HHS Public Access

Author manuscript

Eur J Radiol. Author manuscript; available in PMC 2021 September 01.

Published in final edited form as:

Eur J Radiol. 2020 September ; 130: 109139. doi:10.1016/j.ejrad.2020.109139.

Deep learning evaluation of pelvic radiographs for position, hardware presence, and fracture detection.

Gene Kitamura, M.D.[a]

[a]University of Pittsburgh Medical Center (UPMC) and University of Pittsburgh., UPMC Department of Radiology., 200 Lothrop St., UPMC Montefiore, Room NE 538, Pittsburgh, PA 15213

Abstract

Purpose: Recent papers have shown the utility of deep learning in detecting hip fractures with pelvic radiographs, but there is a paucity of research utilizing deep learning to detect pelvic and acetabular fractures. Creating deep learning models also requires appropriately labeling x-ray positions and hardware presence. Our purpose is to train and test deep learning models to detect pelvic radiograph position, hardware presence, and pelvic and acetabular fractures in addition to hip fractures.

Material and Methods: Data was retrospectively acquired between 8/2009 to 6/2019. A subset of the data was split into 4 position labels and 2 hardware labels to create position labeling and hardware detecting models. The remaining data was parsed with these trained models, labeled based on 6 fracture patterns, and fracture detecting models were created. A receiver operator characteristic (ROC) curve, area under the curve (AUC), and other output metrics were evaluated.

Results: The position and hardware models performed well with AUC of 0.99 – 1.00. The AUC for proximal femoral fracture detection was as high as 0.95, which was in line with previously published research. Pelvic and acetabular fracture detection performance was as low as 0.70 for the posterior pelvis category and as high as 0.85 for the acetabular category.

Conclusion: We successfully created deep learning models that can detect pelvic imaging position, hardware presence, and pelvic and acetabular fractures with AUC loss of only 0.03 for proximal femoral fracture.

Keywords

Deep learning; artificial intelligence; machine learning; radiographs; fracture

Corresponding author: Gene Kitamura, kitamurag@upmc.edu, GitHub: https://github.com/GeneKitamura/machine_learning_pelvis, Phone: 412-648-6062, Fax: 412-692-2615.

CRediT author statement:

Gene Kitamura: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Funding acquisition

The authors do not have any conflict of interest to disclose.

Introduction

Hip fractures are a significant cause of morbidity and mortality in the world with a high economic cost[1,2]. Among elderly patients older than 65 years old, nearly a third will suffer a fall each year with approximately 10–15% of the cases resulting in a fracture[3]. Up to a third of the patients are admitted to nursing facilities within a year of hip fractures, and in these cases, the one year fatality rate exceeds 20%[4,5]. Although not as common, pelvic fractures account for approximately 3% of musculoskeletal injuries with mortality rates of approximately 13%[6,7]. Acetabular fractures are rare, only accounting for about 3 fractures per 100,000 trauma patients, with about 80% occurring after high energy trauma and another 10% occurring from falls in elderly patients[8,9]. Pelvic radiographs are often the initial imaging modality of choice in evaluating fractures, but miss rates have been estimated at approximately 10%[10,11]. Advanced imaging such as computed tomography (CT) and magnetic resonance imaging (MRI) scans can be used to reduce misses, but they cost more, take longer, and are less available[12].

A variety of prior research has described the effectiveness of deep learning in evaluating orthopedic radiographs for extremity fractures[13–15] and bone age[16,17]. In terms of pelvic radiographs, some studies have explored using deep learning for osteoarthritis evaluation[18] and forensic age estimation[19]. Furthermore, several other papers have demonstrated the utility of deep learning with pelvic radiographs for hip fracture detection, specifically evaluating proximal femoral and intertrochanteric fractures[20–23]. By isolating the proximal femur using bounding boxes, one study achieved an AUC of 0.98 for proximal femoral fractures, but without a visualization method such a saliency map or heat map to show the site of abnormality [23]. One study obtained a proximal femoral fracture AUC of 0.98 without using bounding boxes, but only included 100 cases in the test dataset[20]. Another study using bounding boxes demonstrated a proximal femoral fracture AUC of 0.97 but used a vaguely described heat-map visualization method[24]. Finally, one other study obtained an AUC of 0.99 for hip fracture detection with bounding boxes, but did not have any images or visualization methods in the manuscript[21]. Despite several studies assessing deep learning for hip fracture detection, there is a lack of studies evaluating deep learning for pelvic and acetabular fracture detection.

In building a pelvic, acetabular and hip fracture classification deep learning model, initial data cleaning would require a fair amount of effort. Hip x-rays are often acquired with pelvic x-rays, but they are difficult to sort based on the Digital Imaging and Communications in Medicine (DICOM) headers of archival data, as they are often mislabeled or incomplete. Rarely, a completely different study, such as a chest x-ray, may even be associated with the accession of a pelvic x-ray. Finally, depending on the study design, an investigator may want to exclude cases with existing hardware, whether they are arthroplasties or fixation devices.

Therefore, we decided to investigate whether a deep learning model can be constructed to automatically identify pelvic x-ray positions, identify hardware, and detect pelvic and acetabular fractures in addition to hip fractures.

Material and Methods

An institutional review board approval was obtained for this retrospective study with waiver of consent secondary to minimal risk. The case query was between 8/2009 to 6/2019 for patients over 18 years old in the emergency and inpatient setting. All images were obtained consecutively from a vendor-neutral archive after querying pelvic x-rays for the term “fracture” in the impression section of radiology reports. A total of 7440 patients were identified with 14,374 images. Images and radiology reports were evaluated by the Board-certified musculoskeletal Radiologist first author with 3.5 years’ experience as of date of writing. The funding source was only used to obtain the data and had no other role in the study.

For the pelvic position and hardware identification portion of the study, to maximize the proportion of positive hardware cases, we first queried our dataset for hardware terms in the impression of the radiology report, identifying 168 patients, then randomly selected 1009 additional patients for a total of 1177 patients with 2852 images. After manually evaluating the images, 27 images were excluded due to suboptimal quality due to technical factors, ending up with 1175 patients and 2825 images. In our institution, the most commonly mingled study with pelvic radiographs is a chest x-ray, as a trauma series encompasses both chest and pelvic x-rays; therefore, 200 frontal chest x-ray images were also acquired from 200 unique patients. Four position labels were created: pelvis, hip, fail, and chest (CXR). The pelvis label included frontal, oblique, inlet and outlet views of the pelvis. The hip label accounted for frontal, frog-leg lateral, and cross-table lateral views of the hip. The fail position accounted for suboptimal images which did not adequately image the pelvis or hip. Each image was marked as either no hardware or positive hardware. Positive hardware labels were assigned to images with any type of pelvic or hip hardware including arthroplasties, external fixation pins, and internal fixation hardware. A total of 2006 pelvic, 801 hip, and 18 fail positions along with 200 chest x-rays were available. Regarding hardware presence, 2507 had no hardware and 318 had positive hardware. Approximately 70% of the images were used for training and 30% of the images were held out of training as the test set. Equal proportions of each category were allocated to the training and testing datasets, and there was no patient overlap between the training and testing datasets.

For the fracture detection portion of the study, we utilized the remaining 6263 patients and 11522 images. We used the trained position model (which will be described in the Results section) to extract pelvic radiographs from our fracture dataset, resulting in 7520 pelvic radiographs. We then excluded any remaining hardware cases by passing these 7520 pelvic x-rays through the trained hardware model (which will be described in the Results section), ending up with 7357 pelvic x-rays without hardware.

Once the 7357 pelvic x-rays were checked by the first author, an additional 20 cases were excluded due to suboptimal quality. The final 7337 cases were categorized into one of 6 “separate” fracture categories by the first author: normal, anterior pelvis, posterior pelvis, proximal femur, acetabular, and complex. The anterior pelvis category included pubic symphysis diastasis and pubic rami fractures. The posterior pelvis category encompassed posterior iliac fractures, sacroiliac joint diastasis, and sacral fractures. The proximal femur

category included femoral neck, intertrochanteric and subtrochanteric fractures. The acetabular category comprised of acetabular fractures and femoral dislocations. A case was assigned the complex category when it demonstrated more than one type of fracture pattern. There were 3428 normal, 713 anterior pelvis, 123 posterior pelvis, 1902 proximal femur, 410 acetabular, and 761 complex cases. For each of these " different fracture categories, 70% of the cases were assigned to the training dataset and 30% were assigned to the test dataset. There was no patient overlap between the training and testing datasets.

Additional mixtures of the data were also created from the different" labels. A "pelvis consolidated" category was created combining the anterior and posterior pelvis categories into a pelvic ring class. A "femoral/acetabular consolidated" category combined the proximal femur and acetabular groups into a femur/acetabular class. A "femoral vs. non-femoral" category combined the anterior pelvis, posterior pelvis, and acetabular classes into a non-femoral class. Finally, a "binary split" was also created encompassing all abnormal cases.

We trained different models, one for each label and an additional one using the binary split.. For each of the models except the "binary split", we also evaluated the test dataset as a binary output by grouping all fracture types as abnormal and summing the probabilities to compare against the normal cases. In addition, to compare our study to previously published studies, we also trained a model evaluating normal vs. femoral fracture only. Finally, we evaluated each trained model with a test dataset only containing normal and proximal femoral fracture cases to compare performance among the models.

Deep learning models were created based on the Densenet-121 architecture[25] using Tensorflow version 1.12[26] and Python 3.6.2. The ImageNet weights were loaded to the model, the initial learning rate was set to 0.001, and an Adam optimizer was selected to decay the learning rate. A dropout layer was not incorporated. Data parsing was done using SciPy version 1.1.0[27] and Scikit-learn version 0.20.0[28]. Imaging data was extracted using Pydicom[29]. Each image was initially resized to 238 by 238 pixels, then augmented with random brightness, random contrast, random horizontal flipping, and random cropping to the final size of 224 by 224 pixels. Pixel intensity normalized to [−1,1] range. Training was monitored using the softmax cross-entropy loss and terminated when the loss no longer decreased after 100 batch-steps (Fig. 1). Training took approximately 1 day for each model. The models were trained on a GeForce 1080 GTX Graphical Processing Unit (GPU) on a Linux cluster.

Performance of the models were evaluated using a receiver operator characteristic (ROC) curve and area under the curve (AUC) of the held-out test data. Confidence intervals (CI) were set at 95% and noted as ± or +- value. Chi-square was used to evaluate sex distribution and a t-test was used to compare ages. AUC[30] and output metrics, including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)[31], were compared between models. Visual evaluation of the test set was done by comparing output probabilities of each class and checking for concordance between the images and the proposed labels. Both Grad-CAM and guided Grad-CAM heatmaps[32] were created for each fracture class to visualize the model attention.

Results:

As can be seen from the ROC curves of the position and hardware model (Fig 2), the models performed exceptionally well at classifying pelvic imaging position and hardware presence, with an AUC of 0.99 ± 0.01 and 1.00 ± 0.01, respectively. A sample of images with their probability outputs can be seen in Fig 3. The model's performance at detecting the 6 "fail" cases was expectedly low, only correctly labeling one of the cases; the "fail" images with their output probabilities is shown in Fig 4. Finally, we evaluated both models using an intersection of the position and hardware test dataset, resulting in an AUC of 0.99 ± 0.01 for both the models. The high AUC of both the position and hardware models allowed us to use these models to subsequently parse our fracture dataset as noted in the Material and Methods section.

The demographic of the fracture dataset is noted on Table 1; there was no statistical difference in age, but there was a statistical difference in sex distribution between the two groups. The ROC curves for the various models created for fracture detection, along with the AUC and 95% CI, are shown in Fig 5 and Table 2. The models performed the best with proximal femoral fracture classification, with an AUC between 0.93 and 0.95; the model trained only on normal vs. femoral fracture demonstrated the best performance with an AUC of 0.95. When each of the fracture classification models were evaluated as a binary output of normal vs abnormal cases, they performed as well as a model initially trained on a binary split, demonstrating an AUC between 0.85 and 0.86. For the fracture model, non-hip fracture detection performance was as low as 0.70 for the posterior pelvis category and as high as 0.85 for the acetabular category.

Output metrics of each trained model evaluated using a test dataset containing only normal and proximal femoral fracture cases are noted in Table 3. Sensitivity, specificity, PPV and NPV were calculated using the threshold value at the upper left corner of the ROC curves. Each of the metric values was compared to the output of the model only trained on normal vs. femoral fracture cases. As can be seen in Table 3, there were some statistically significant differences in the output metrics, but the AUC values were generally similar except for the "femoral/acetabular consolidated" model, which had the lowest AUC at 0.92 ($p<0.001$).

Discussion:

Initially, we successfully created deep learning models that can correctly categorize pelvic position and the presence of hardware, evidenced by a high AUC with the held-out test set. Understandably, the position model did not perform well with the "fail" cases since there was a paucity of training data with only 12 images. Increasing the performance of the "fail" class may be difficult since only a very small proportion of pelvic x-rays acquired are failures. Eventually being able to detect these "fail" cases reliably should aid in machine learning workflows by excluding malpositioned cases and may even aid Radiology workflow by providing real-time quality control by flagging cases that may need to be repeated.

Regarding the fracture classification portion of the study, to be comparable to previous studies, one model was trained only on fractured proximal femoral cases vs. normal cases. Our AUC of 0.95 was slightly lower compared to previous studies with AUC's of 0.97–0.99 using bounding boxes[21,23,24], but showed that models can perform well without manually isolating the proximal femur. Furthermore, our femoral fracture test set (1274 non-fractured and 620 fractured cases) was randomly selected from the total dataset with a significantly higher number of cases compared to a previous study with only 100 test cases[20]. In addition, unlike most of the other hip fracture machine learning research, we evaluated the demographics of our population. The only study that reported demographics was done by Cheng et al.[20], but there was a large demographic difference in the fractured cases (mean 45 years old and 68% males) versus the non-fractures cases (mean age 72 and 42% males) with $p<0.001$. We did not have a statistical difference in age between the non-fractured and fracture groups. Although the sex distribution was statistically significant, there was only a 2% difference between the two groups (40% vs. 38% with $p=0.041$). This analysis shows that our non-fractured and fractured groups were relatively similar, contrary to the other study.

Unlike any of the previous studies, we also evaluated pelvic and acetabular fractures in addition to hip fractures using deep learning. Detecting other types of fractures slightly lowered the AUC for proximal femoral fractures, from 0.95 to 0.93 with the "separate" model. Grouping the proximal femoral fractures together with acetabular fractures further lowered the AUC of fracture classification to 0.88, suggesting that machine learning models perform best at detecting proximal femoral fractures. Grouping the anterior and posterior pelvic ring fractures increased the AUC from as low as 0.70 to 0.81, which was understandable as the number of model classes decreases from 6 to 5. In nearly all data mixtures excluding the "separate" dataset, the complex class showed the lowest AUC between 0.61 to 0.77, which may be due to the model predicting the class based on the predominant fracture type rather than on a holistic view. In the future, a multi-label rather than a multi-class approach to tag the images to eliminate the complex class may improve performance. Clinically, our study implies that machine learning models may have utility as a general fracture detector for pelvic x-rays rather than a narrow hip fracture detector.

One point of emphasis from this study is the importance of image labeling. Current pelvic ring fracture types are largely divided into anterior-posterior compression, lateral compression, and vertical shear injuries[33–35], which directed our split of fractures into anterior and posterior pelvic categories and grouping them together as a pelvic ring class in the "pelvis consolidated" model. We also consciously grouped fractures in unconventional ways. For example, with the "femoral/acetabular consolidated" category, we understand that proximal femoral and acetabular fractures are very different in terms of biomechanics, but they are spatially in a similar location, and depending on the clinical need, may be to provide a fracture classification model should be enough. We also showed that initially training models on granular labels and outputting them as binary classes performed as well as models initially trained on binary data, indicating simply labeling images as normal vs abnormal is not worthwhile. Finally, when each of the trained models was evaluated using a test dataset only containing proximal femoral fractures and normal cases, the AUC value differences were mostly not statistically significant (Table 3), showing that increasing model

capabilities to detect other types of fractures on pelvic x-rays does not notably decrease proximal femoral fracture detection capabilities.

To visualize the model attention, a gradient-weighted class activation mapping (Grad-CAM) technique was applied[32]. This technique emphasizes the important feature maps of a neural network layer for a given class. As can be seen from Figure 6, for each of the different fracture classes, the heatmap focuses on the fracture regions of interest/locations . To further refine the heatmap, a guided-Grad-CAM[32] technique is utilized to apply a high resolution visualization in addition to the class-discriminative heatmap. For each of the classes in Figure 6, the guided-Grad-CAM map shows a higher resolution localization to the bones, rather than simply to a region. Visualization methods help investigators get an idea of where the model is focusing and may increase the believability of the outputs.

Limitations of this study include broad position and hardware categories; in the future, we expect to tease out the pelvic and hip position and hardware to be more granular. For fracture detection, we did not employ any bounding boxes for the fracture site, which may improve the model performance for fracture detection. To accommodate the DenseNet architecture, we resized all images to 224 × 224 pixels; however, subtle fractures often only occupy a very small space in the image, and downsizing the images likely led to distorting these areas and lowering the classification performance. Image downsizing is performed due to computational limits, but evolving hardware and new architectures allowing a larger image input size may increase the detection of subtle fractures in the future. In addition, we only employed a single reader as the gold-standard, rather than a group of readers. Furthermore, we did not perform any hyperparameter tuning using a validation dataset; the test dataset was used to obtain a one-shot output metric to demonstrate the proof of the concept. Therefore, it is conceivable that with training iterations using a validation dataset, improved model performance could have been achieved. Finally, this was a single institution study, and we hope to eventually expand to a multi-institutional study to increase our dataset to create more robust and effective models.

Conclusion

In conclusion, we were able to successfully create deep learning models that can accurately categorize pelvic imaging position and hardware presence, which may aid in future machine learning projects or radiology workflow implementation. Regarding fracture detection, we created a model that can classify pelvic and acetabular fractures with very little performance loss for proximal femoral fracture detection compared to prior research. Though the AUC for the pelvic and acetabular fractures is not as high as the proximal femoral fracture, we hope that with continued research, we can get the values closer to proximal femoral values.

Acknowledgement:

This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

Funding:

The project described was supported by the National Institutes of Health through Grant Number UL1 TR001857.

Abbreviations:

DICOM	Digital Imaging and Communications in Medicine
ROC	receiver operator characteristic
AUC	area under the curve
CT	computed tomography
MRI	magnetic resonance imaging
GPU	Graphical Processing Unit
Grad-CAM	gradient-weighted class activation mapping
CI	confidence interval
PPV	positive predictive value
NPV	negative predictive value

References

[1]. Kanis JA, Odén A, McCloskey EV, Johansson H, Wahl DA, Cooper C, IOF Working Group on Epidemiology and Quality of Life, A systematic review of hip fracture incidence and probability of fracture worldwide, Osteoporos. Int 23 (2012) 2239–2256. [PubMed: 22419370]

[2]. Dhanwal DK, Dennison EM, Harvey NC, Cooper C, Epidemiology of hip fracture: Worldwide geographic variation, Indian J. Orthop 45 (2011) 15–22. [PubMed: 21221218]

[3]. Berry SD, Miller RR, Falls: Epidemiology, pathophysiology, and relationship to fracture, Current Osteoporosis Reports. 6 (2008) 149–154. 10.1007/s11914-008-0026-4. [PubMed: 19032925]

[4]. Orsini LS, Rousculp MD, Long SR, Wang S, Health care utilization and expenditures in the United States: a study of osteoporosis-related fractures, Osteoporos. Int 16 (2005) 359–371. [PubMed: 15340799]

[5]. Blume SW, Curtis JR, Medical costs of osteoporosis in the elderly Medicare population, Osteoporos. Int 22 (2011) 1835–1844. [PubMed: 21165602]

[6]. Rommens PM, Hessmann MH, Staged reconstruction of pelvic ring disruption: differences in morbidity, mortality, radiologic results, and functional outcomes between B1, B2/B3, and C-type lesions, J. Orthop. Trauma 16 (2002) 92–98. [PubMed: 11818803]

[7]. Rommens PM, Pelvic ring injuries: a challenge for the trauma surgeon, Acta Chir. Belg 96 (1996) 78–84. [PubMed: 8686407]

[8]. Laird A, Keating JF, Acetabular fractures: a 16-year prospective epidemiological study, J. Bone Joint Surg. Br 87 (2005) 969–973. [PubMed: 15972913]

[9]. Dakin GJ, Eberhardt AW, Alonso JE, Stannard JP, Mann KA, Acetabular fracture patterns: associations with motor vehicle crash information, J. Trauma 47 (1999) 1063–1071. [PubMed: 10608534]

[10]. Hakkarinen DK, Banh KV, Hendey GW, Magnetic resonance imaging identifies occult hip fractures missed by 64-slice computed tomography, J. Emerg. Med 43 (2012) 303–307. [PubMed: 22459594]

[11]. Chellam WB, Missed subtle fractures on the trauma-meeting digital projector, Injury. 47 (2016) 674–676. [PubMed: 26653270]

[12]. Rehman H, Clement RGE, Perks F, White TO, Imaging of occult hip fractures: CT or MRI?, Injury. 47 (2016) 1297–1301. 10.1016/j.injury.2016.02.020. [PubMed: 26993257]

[13]. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT, Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs, Radiology: Artificial Intelligence. 1 (2019) e180001.

[14]. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A, Binomial Classification of Pediatric Elbow Fractures Using a Deep Learning Multiview Approach Emulating Radiologist Decision Making, Radiology: Artificial Intelligence. 1 (2019) e180015.

[15]. Kitamura G, Chung CY, Moore BE 2nd, Ankle Fracture Detection Utilizing a Convolutional Neural Network Ensemble Implemented with a Small Sample, De Novo Training, and Multiview Incorporation, J. Digit. Imaging 32 (2019) 672–677. [PubMed: 31001713]

[16]. Kim JR, Shim WH, Yoon HM, Hong SH, Lee JS, Cho YA, Kim S, Computerized Bone Age Estimation Using Deep Learning Based Program: Evaluation of the Accuracy and Efficiency, AJR Am. J. Roentgenol (2017) 1–7.

[17]. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, Choy G, Do S, Fully Automated Deep Learning System for Bone Age Assessment, J. Digit. Imaging 30 (2017) 427–441. [PubMed: 28275919]

[18]. Xue Y, Zhang R, Deng Y, Chen K, Jiang T, A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis, PLoS One. 12 (2017) e0178992. [PubMed: 28575070]

[19]. Li Y, Huang Z, Dong X, Liang W, Xue H, Zhang L, Zhang Y, Deng Z, Forensic age estimation for pelvic X-ray images using deep learning, Eur. Radiol 29 (2019) 2322–2329. [PubMed: 30402703]

[20]. Cheng C-T, Ho T-Y, Lee T-Y, Chang C-C, Chou C-C, Chen C-C, Chung I-F, Liao C-H, Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs, Eur. Radiol 29 (2019) 5469–5477. [PubMed: 30937588]

[21]. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ, Detecting hip fractures with radiologist-level performance using deep neural networks, arXiv [cs.CV] (2017). http://arxiv.org/abs/1711.06504.

[22]. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, McConnell MV, Percha B, Snyder TM, Dudley JT, Deep learning predicts hip fracture using confounding patient and healthcare variables, NPJ Digit Med. 2 (2019) 31. [PubMed: 31304378]

[23]. 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. 48 (2019) 239–244. [PubMed: 29955910]

[24]. Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, Zaid M, McGill KC, Patel R, Sohn JH, Wright A, Darger BF, Padrez KA, Ozhinsky E, Majumdar S, Pedoia V, Automatic Hip Fracture Identification and Functional Subclassification with Deep Learning, arXiv [q-bio.QM] (2019). http://arxiv.org/abs/1909.06326.

[25]. Huang G, Liu Z, van der Maaten L, Weinberger KQ, Densely Connected Convolutional Networks, arXiv [cs.CV] (2016). http://arxiv.org/abs/1608.06993.

[26]. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv [cs.DC] (2016). http://arxiv.org/abs/1603.04467.

[27]. Jones E, Oliphant T, Peterson P - URL http://scipy.org, 2011, SciPy: Open source scientific tools for Python, 2009, (2011).

[28]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res 12 (2011) 2825–2830.

[29]. Mason D - Medical Physics, 2011, SU- E- T- 33: Pydicom: An Open Source DICOM Library, Wiley Online Library. (2011). http://onlinelibrary.wiley.com/doi/10.1118/1.3611983/full.

[30]. Hanley JA, McNeil BJ, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology. 148 (1983) 839–843. [PubMed: 6878708]

[31]. Kosinski AS, A weighted generalized score statistic for comparison of predictive values of diagnostic tests, Stat. Med 32 (2013) 964–977. [PubMed: 22912343]

[32]. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, International Journal of Computer Vision. (2019). 10.1007/s11263-019-01228-7.

[33]. Alton TB, Gee AO, Classifications in Brief: Young and Burgess Classification of Pelvic Ring Injuries, Clinical Orthopaedics and Related Research®. 472 (2014) 2338–2342. 10.1007/s11999-014-3693-8. [PubMed: 24867452]

[34]. Tile M, Helfet DL, Kellam JF, Fractures of the Pelvis and Acetabulum: Principles and Methods of Management, Thieme, 2015.

[35]. Young JW, Burgess AR, Brumback RJ, Poka A, Pelvic fractures: value of plain radiography in early assessment and management, Radiology. 160 (1986) 445–451. [PubMed: 3726125]

Highlights

- Deep learning is effective at classifying pelvic radiograph positions.
- Deep learning can accurately detect hardware on pelvic radiographs.
- Deep learning models can detect hip, pelvic and acetabular fractures.

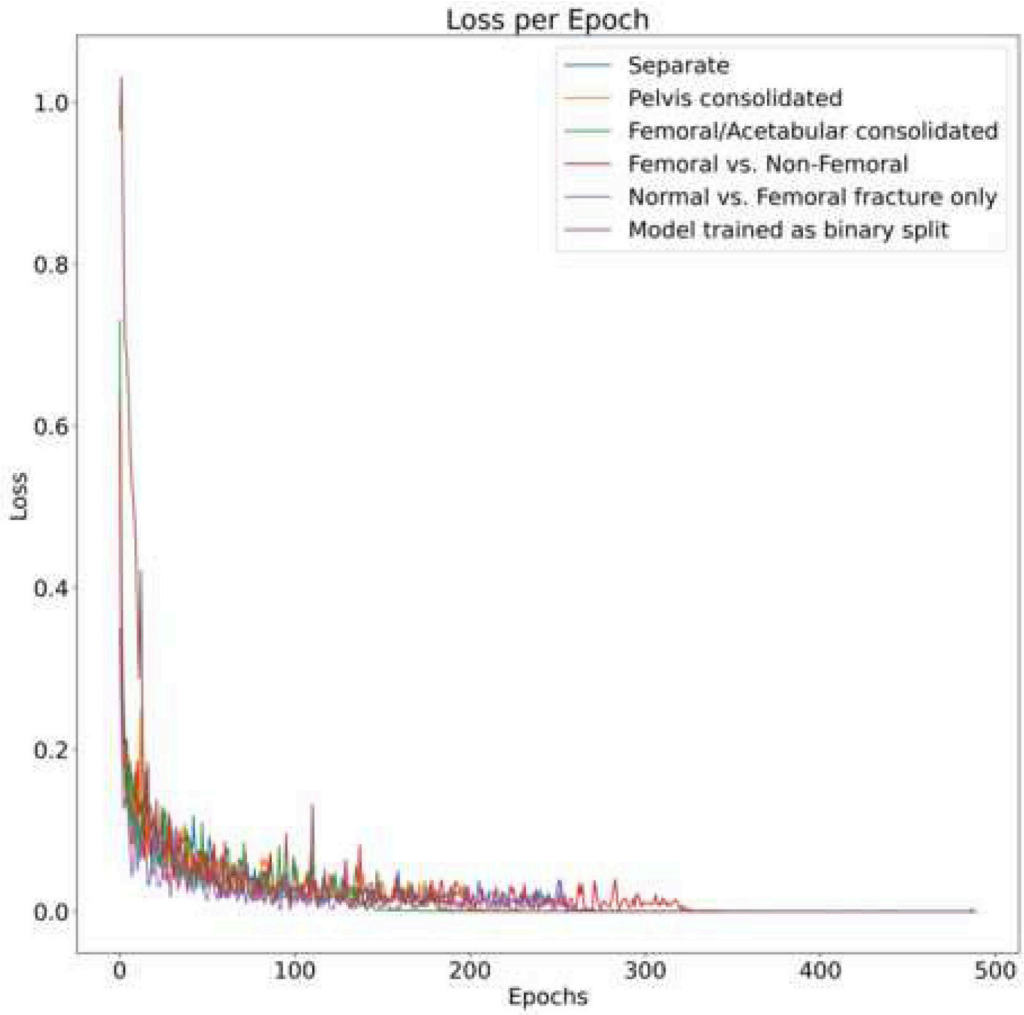

Fig.1.

Losses of each model plotted against the epochs. With all models, training convergence is monitored as loss plateauing. By epoch 400, the loss of all models has stop decreasing and the model is considered converged.

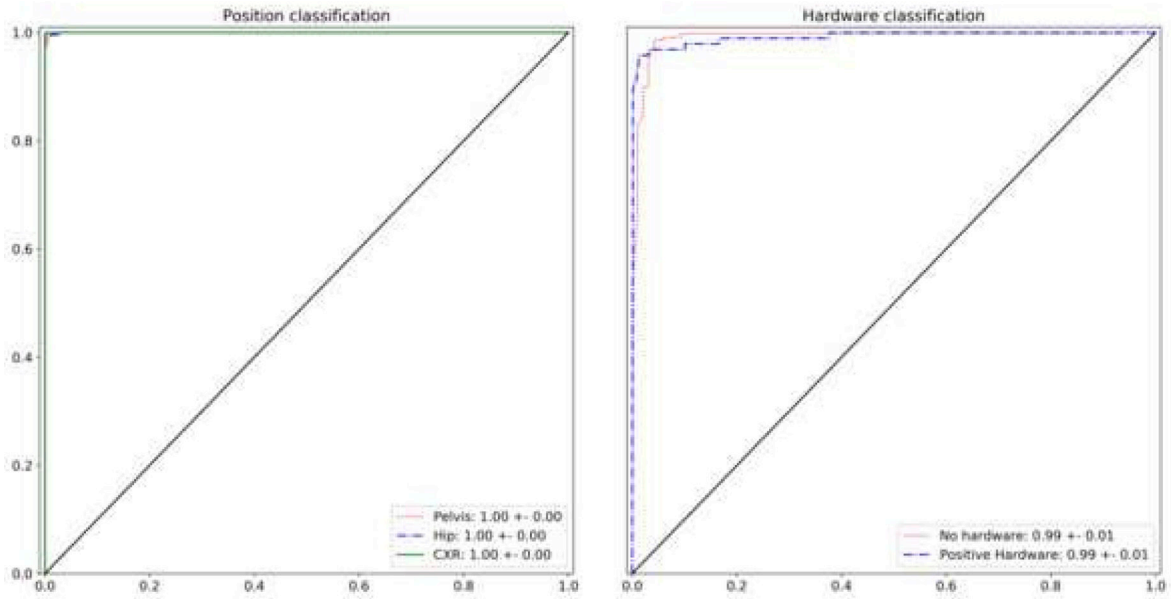

Fig.2.

Receiver Operator Curve (ROC) curve for pelvic image position and hardware presence classification using the test dataset. The area under the curve (AUC) value for the pelvis, hip, and chest x-ray classification was nearly perfect at 1.0. We did not plot the “fail” class ROC since it only contained 6 cases and the curve would have been misleading. AUC values for hardware presence and absence were also very high at 0.99. The 95% confidence interval is noted.

Fig.3.

Sample images and their probability values for the positions or hardware using the test dataset. The output label can be compared with the actual image to confirm concordance.

Fig.4.

Images of the “fail” position cases from the test dataset. These are suboptimally positioned images that neither fit the pelvis or hip position. Only one of the images is accurately classified as “fail” based on the probability output.

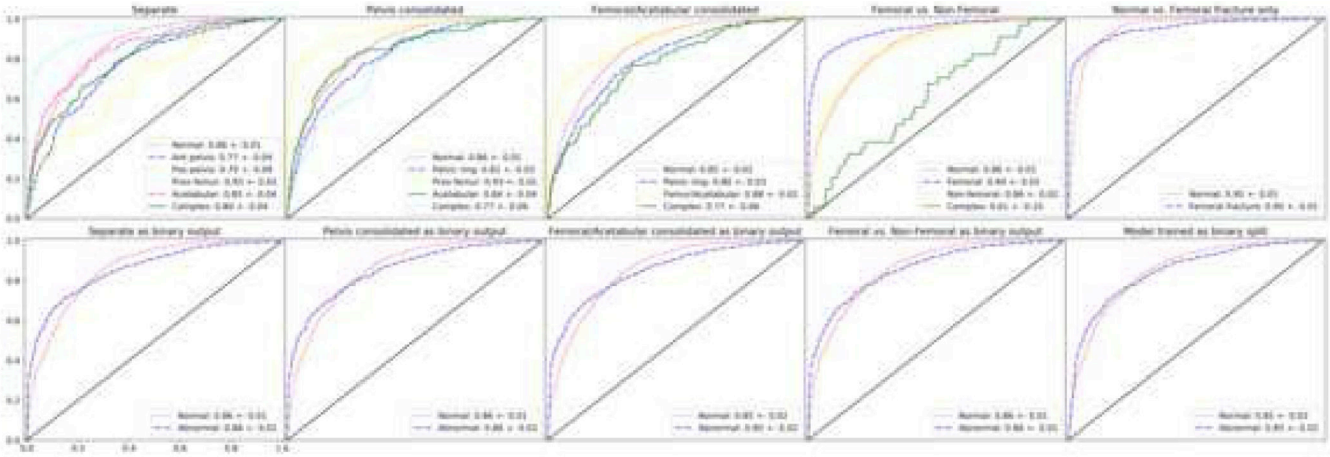

Fig.5.

Receiver Operator Curve (ROC) curve for pelvic fracture detection and classification using the test dataset. The area under the curve (AUC) value for the fracture classes are noted in the bottom right corner for each graph. For the “Separate”, “Pelvis consolidated”, “Femoral/Acetabular consolidated”, and “Femoral vs. Non-Femoral” models, the classes were evaluated separately on the top row and as a binary split of normal vs. abnormal on the corresponding bottom row. The far upper right graph shows the ROC curve for the model created using only proximal femoral fractures vs. non-fractured cases. Finally, the far bottom right plot is the ROC curve when a model was trained with an initial binary split. The 95% confidence interval is noted.

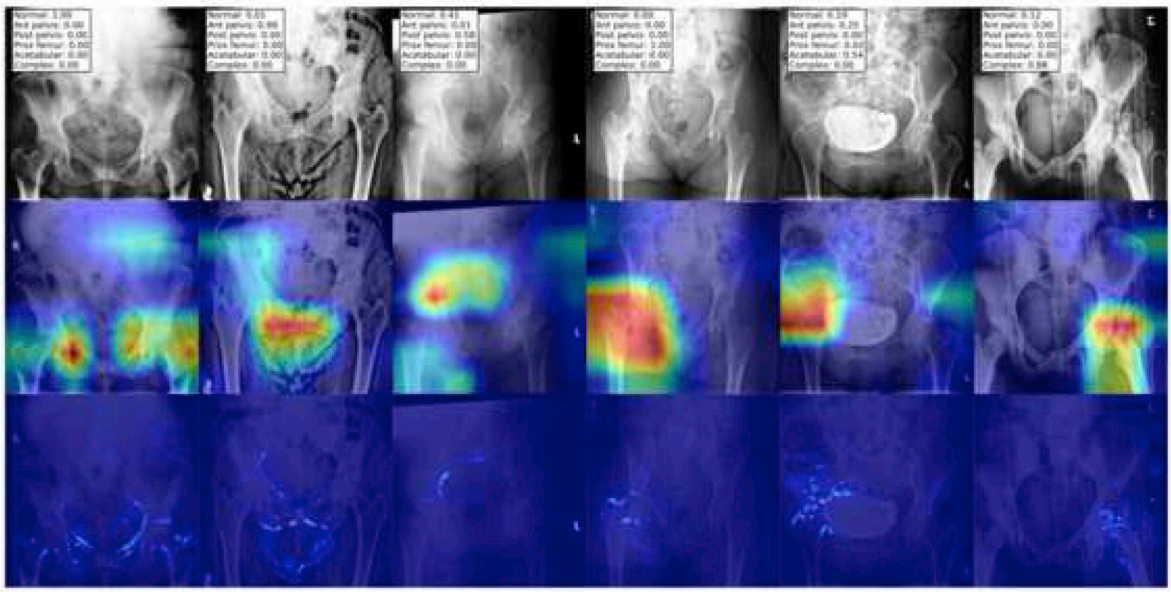

Fig.6.

Model visualization of fractures. A representative image for each of the “separate” fracture types are shown with the highest probabilities denoting the appropriate fracture types in the top row. The second row shows the heatmap of the fracture class using the Gradient-weighted class activation mapping (Grad-CAM) technique. The third row shows the guided-Grad-CAM technique, illustrating a high resolution heatmap focusing on the part of the image contributing most to the fracture classification.

Table 1:

The demographic for the fracture detection dataset.

	Non-fractured	Fractured	p value
Total number	3428	3909	
Mean age	66.9	67.8	0.062
% Males	40.4%	38.0%	0.041

Table 2:

Area under the curve values with 95% confidence intervals for the various trained fracture detection models evaluating the different fracture categories.

Fracture category	Model					
	Separate	**Pelvis consolidated**	**Femoral / Acetabular consolidated**	**Femoral vs. Non-Femoral**	**Normal vs. Femoral fracture only**	**Binary split**
Normal	0.86 ± 0.01	0.86 ± 0.01	0.85 ± 0.02	0.86 ± 0.01	0.95 ± 0.01	n/a
Ant. pelvis	0.77 ± 0.04	n/a	n/a	n/a	n/a	n/a
Pos. pelvis	0.70 ± 0.09	n/a	n/a	n/a	n/a	n/a
Pelvic ring	n/a	0.81 ± 0.03	0.80 ± 0.03	n/a	n/a	n/a
Prox. femur	0.93 ± 0.01	0.93 ± 0.01	n/a	0.94 ± 0.01	0.95 ± 0.01	n/a
Acetabular	0.85 ± 0.04	0.84 ± 0.04	n/a	n/a	n/a	n/a
Femur/ Acetabular	n/a	n/a	0.88 ± 0.02	n/a	n/a	n/a
Non-femoral	n/a	n/a	n/a	0.86 ± 0.02	n/a	n/a
Complex	0.80 ± 0.04	0.77 ± 0.06	0.77 ± 0.06	0.61 ± 0.10	n/a	n/a
Binary output: Normal	0.86 ± 0.01	0.86 ± 0.01	0.85 ± 0.02	0.86 ± 0.01	n/a	0.85 ± 0.02
Binary output: Abnormal	0.86 ± 0.01	0.86 ± 0.02	0.85 ± 0.02	0.86 ± 0.01	n/a	0.85 ± 0.02

Table 3:

Output metrics with 95% confidence intervals for the fully trained models when evaluated using a test dataset only containing proximal femoral fractures and normal cases.

Model	Output Metrics and p values									
	Sensitivity	**p value**	**Specificity**	**p value**	**PPV**	**p value**	**NPV**	**p value**	**AUC**	**p value**
Normal vs. Femoral fracture only^	0.86±0.02*	N/A	0.90±0.01	N/A	0.81±0.02	N/A	0.93±0.01*	N/A	0.95±0.01*	N/A
Separate	0.79±0.02	<0.001	0.94±0.01*	<0.001	0.87±0.2*	<0.001	0.90±0.01	<0.001	0.94±0.01	0.057
Pelvis consolidated	0.84±0.02	0.018	0.90±0.01	0.929	0.80±0.02	0.58	0.92±0.01	0.013	0.94±0.01	0.076
Femoral / Acetabular consolidated	0.77±0.02	<0.001	0.94±0.01*	<0.001	0.85±0.02	0.007	0.89±0.01	<0.001	0.92±0.02	<0.001
Femoral vs. Non-Femoral	0.83±0.02	0.004	0.92±0.01	0.017	0.84±0.02	0.042	0.92±0.01	0.008	0.94±0.01	0.192

*highest output metric values

^p-values are compared against the model trained only on normal vs. femoral fracture cases.