilized

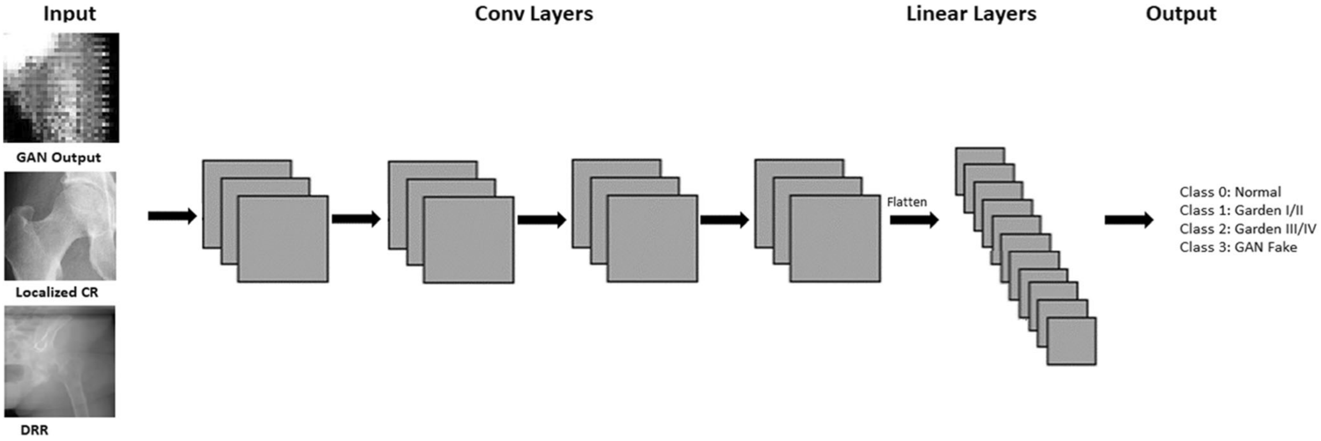

Fig. 2 Training network architecture: Three different inputs are utilized. Residual convolutions are utilized with two embedded convolutional operations followed by batch normalization and ReLu nonlinearity. Downsampling is achieved by strided convolutions. No GAN or DRR inputs were used for analyzing the test dataset. In the case of a network prediction where the GAN Fake class had the highest probability, the second highest predicted class was used as the network prediction

one such method, described by Selvaraju et al. [14], called gradient-weighted class activation mapping (GRAD-CAM). By analyzing the gradients flowing into the final convolutional layer, this method highlights the important regions of the input image that are used by the network in obtaining final predictions. Additionally, we employed guided backpropagation, initially described by Springenberg et al. [15] in order to highlight all contributing features of an input image which are used by the network to make a final decision. Figure 3 illustrates these techniques.

Results

When distinguishing between a fracture of any Garden classification and a normal radiograph in the test dataset, the network accuracy was 92%. Sensitivity (SN) and specificity (SP) were 0.91 and 0.93 respectively. Positive predictive value (PPV) and negative predictive value (NPV) were 0.96 and 0.86 respectively and area under the receiver operating curve (AUC) was 0.92. For three class discrimination between Garden I + II versus Garden III + IV versus normal radiographs, network accuracy was 86%.

"One versus all" sensitivity and specificity were 0.79 and 0.90 respectively with subsequent PPV and NPV of 0.80 and 0.90 and AUC of 0.96. One versus all statistical testing for multiple classes involves counting the correct class as the positive class and any other guess as the negative class. Class-specific statistics are summarized in Table 1.

For Garden I/II fractures, SN, SP, PPV, and NPV were 0.54, 0.93, 0.79, and 0.80, respectively, with accuracy of 80%. For Garden III/IV fractures, SN, SP, PPV, and NPV were 0.91, 0.83, 0.73, and 0.95 respectively, with accuracy of 86%. When comparing network performance for discrimination between normal radiographs and nondisplaced Garden I/II fractures, SN, SP, PPV, and NPV were 0.80, 0.94, 0.91, and 0.94, respectively, with an accuracy of 88%.

Without any advanced augmentation, the network AUC for distinguishing between a fracture of any kind and a normal radiograph was 0.80. With DRR outputs as the only advanced augmentation, AUC was 0.91. With the GAN outputs as the only advanced augmentation, AUC was 0.87.

The network misclassified 15 of 105 cases. Fourteen involved the network missing a classification by one class (for example, predicting class 0 instead of class 1 or class 1 instead of class 2). Of these, eight misses involved failure to distinguish normal study and a nondisplaced fracture. Examples of network misses are shown in Fig. 4.

Discussion

The results of this study support the hypothesis that a series of convolutional neural networks can be trained to differentiate

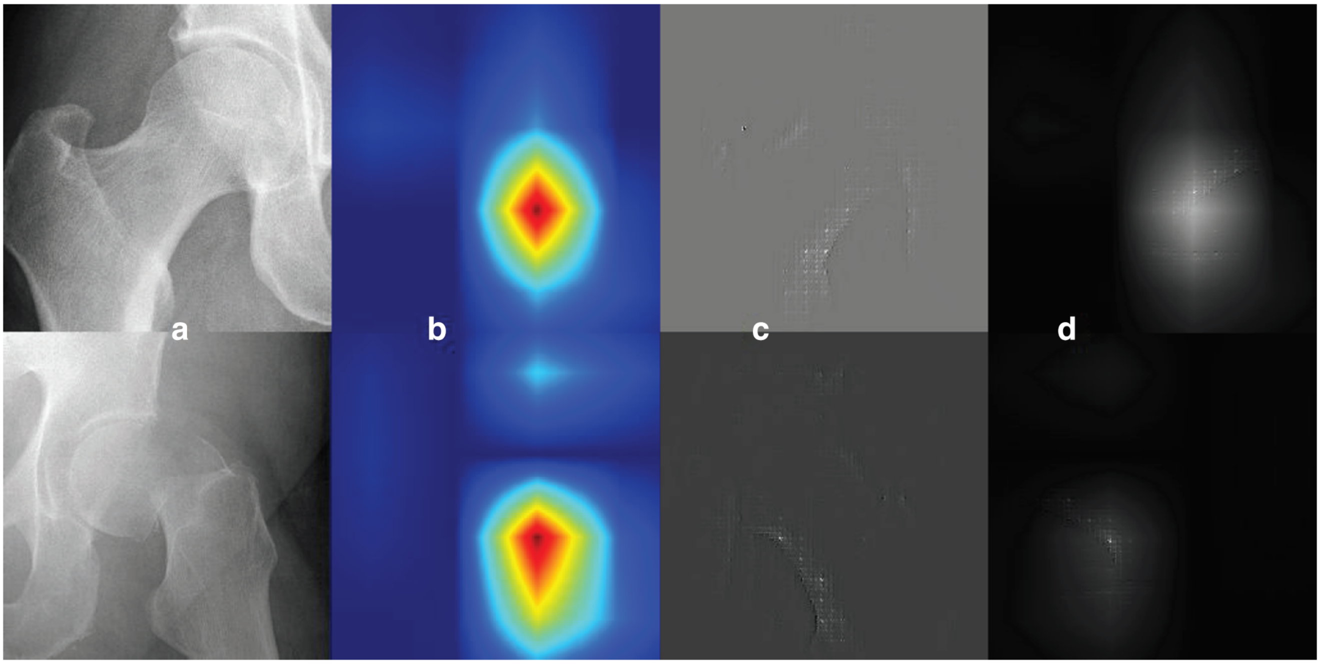

Fig. 3 Saliency maps of two different patients represented by the two rows. **a** The input images after localization of the femoral neck by the localization network. **b** Grad-CAM highlighting the input regions of interest utilized in generating a positive identification for a specific class. **c** Guided backpropagation results highlighting every pixel utilized in making a decision, positive or negative. **d** Guided Grad-CAM results which combine the previous two methods to highlight the pixels the network uses to make a positive prediction of a specific class

Table 1 Class specific statistics on the test dataset which contained 105 patients, 35 in each class. "Normal" refers to network performance when detecting the normal radiograph class versus all other classes. "G I/II" refers to network performance for detecting a nondisplaced fracture class (Garden I or II fractures) versus any other class. "G III/IV" refers to network performance for detecting a displaced fracture (Garden III or IV) versus any other class. Three class average AUC was 0.96

	Accuracy	SN	SP	PPV	NPV
Normal	0.92	0.91	0.93	0.86	0.96
G I/II	0.80	0.54	0.93	0.79	0.80
G III/IV	0.86	0.91	0.83	0.73	0.95

between no femoral neck fracture, Garden I/II fracture, and Garden III/IV fracture on radiographs with good accuracy. We also show that the femoral neck can be automatically localized on raw input AP radiographs and this information can be fed automatically into the classification network. Finally, we demonstrate that two advanced data augmentation techniques, digitally reconstructed radiographs (DRRs) and generative adversarial networks (GANs), can be used for radiography-based deep learning projects.

The sensitivity of the network for detecting and correctly classifying nondisplaced Garden I/II fractures was lower than for Garden III/IV fractures. However, only one-third of the misses from this group involved incorrectly labeling a hip fracture as having no fracture. The remaining misses from this group involved predicting a displaced fracture in the cases when there was a nondisplaced fracture. The relative difficulty in differentiating Garden I/II fractures from normal hip radiographs was expected as this diagnosis can be challenging with subtle cases even for a radiologist. Fourteen of the 15 misses by the network involved a predicted classification that was one grouping off from the human annotated label. There was only one instance where the network misclassified a fracture by a difference of more than one class, where it predicted a class 2 displaced fracture when there was no fracture seen by the human annotator. This miss may have been due to a calcified granuloma partially overlying the femoral neck. The relatively lower sensitivity of the network to detect nondisplaced fractures may be addressed in a future implementation of the network by training on a larger input matrix, as downscaling the input images may obscure subtle fracture lines and cortical irregularities. This would require a larger training dataset as using large input sizes on small datasets leads to overfitting of the model.

Our testing dataset contained a class balanced population with 35 AP hip radiographs each in class 0, 1, and 2. In a real-world setting, where fractures on hip radiographs are relatively uncommon even in patients in the emergency department after hip trauma, this would lead to a reduction of actual PPV for detecting a fracture when compared to that calculated based on our test dataset. However, real-life negative predictive value would likely be higher. Ground truth labelling of cases also potentially affects test dataset statistics. We utilized one human annotator for labelling of the Garden fracture class. It is possible that this may introduce a bias in borderline, equivocal cases. This potential pitfall could be mitigated by using an expert consensus methodology. Finally, the average age of the patients in our datasets is 75, which could bias the network to increase its pretest probability of detecting fractures when it recognizes imaging features more common in elderly patients.

Classical machine learning techniques, such as support vector machines, have been widely studied in radiologic imaging analysis and in other tasks such as segmentation of lesions and predicting patient outcomes [16–18]. The limitation of these older techniques is the need for operators to prespecify the variables to be extracted from the images, when in fact there may be many other "hidden" variables not apparent to human operators but superior for the classification task at hand. With deep learning, convolutional neural networks are able to classify data through a hierarchical process without pre-specifying discriminating image variables.

Two major limitations of deep learning are the need for large datasets to prevent overfitting and the non-transparency associated with feature selection by the CNN. The first may be overcome through data augmentation and generation where data is manipulated to artificially enlarge the dataset. We chose to assess for femoral neck fractures on the AP hip radiograph as this view is consistently obtained in the trauma setting. Including lateral hip radiographs in future

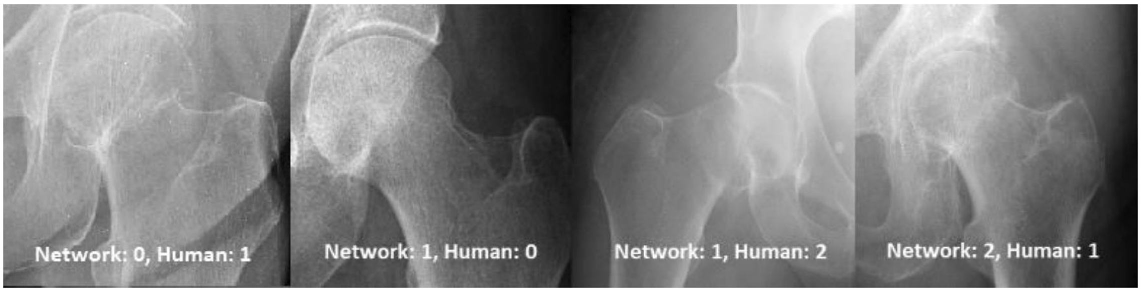

Fig. 4 Example images of network misses. Representative examples of the misclassified fractures from the network. Class 0 represents cases with no fracture, class 1 represents nondisplaced Garden I/II fractures, and class 2 represents displaced Garden III/IV fractures

iterations of the network would provide more data to train the network. The non-transparency limitation can be addressed by utilizing techniques such as (1) guided backpropagation which display the pixels on the input image that the network weights the most when making a final classification; and (2) class activation mapping techniques that highlight regions of interest in the input pictures which contribute to positive identifications of a class. This study used both data augmentation and visualization techniques to mitigate the inherent limitations of deep learning.

Using the guided backpropagation and the gradient-weighted class activation mapping techniques (Fig. 3), we can infer that the network found it important to focus most on the medial aspect of the femoral neck. An additional consistent point of focus was on the lateral aspect of the acetabular rim. The network may have found the relationship between the femoral neck and acetabular rim important for positive identification of a displaced femoral neck fracture.

The decision to use a localizer neural network to hone in on the femoral neck, rather than using the full AP radiograph, was made for two reasons. First, there is a theoretical improvement in training time and performance over utilizing full images. This is because the network does not have to waste computations deciding that likely extraneous features outside the femoral neck, such as those in the pelvis, are irrelevant to the problem of a femoral neck fracture. Additionally, we hypothesize that this would improve regularization accuracy since Zech et al. [19] showed that spurious details such as the style of side marker used in specific hospitals can bias neural network predictions. By automatically cropping to the femoral neck, we could minimize the effect of these irrelevant details in the model and focus the CNN on femoral neck texture and morphology.

In the literature, machine learning techniques have been previously applied to detect proximal femur fractures [11, 12, 20]. The study from Tian et al. [16] relied on handcrafted features, in this case, femoral neck-shaft angle, for detection of hip fracture. Although handcrafted techniques have shown promise, they suffer from a variety of drawbacks related to pre-specified feature extraction which are mitigated by deep learning techniques. Gale et al. [11] applied a system of CNNs to detect all types of hip fractures, including extracapsular fractures. Urakawa et al. [20] focused on intertrochanteric fracture detection and achieved slightly higher accuracy of fracture detection by a CNN than our study. This may in part be attributable to a generally larger image region containing pathology in intertrochanteric fractures which facilitates learning. Our study focused on detecting and classifying only femoral neck fractures which undergo a specific surgical intervention depending on the type.

The importance of this work is threefold: (1) artificial intelligence tools have a potential role in triaging emergency radiology worklists. Femoral fractures should be urgently evaluated and a fracture detection/classification algorithm can expedite final image interpretation and treatment; (2) deep learning algorithms can augment the work of the radiologist and function as a second subspecialist reader. Prior studies such as the one from Dominguez et al. [21] reported a 4.4% miss rate of fractures on the initial radiograph. Thus, these algorithms can help to confirm pathology which may be challenging to a trainee or to a non-specialty trained radiologist or suggest the use of advanced imaging in equivocal cases; and (3) advanced data augmentation techniques demonstrated in this study, GANs and DRRs, can be employed to mitigate the limitations of small datasets in medical image recognition.

Conclusion

Deep learning using a CNN is able to predict the presence of femoral neck fractures on AP radiographs with good accuracy. Furthermore, a CNN can learn to differentiate nondisplaced Garden I/II fractures from displaced Garden III/IV fractures. Several techniques can be employed when faced with small datasets to improve neural network performance. Future research may focus on expanding the dataset, prospective validation of the algorithm, and testing on additional neural network architectures.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Data Augmentation

Data augmentation used in this study entailed several real-time modifications to the source images at the time of training. These modifications included (1) random horizontal and vertical flipping of the input image; (2) random rotation of the input image by − 30° to 30°; and (3) random contrast jittering of the input image and addition of a random Gaussian noise matrix, performed to simulate different acquisition parameters for each image. This resulted in roughly an additional 3.2×10^8 variations of each input image.

For further data augmentation, 6000 digitally reconstructed radiographs (DRRs) were generated. DRR volume rendering, also called simulated x-ray volume rendering, is a direct volume rendering technique that consists of simulating x-rays passing through the reconstructed CT volume based on the absorption properties of the tissue. DRR generation is a popular technique in simulating radiation therapy treatments. For this project, we recreated a cone-beam radiographic acquisition with a ray-tracing DRR generating algorithm utilizing the thin slice acquisitions, where available. Forty-five hip CT volumes from 34 patients (11 patients had a pelvic CT scan performed which contributed two hips) were obtained that were performed within 3 days of the hip radiographs but before surgery. Using 3D input volumes allowed for a far greater range of augmentation capabilities. We applied rigid affine transformation and multiple slightly different acquisition parameters, allowing us to effectively turn one input example into multiple slightly different radiographs. We were able to take one patient volume and simulate (1) making the patient slightly thinner, larger, taller or shorter; (2) varying radiographic KvP, mA, and source to object distance; and (3) simulating internal/external rotation of the hip and minor variations in apical angulation of the patient.

Additional training examples were generated utilizing a generative adversarial network (GAN). GANs are a family of deep generative models which balance training of a generative network and a discriminative network in order to generate realistic examples based on a training set data distributions. To summarize our design choices in this paper, a deep convolutional GAN with a residual network generator was utilized. Earthmover distance based on the paper by Arjovsky et al. [22] was used as the discriminatory function with gradient clipping. The GAN was trained for 300 epochs which took 72 h on the research computer until convergence. Then, 2000 generated outputs were subsequently used as an additional input into the network. All artificially generated images (using DRR and GAN techniques) were utilized as additional data for training only, and not for testing or validation.

Neural Network Implementation Details

The overall network architecture is shown in Table 1 and Fig. 2. The CNN was implemented by a series of 3 × 3 convolutional kernels to maximize computational efficiency while preserving nonlinearity [23]. After an initial standard convolutional layer, a series of residual layers are utilized in the network. Originally described by He et al. [24], residual neural networks can stabilize gradients during backpropagation, leading to improved optimization and facilitating greater network depth. A spatial transformer module was inserted after the 11th hidden layer. Initially applied to convolutional neural networks by Jaderberg et al. [25], spatial transformer layers allow a network to explicitly learn affine transformation parameters that regularize global spatial variations in feature space. This is important for the network to be robust against significant variations in imaging technique or positioning commonly seen in practice.

Downsampling of feature map size was implemented by means of strided convolutions. All nonlinear functions utilize the rectified linear unit (ReLU) which allows training of deep neural networks by stabilizing gradients on backpropagation [26]. Additionally, batch normalization was used between the convolutional and ReLU layers to prevent covariate shift [27]. Upon downsampling, the number of feature channels is doubled, preventing a representation bottleneck. Dropout with a keep probability of 75% was applied to the first fully connected layer to limit over-fitting and add stochasticity to the training process [28].

In addition to the customized network described above, several additional network architectures were tested. This includes (1) ResNet 52 network architecture initialized both randomly and with pre-trained weights from Imagenet; (2) custom-built networks, initialized from random weights, and with varying numbers of convolutional layers based on the Inception v4 architecture; and (3) 100 layer network based on a randomly initialized DenseNet architecture. Performance for the networks was best when initializing weights randomly across the board. We found that when using

Table 2 Network architecture: Dimensions of all the intermediate layers of the convolutional neural network. Residual layers contain two feature maps per layer

Input Layer	Input layer dimensions	Filter type	Filter size	Output layer
Input	256 × 256 × 1	Convolutional	3 × 3 × 16	Hidden layer 1
Hidden layer 1	128 × 128 × 16	Residual	3 × 3 × 32	Hidden layer 2/3
Hidden layer 2/3	64 × 64 × 32	Residual	3 × 3 × 64	Hidden layer 4/5
Hidden layer 4/5	32 × 32 × 64	Residual	3 × 3 × 128	Hidden layer 6/7
Hidden layer 6/7	16 × 16 × 128	Residual	3 × 3 × 128	Hidden layer 8/9
hidden layer 8/9	16 × 16 × 128	Residual	3 × 3 × 256	Hidden layer 10/11
Hidden layer 10/11	8 × 8 × 256	Spatial Transform	N/A	Hidden layer 12
Hidden layer12	8 × 8 × 256	Inception	×256	Hidden layer 13
Hidden 13	8 × 8 × 256	Residual	3 × 3 × 512	Hidden layer 14/15
Hidden layer 14/15	4 × 4 × 512	Residual	3 × 3 × 512	Hidden layer 16/17
Hidden layer 16/17	4 × 4 × 512	Residual	3 × 3512	Hidden layer 18/19
Hidden layer 18/19	4 × 4 × 512	Linear	× 16	Hidden layer 20
Hidden layer 20	1 × 16	Linear	16 × 8	Hidden layer 21
Hidden layer 21	1 × 8	Softmax	8 × 3	Classification

greater than 14 hidden layers in two dimensional residual networks overfitting occurred.

Training was implemented using the parameterized Adam optimizer, combined with the Nesterov accelerated gradient described by Dozat [29]. Parameters were initialized to equalize input and output variance utilizing the heuristic described by Glorot et al. [13]. L2 regularization was implemented to prevent overfitting of data by limiting the squared magnitude of the kernel weights. Final hyperparameter settings included a learning rate set to 1e-3, keep probability for dropout of 50%, moving average weight decay of 0.999, and L2 regularization weighting of 1e-4.

Software code for this study was written in Python (v3.5) using the TensorFlow module (v1.5). Experiments and CNN training were performed on a Linux workstation with NVIDIA Titan X Pascal GPU with 12 GB on chip memory, i7 CPU and 32 GB RAM. The classification network was trained for 200 epochs, taking 5 h. Inference time was 22 s per image for localization and classification. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Papadimitriou N, Tsilidis KK, Orfanos P, Benetou V, Ntzani EE, Soerjomataram I, et al. Burden of hip fracture using disability-adjusted life-years: a pooled analysis of prospective cohorts in the CHANCES consortium. Lancet Public Heal [Internet]. 2017 [cited 2018 Sep 13];2(5):e239–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29253489
2. Kani KK, Porrino JA, Mulcahy H, Chew FS. Fragility fractures of the proximal femur: review and update for radiologists. Skeletal Radiol [Internet]. 2018 Jun 29 [cited 2018 Sep 13]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/29959502
3. Sozen T, Ozisik L, Calik Basaran N. An overview and management of osteoporosis. Eur J Rheumatol [Internet]. 2017 [cited 2018 Sep 13];4(1):46–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28293453
4. Natasha Morrissey, Efthymios Iliopoulos, Ahmad Wais Osmani, Kevin Newman. Injury, Int. J. Care Injured. 2017; 48: 1155–1158.
5. Ryan DJ, Yoshihara H, Yoneoka D, Egol KA, Zuckerman JD. Delay in hip fracture surgery. J Orthop Trauma [Internet]. 2015 [cited 2018 Sep 13];29(8):343–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25714442
6. Garden RS. Low-angle fixation in fractures of the femoral neck. J Bone Joint Surg Br [Internet]. The British Editorial Society of Bone and Joint Surgery; 1961 1 [cited 2018 Sep 13];43–B(4):647–63. Available from: https://doi.org/10.1302/0301-620X.43B4.647
7. Florschutz A V., Langford JR, Haidukewych GJ, Koval KJ. Femoral Neck fractures. J Orthop Trauma [Internet]. 2015 [cited 2018 Sep 13];29(3):121–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25635363
8. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE [Internet]. 1998 [cited 2018 13];86(11):2278–324. Available from: http://ieeexplore.ieee.org/document/726791/
9. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature [Internet]. Nature Publishing Group; 2017 25 [cited 2018 Sep 13];542(7639):115–8. **Available from:** http://www.nature.com/articles/nature21056
10. Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J Digit Imaging [Internet] 2017 Aug 10 [cited 2018 Sep 13];30(4):477–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28695342
11. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. 2017 [cited 2018 Sep 13]; Available from: http://arxiv.org/abs/1711.06504
12. Kazi A, Albarqouni S, Sanchez AJ, Kirchhoff S, Biberthaler P, Navab N, et al. Automatic classification of proximal femur fractures based on attention models. In Springer, Cham; 2017 [cited 2018 Sep 13]. p. 70–8. Available from: https://doi.org/10.1007/978-3-319-67389-9_9
13. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [Internet]. [cited 2018 Sep 13]. Available from: http://www.iro.umontreal.
14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2016 [cited 2018 Sep 13]; Available from: http://arxiv.org/abs/1610.02391
15. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. 2014 21 [cited 2018 Sep 13]; Available from: http://arxiv.org/abs/1412.6806
16. Tian TP, Chen Y, Leow WK, Hsu W, Howe T Sen, Png MA. Computing neck-shaft angle of femur for x-ray fracture detection. In Springer, Berlin, Heidelberg; 2003 [cited 2018 Sep 13]. p. 82–9. **Available from:** https://doi.org/10.1007/978-3-540-45179-2_11
17. Zhou J, Chan KL, Chong VFH, Krishnan SM. Extraction of brain tumor from MR images using one-class support vector machine. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference [Internet]. IEEE; 2005 [cited 2018 Sep 13]. p. 6411–4. Available from: http://ieeexplore.ieee.org/document/1615965/
18. Hu X, Wong KK, Young GS, Guo L, Wong ST. Support vector machine multiparametric MRI identification of pseudoprogression from tumor recurrence in patients with resected glioblastoma. J Magn Reson Imaging [Internet]. 2011 [cited 2018 Sep 13];33(2):296–305. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21274970
19. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. 2018 [cited 2018 Sep 13]; Available from: http://arxiv.org/abs/1807.00431
20. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol [Internet]. 2019 [cited 2019 Jan 16];48(2):239–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29955910
21. Dominguez S, Liu P, Roberts C, Mandell M, Richman PB. Prevalence of traumatic hip and pelvic fractures in patients with suspected hip fracture and negative initial standard radiographs—a study of emergency department patients. Academic emergency medicine. 2005 Apr;12(4):366-9.

Appendix References

22. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN 2017 Jan 26 [cited 2018]; Available from: http://arxiv.org/abs/1701.07875

23. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved Training of Wasserstein GANs Montreal Institute for Learning Algorithms [Internet]. [cited 2018 Sep 13]. Available from: https://github.com/igul222/improved_wgan_training.
24. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2015 10 [cited 2018 Sep 13]; Available from: http://arxiv.org/abs/1512.03385
25. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial Transformer Networks. 2015 5 [cited 2018 Sep 13]; Available from: http://arxiv.org/abs/1506.02025
26. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going Deeper with Convolutions. 2014 16 [cited 2018 Sep 13]; Available from: http://arxiv.org/abs/1409.4842
27. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015 [cited 2018 Sep 13]; Available from: http://arxiv.org/abs/1502.03167
28. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res [Internet]. 2014 [cited 2018 13];15:1929–58. Available from: http://jmlr.org/papers/v15/srivastava14a.html
29. Dozat T. Incorporating Nesterov Momentum into Adam [Internet]. [cited 2018 Sep 13]. Available from: http://mattmahoney.net/dc/text8.zip

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.