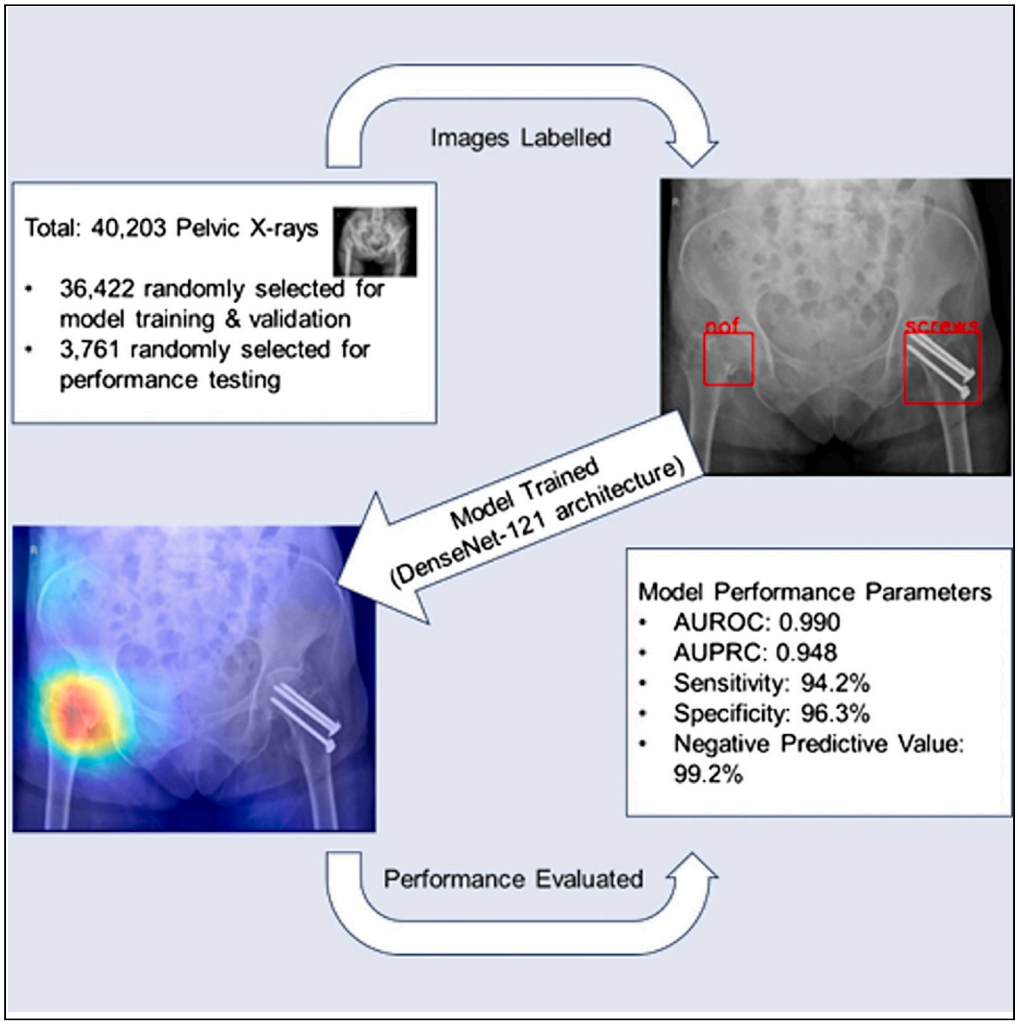
iScience

CellPress
OPEN ACCESS

Article

Application of a deep learning algorithm in the detection of hip fractures



Yan Gao, Nicholas Yock Teck Soh, Nan Liu, ..., Narayan Venkataraman, Siang Hiong Goh, Yet Yen Yan

yan.yet.yen@singhealth.com.sg

Highlights

Deep learning model developed with >40k images predicts hip fractures on pelvic X-rays

All x-rays included regardless of technical quality, other pathologies or implants

Differs from previous work which tended to have selective criteria for image inclusion

Model achieved high sensitivity (94.2%) and specificity (96.3%)

Gao et al., iScience 26, 107350
August 18, 2023
https://doi.org/10.1016/j.isci.2023.107350

iScience

CellPress
OPEN ACCESS

Article

Application of a deep learning algorithm in the detection of hip fractures

Yan Gao,[1] Nicholas Yock Teck Soh,[2] Nan Liu,[3] Gilbert Lim,[3] Daniel Ting,[4] Lionel Tim-Ee Cheng,[5,6] Kang Min Wong,[2,6] Charlene Liew,[2,6] Hong Choon Oh,[1] Jin Rong Tan,[5] Narayan Venkataraman,[7] Siang Hiong Goh,[8] and Yet Yen Yan[2,6,9,*]

SUMMARY

This paper describes the development of a deep learning model for prediction of hip fractures on pelvic radiographs (X-rays). Developed using over 40,000 pelvic radiographs from a single institution, the model demonstrated high sensitivity and specificity when applied to a test set of emergency department radiographs. This study approximates the real-world application of a deep learning fracture detection model by including radiographs with sub-optimal image quality, other non-hip fractures, and metallic implants, which were excluded from prior published work. The study also explores the effect of ethnicity on model performance, as well as the accuracy of visualization algorithm for fracture localization.

INTRODUCTION

Hip fractures are a major public health problem, with global incidence increasing due to population aging and estimated to reach 6.3 million by 2050.[1] These fractures commonly occur in the elderly, with falls from standing height the most frequent mechanism of injury.[2,3] Mortality associated with hip fractures has remained relatively high and largely unchanged in the past decade, with overall one-year mortality in North America reported at approximately 27%.[4,5] Of those who survive the initial hospitalization after a hip fracture, a large proportion suffer from permanent disability, reduced independence, and social isolation.[6,7] Up to 40% of patients are unable to walk independently one year after a hip fracture, and up to 20% will be permanently institutionalized.[8,9]

Hip fractures can be divided into intracapsular (neck of femur) and extracapsular (trochanteric and subtrochanteric) types.[10] The primary modality of diagnosis of these fractures remains conventional radiography, which is relatively low-cost and readily available at emergency departments and urgent care facilities.[11] The frontal pelvic radiograph (PXR) is the most frequently performed projection and allows for evaluation of the bony pelvis as well as both proximal femora.[12] Computed tomography, magnetic resonance imaging, and nuclear scintigraphy are options for further evaluation of suspected occult fractures.[11] When interpreted by radiologists, sensitivity of PXR for hip fracture is high, having been reported as between 90 and 98%, with 1.6–4% patients having occult fractures subsequently diagnosed on other modalities.[11,13] However, many medical facilities do not have round-the-clock radiology staff coverage, potentially contributing to delays in image interpretation and diagnosis.[14] Delayed diagnosis of hip fractures and resultant prolonged time to hospital admission and corrective surgery have been demonstrated to increase patient mortality and morbidity.[15–17]

There is potential for computer aided diagnosis (CAD) to fill such gaps in radiology expertise and availability. Computed radiography and picture archiving and communication systems are commonplace in radiology facilities today and are frequently integrated with radiology information systems and hospital electronic medical records.[18] Beyond enabling the trends of remote work and teleradiology over the past decades, they offer potential for implementation of deep learning in CAD.[19,20] Deep convolutional neural networks (DCNNs) are a class of deep learning algorithms which can accurately classify images and perform object recognition.[21] Rapid advancements in the field have been driven by availability of large image sets for training and increasing computational power.[21,22] In recent years, applications of DCNNs in CAD have been explored as a tool to augment physicians and improve patient care. These implementations have covered a wide range of image modalities and pathological conditions, from optical coherence tomography in age-related macular degeneration to chest radiographs in detection of pneumonia.[23,24] These

[1]Health Services Research, Changi General Hospital, Singapore Health Services (SingHealth), Singapore, Singapore

[2]Department of Diagnostic Radiology, Changi General Hospital, Singapore Health Services (SingHealth), Singapore, Singapore

[3]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore

[4]Singapore Health Services (SingHealth), Duke-NUS Medical School, Singapore, Singapore

[5]Department of Diagnostic Radiology, Singapore General Hospital, Singapore Health Services (SingHealth), Singapore, Singapore

[6]Radiological Sciences ACP, Duke-NUS Medical School, Singapore, Singapore

[7]Department of Medical Informatics, Changi General Hospital, Singapore Health Services (SingHealth), Singapore, Singapore

[8]Department of Emergency Medicine, Changi General Hospital, Singapore Health Services (SingHealth), Singapore, Singapore

[9]Lead contact

*Correspondence: yan.yet.yen@singhealth.com.sg

https://doi.org/10.1016/j.isci.2023.107350

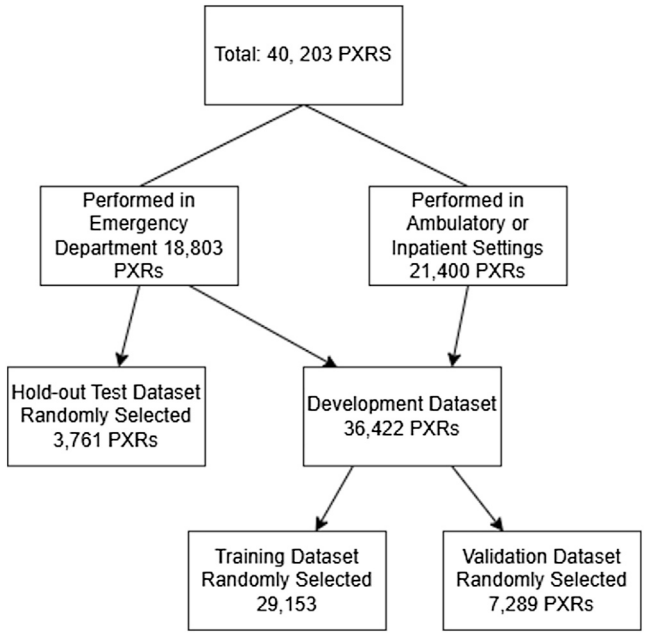


Figure 1. Allocation of PXRs into training, validation and test datasets

studies have shown promising results, with sensitivity and specificity of some narrow CAD solutions equaling or exceeding human diagnostic radiologists.[25–27]

Recent studies have explored the use of DCNN in CAD of hip fractures. A major hurdle in developing highly accurate CAD solutions for medical image interpretation remains the availability of large volume medical image sets with high-quality annotations. Most previous studies have either relied on relatively small numbers of images, simple methods of annotation (e.g., point based annotation), and automated solutions for annotation, or utilized weak labels (e.g., presence or absence of pathology without specifying location).[28–35] Several prior studies with large training sets have a wide exclusion criterion and exclude radiographs with implants, other non-hip fractures, poor positioning or suboptimal image quality, potentially introducing selection bias.[28,29,36] A few studies have also excluded specific subtypes of hip fractures (e.g., trochanteric, sub-trochanteric), limiting the real-world utility of their models.[37,38] The absence of sub-group analysis between ethnicities and limited provision of patient characteristics in these studies also may restrict the applicability of reported results to a real-world population.

The aim of this study is to develop and examine the performance of a DCNN solution, constructed on DenseNet-121 architecture with pre-trained ImageNet weights, for CAD of hip fractures on PXRs, utilizing a large image set of over 40,000 images and trained on TensorFlow using image-level labels and an image classifier approach.[39–41] All PXRs will be included in this study regardless of perceived image quality, presence of other non-hip fractures or metallic implants. In addition, the performance of Grad-CAM used as a visual adjunct to highlight regions of interest within the images and examine validity of the algorithm will be evaluated.

RESULTS

Cohort characteristics

Of the 36,422 PXRs used for training (29,153 or 80%) and validation (7,289 or 20%) of the model, 2,672 (7.3%) were positive for hip fracture Figure 1. In contrast, of the 3,761 PXRs in the hold-out test dataset, 463 were positive for hip fracture (12.3%). Orthopedic implants in either proximal femur or the bony pelvis was present in a larger proportion of training and validation set PXRs (34.3%) as opposed to training set PXRs (10.3%). Both of these variations are related to differences in case-mix between the emergency department versus ambulatory and inpatient settings.

Table 1. Comparison of characteristics between PXRs grouped by dataset cohorts

	Datasets	
Characteristics	Training and Validation	Hold-out Test
Total number	36,442	3,761
Gender; n (%)		
Male	14,532 (40)	1,734 (46)
Female	20,450 (56)	2,024 (54)
Not classified	1,460 (4)	3 (0)
Ethnicity; n (%)		
Chinese	23,398 (64)	2,397 (64)
Malay	5,881 (16)	674 (18)
Indian	2,447 (7)	292 (8)
Others	3,252 (9)	395 (11)
Not classified	1,464 (4)	3 (0)
Age in years; mean (std dev)	68.0 (22.3)	65.4 (26.7)
Performed in; n (%)		
Emergency Department	15,042 (41)	3,761 (100)
Ambulatory or Inpatient	21,400 (59)	0 (0)
Hip Fracture; n (%)	2,672 (7.3)	463 (12.3)
Neck of Femur	1,371 (3.8)	225 (6.0)
Trochanteric	1,239 (3.4)	213 (5.7)
Subtrochanteric	391 (1.1)	81 (2.2)
Atypical fracture	16 (0.0)	2 (0.1)
Orthopedic implant; n (%)	12,489 (34.3%)	389 (10.3%)

Demographics between the cohorts of PXRs in the training and validation sets compared to the hold-out test set are as shown in Table 1, with differences in gender, ethnicity, age, hip fracture prevalence and presence of orthopedic implants related to the differences of these cohort characteristics in ambulatory, inpatient and emergency department settings.

Model performance

Performance of the hip fracture detection model was evaluated with a hold-out test dataset of 3,761 PXRs of which 463 (12.3%) were positive for hip fracture. Our hip fracture detection network achieved AUROC of 0.990 (95% confidence interval [CI]:0.986, 0.993) and AUPRC of 0.948 (95% CI: 0.926, 0.965). Using Youden's index, the operating point is determined to be 0.127.[43] At this threshold, the model predicted 27 false negatives and 121 false positives (Table 2). The model detected 7 of 7 undisplaced fractures (sensitivity 100%) and 429 of 456 displaced fractures (sensitivity 94.1%).

Subgroup analysis of model performance was performed to examine performance parameters in the presence of orthopedic implants within the PXR (Table 3). Within the hold-out test set, there were 389 PXRs with at least one implant, of which 32 (8.2%) were positive for hip fracture. Of the other 3,372 PXRs without implants, 431 (12.8%) were positive for hip fracture.

Subgroup analysis of model performance was also performed to examine performance parameters between the various ethnicities (Table 4). The hold-out test set included PXRs of 2,397 Chinese, 674 Malays, 292 Indians, and 395 others, of which 354 (14.7%), 56 (8.3%), 22 (7.5%) and 30 (7.6%) were positive for hip fracture.

Localization of predicted fractures with Grad-CAM

Grad-CAM heatmaps were generated for all 557 hip fractures predicted by our model. Of the 436 true positive hip fractures accurately predicted by our model, there were 101 (23.2%) instances where the model identified an incorrect activation site. Fused heatmaps for the other 335 (76.8%) predicted fractures

Table 2. Performance of hip fracture detection model on hold-out test dataset

	Model Predictions for Hip Fracture					
Ground truth	Predicted Absent			Predicted Present		
Hip Fracture Absent	3177 (True Negative)			121 (False Positive)		
Hip Fracture Present	27 (False Negative)			436 (True Positive)		
Performance Parameters						
AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV
0.990 (95% CI: 0.986, 0.993)	0.948 (95% CI:0.926, 0.965)	96.1%	94.2%	96.3%	78.3%	99.2%

demonstrated activation area correctly located at the actual hip fracture, as assessed by two consultant musculoskeletal radiologists. Interestingly, it was observed that Grad-CAM heatmaps frequently outlined the lateral femoral neck even in the presence of unequivocal cortical disruption and fracture line.

Figures 2 and 3 demonstrate examples of Grad-CAM heatmaps in true positive predicted fractures. Figure 4 demonstrates an example of Grad-CAM heatmaps in true positive predicted fractures delineating the lateral cortical outline of the proximal femur.

DISCUSSION

Misinterpretation of PXRs can contribute to missed diagnosis and delay surgical repair, and hip fractures represent an ideal target for DCNN solutions for CAD.[44] This study utilized a DCNN and relied on a combination of independent board certified musculoskeletal radiologist reads, index radiology reports and subsequent advanced imaging reports (where available) to determine ground-truth labeling, similar to previous literature.[29]

This study demonstrates the ability for a DCNN solution to identify hip fractures on PXRs with extremely high accuracy. The study included 40,203 PXRs and the model achieved AUROC of 0.990 (95% CI: 0.986, 0.993) and AUPRC of 0.948 (95% CI: 0.926, 0.965) when applied to the hold-out test set comprising emergency department radiographs. These results are comparable with previous large volume studies by Gale et al.,[29] Oakden-Rayner et al.,[27] Kitamura et al.,[36] and Cheng et al.,[28] who reported AUROC values between 0.98 and 0.99, and is markedly improved compared to earlier smaller studies which reported relatively low sensitivity and specificity below that of human radiologists.[31–33] The model also achieved high sensitivity and specificity of 94.2% and 96.3% respectively, and extremely high negative predictive value (NPV) of 99.2%. These parameters compare favorably to the mean sensitivity and specificity of 89.3% and 87.5% reported by a recent meta-analysis of 18 studies with a total of 39,598 radiographs.[34] The high sensitivity and negative predictive value of our model underscores the potential for DCNN CAD solutions like ours to be particularly useful in urgent or emergency care settings, where emphasis is on avoiding missed diagnoses.

A major strength of this study was its inclusion of all PXRs performed within the recruitment period, regardless of perceived technical and diagnostic difficulty, existence of metallic implants or presence of other radiographically identified pathologies, for example pelvic fractures or bone tumors. This contrasts with most previous studies which excluded certain subsets of PXRs or pathologies. For example, Cheng et al. and Kitamura et al. excluded PXRs deemed to have poor image contrast, positioning errors, foreign body interference and those with other fractures.[28,36] Gale et al. and subsequently Oakden-Rayner et al. excluded PXRs with metallic implants.[27,29] Murphy et al. and Bae et al. excluded certain types of hip fractures (trochanteric and/or subtrochanteric).[37,38] Although the effects of these exclusions are difficult to confidently predict, they potentially introduce selection

Table 3. Subgroup analysis of model performance in presence of orthopedic implants

	Performance Parameters (Orthopedic Implants Present/Absent)						
	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV
Implant present	0.969 (95% CI: 0.926, 0.998)	0.914 (95% CI: 0.822, 0.983)	0.954	0.875	0.961	0.667	0.988
Implant absent	0.991 (95% CI: 0.988, 0.994)	0.950 (95% CI: 0.927, 0.968)	0.959	0.947	0.961	0.779	0.992

Table 4. Subgroup analysis of model performance in different ethnic groups

	Performance Parameters (Ethnicity Subgroups)						
	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV
Chinese	0.987 (95% CI: 0.981, 0.991)	0.952 (95% CI: 0.937, 0.967)	0.953	0.932	0.956	0.788	0.988
Malay	0.997 (95% CI: 0.993, 1.000)	0.975 (95% CI: 0.946, 0.995)	0.985	0.964	0.987	0.871	0.997
Indian	0.993 (95% CI: 0.983, 1.000)	0.851 (95% CI: 0.659, 1.000)	0.986	0.909	0.993	0.909	0.993
Others	0.997 (95% CI: 0.991, 1.000)	0.972 (95% CI: 0.930, 1.000)	0.975	0.933	0.978	0.778	0.994

bias and possibly undermine the robustness of developed algorithms. Hence, results reported in such studies may be overly optimistic or not generalizable to excluded patient populations, limiting the algorithms' potential real-world use.[45] This study's broad inclusion of all PXRs while still achieving comparably high performance lends credibility to this study's results, their generalizability and clinical applicability of the developed model, although the clinical utility of this algorithm requires external validation.

Another strength of this study was its utilization of a hold-out test set comprising exclusively of emergency department PXRs, as opposed to radiographs performed in other ambulatory or inpatient settings. This was a deliberate decision to better approximate the potential real-world use scenario of such a model, where in routine clinical practice, the majority of patient falls and injuries occur outside of the hospital, and patients with suspected hip fractures predominantly present to emergency departments or urgent care facilities.[46] Comparatively, a sizable proportion of PXRs performed in the inpatient and ambulatory settings may be for follow-up of prior injuries or post-surgical implants. This is supported by the study's findings, with PXRs performed in the emergency department having a higher prevalence of hip fractures and lower prevalence of surgical implants.

Subgroup analysis of PXRs with orthopedic implants found that model performance was slightly lower when an implant was present in the radiograph, but still relatively comparable with AUROC of 0.969 (versus 0.991 when no implant was present) and AUPRC of 0.914 (versus 0.950 when no implant was present). Nonetheless, NPV remained extremely high at 0.988 (versus 0.992 when no implant was present). This study differs from previous work in examining differences in model performance in detecting hip fractures in the presence of orthopedic implants. These results suggest that DCNN solutions can maintain excellent performance even in the presence of implants.

By virtue of geographical location, this study enrolled PXRs from a multi-ethnic local population. Considering recent studies demonstrating the ability of artificial intelligence deep learning models to predict ethnicity from medical images where human interpreters cannot, the authors were interested to examine if this model would perform differently based on ethnicity.[47] Nevertheless, sub-group analysis of model performance on the test set demonstrated sustained high model performance across the different ethnic groups aside from lower AUPRC in the Indian ethnic group possibly attributed to its small sample size. This study differs from previous work in examining the potential effect of ethnicity in the context of deep learning solutions for hip fracture detection and is the largest study to be conducted in an Asian population.

Grad-CAM allows the visualization of input image areas considered most important by the model in its predictions. While traditionally used for localization and identifying model bias,[48] prior studies have suggested Grad-CAM may assist physicians in identifying pathologic regions and improve user confidence by providing insight into the 'black box' model.[49,50] In this study, examination of Grad-CAM heatmaps showed that in 76.8%, class-discriminative regions correctly localized the fracture site. While relatively accurate, this is lower than the up to 95.9% concordance rate reported by prior studies.[27,28] The authors postulate that this may be related to other studies' exclusion of PXRs deemed to be of poor quality, including foreign bodies or with other fractures, compared to this study's comparatively wider inclusion criteria. Furthermore, this study found that Grad-CAM heatmaps have a tendency to outline the outer femoral neck despite the presence of a cortical disruption and fracture line, as opposed to delineating Shenton's line as observed by Oakden-Rayner et al.[27] (Figure 4). In several instances, the Grad-CAM heatmaps failed to localize the hip fracture entirely (Figure 2), in keeping with findings from other studies.[27,42] Given its unpredictable localization, Grad-CAM should be interpreted with caution. Future studies may wish to evaluate the effect of these heatmaps in user confidence and combined physician-model performance.

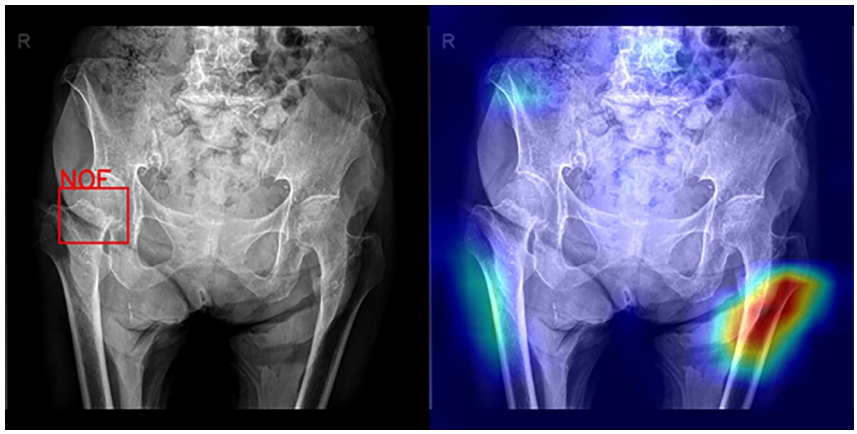


Figure 2. True positive predicted hip fracture with incorrect heatmap activation at the contralateral limb

This study utilized a large volume of 40,203 PXRs annotated with image-level labels for DCNN training. Compared to object detection algorithms, which have conventionally utilized bounding boxes to provide location supervisory signals, this study adopted an image classification approach with image-level labels. This allowed for a less labor-intensive process of annotation, which could be performed by just 2 radiologists. Such an approach is consistent with other recent studies, which have either used image classifiers or automated and simple methods of location annotation (e.g., point based annotation) for object detection.[28–31] Earlier studies using strongly supervised object detection with bounding box annotations of location supervisory signals have been limited to small numbers, with less favorable results.[32]

Beyond image labeling and training supervision, this network also demonstrates that prediction of hip fractures on PXR can be achieved with high accuracy when using the entire radiograph as the input, without the need for a separate localization network solution to first identify the hip joint, as in the work by Gale et al.[29] This study also demonstrates the ability to achieve comparable performance without the implementation of model pre-training, as in Cheng et al.'s work where a separate image set of limb radiographs was used to pre-train the algorithm.[28]

The authors were able to develop their own DCNN model using the publicly available DenseNet-121 architecture, along with PXRs and radiology reports obtained from a single tertiary institution. The entire study was carried out over a period of 12 months. The favorable performance achieved by this model demonstrates that it may be feasible for institutions to develop their own deep learning algorithms for computer aided diagnoses, based on patterns of local prevalence and local imaging parameters.

Limitations of the study

Limitations of this study include the inability to directly compare performance of our model against other models, particularly due to heterogeneity in inclusion criteria where patients with metallic implants and other non-hip fractures were often excluded from other studies. Performance of our model was also not externally validated using dataset from other institutions, to evaluate for generalizability. This may be of concern as our study benefitted from relative technical homogeneity of the PXR inputs, with all the radiographs performed using either one of two digital radiography systems from the same manufacturer (Philips Healthcare, Netherlands). Future studies may consider external validation such as in the work of Oakden-Rayner et al.,[27] and potentially comparison of different models using a common external test set. Finally, this study did not compare the performance of our network directly with physician performance.

Conclusions

In conclusion, this study demonstrates that a DCNN solution for CAD of hip fractures (neck of femur, trochanteric, and subtrochanteric) on PXRs developed using an image classifier approach can achieve good performance with high sensitivity, specificity and negative predictive value. This is possible even when including all PXRs regardless of technical quality, metallic implants or other concomitant pathology, unlike most previous work, and underscores the potential for DCNN solutions to reduce missed or delayed

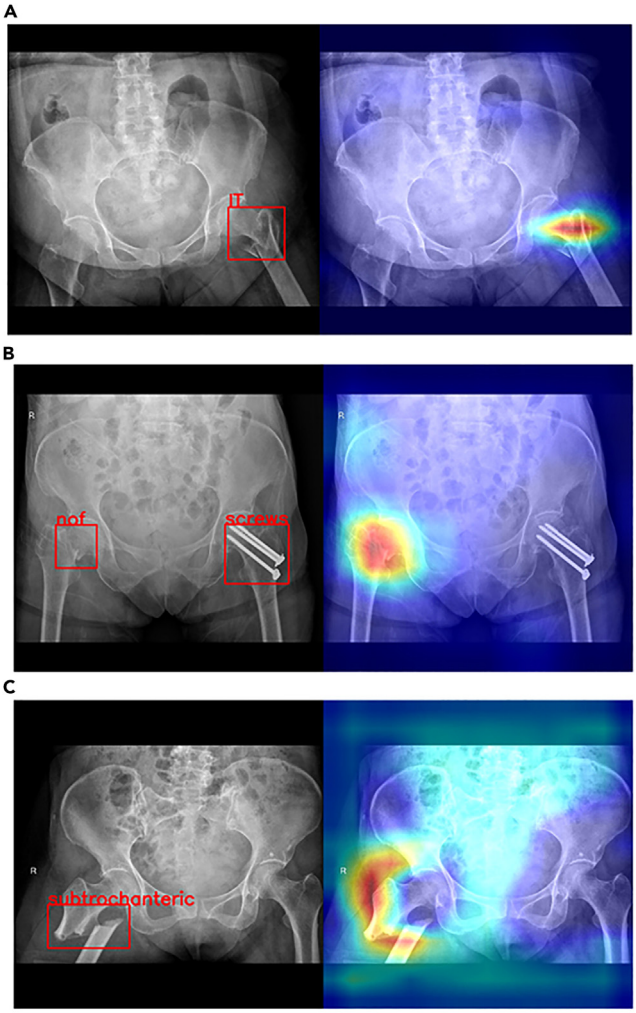


Figure 3. True positive predicted hip fractures with heatmap activation correctly located
(A) Intertrochanteric fracture.
(B) Femoral neck fracture.
(C) Subtrochanteric fracture.

diagnoses of hip fractures, particularly in the urgent and emergency care settings. This study also demonstrates that it is feasible for institutions to develop their own deep learning models based on local imaging parameters and disease patterns.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Study cohort and data
  - Labeling of PXRs and image pre-processing

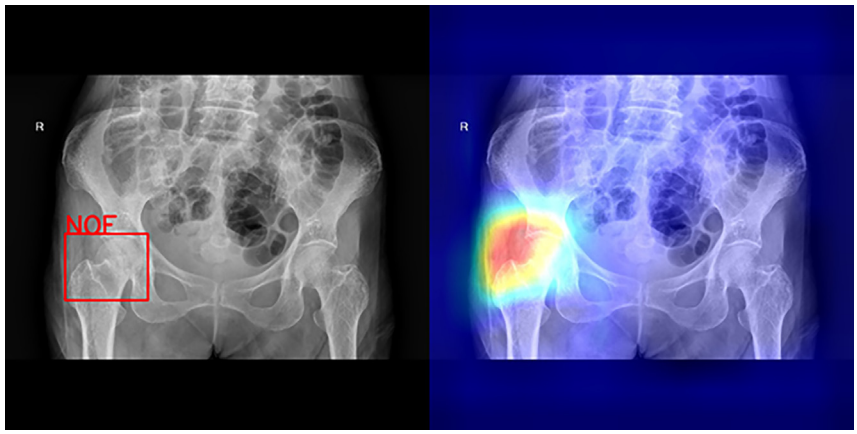


Figure 4. Example of Grad-CAM heatmaps' tendency to outline the lateral femoral neck

- Dataset preparation
- Model training and selection
- Model evaluation and statistical analysis

● QUANTIFICATION AND STATISTICAL ANALYSIS

ACKNOWLEDGMENTS

The authors have received funding from AI Singapore (AISG2-100E-2022-095) in the creation of this manuscript. The funder played no role in study design, data collection, analysis and interpretation of data or the writing of this manuscript.

AUTHOR CONTRIBUTIONS

List of Authors

Yan Gao (Y.G.), Nicholas Yock Teck Soh (N.S.), Nan Liu (N.L.), Gilbert Lim (G.L.), Daniel Ting (D.T.), Lionel Cheng (L.C.), Kang Min Wong (K.W.), Charlene Liew (C.L.), Hong Choon Oh (H.O.), Jin Rong Tan (J.R.), Narayan Venkataraman (N.V.), Siang Hiong Goh (S.G.), Yet Yen Yan (Y.Y.)

Indicated in the following rows are a list of contributions and the name(s) of the author(s) to whom they apply:

1. Conceptualization.

G.L., D.T., N.L., L.C., K.W., C.L., J.R., N.V., S.G., and Y.Y.

2. Methodology.

G.L., D.T., N.L., L.C., K.W., C.L., J.R., N.V., S.G., and Y.Y.

3. Software, Validation, Formal Analysis.

Y.G., N.S., N.L., H.O., and Y.Y.

4. Investigation, Resources, Data curation.

Y.G., N.S., G.L., D.T., H.O., L.C., K.W., C.L., J.R., N.V., S.G., and Y.Y.

5. Writing – Original Draft.

N.S.

6. Writing – Review & Editing.

Y.G., N.L., H.O., S.G., and Y.Y.

7. Visualization.

Y.G. and N.S.

8. Supervision.

Y.Y. and Y.G.

9. Project administration, Funding Acquisition.

Y.Y., N.L., and D.T.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: February 28, 2023
Revised: May 30, 2023
Accepted: July 6, 2023
Published: July 11, 2023

REFERENCES

1. Cooper, C., Cole, Z.A., Holroyd, C.R., Earl, S.C., Harvey, N.C., Dennison, E.M., Melton, L.J., Cummings, S.R., and Kanis, J.A.; IOF CSA Working Group on Fracture Epidemiology; Group on Fracture Epidemiology (2011). Secular trends in the incidence of hip and other osteoporotic fractures. Osteoporos. Int. *22*, 1277–1288.

2. Endo, Y., Aharonoff, G.B., Zuckerman, J.D., Egol, K.A., and Koval, K.J. (2005). Gender differences in patients with hip fracture: a greater risk of morbidity and mortality in men. J. Orthop. Trauma *19*, 29–35.

3. Hagino, H., Endo, N., Harada, A., Iwamoto, J., Mashiba, T., Mori, S., Ohtori, S., Sakai, A., Takada, J., and Yamamoto, T. (2017). Survey of hip fractures in Japan: recent trends in prevalence and treatment. J. Orthop. Sci. *22*, 909–914.

4. Mundi, S., Pindiprolu, B., Simunovic, N., and Bhandari, M. (2014). Similar mortality rates in hip fracture patients over the past 31 years: A systematic review of RCTs. Acta Orthop. *85*, 54–59.

5. Brauer, C.A., Coca-Perraillon, M., Cutler, D.M., and Rosen, A.B. (2009). Incidence and mortality of hip fractures in the United States. JAMA *302*, 1573–1579.

6. Braithwaite, R.S., Col, N.F., and Wong, J.B. (2003). Estimating hip fracture morbidity, mortality and costs. J. Am. Geriatr. Soc. *51*, 364–370.

7. Craik, R.L. (1994). Disability following hip fracture. Phys. Ther. *74*, 387–398.

8. Cooper, C. (1997). The crippling consequences of fractures and their impact on quality of life. Am. J. Med. *103*. discussion 17S-19S.

9. Ibrahim, N.I., Ahmad, M.S., Zulfarina, M.S., Zaris, S.N.A.S.M., Mohamed, I.N., Mohamed, N., Mokhtar, S.A., and Shuid, A.N. (2018). Activities of daily living and determinant factors among older adult subjects with lower body fracture after discharge from hospital: a prospective study. Int. J. Environ. Res. Publ. Health *15*, 1002.

10. Dinamarca-Montecinos, J.L., Prados-Olleta, N., Rubio-Herrera, R., Castellón-Sánchez del Pino, A., and Carrasco-Buvinic, A. (2015). Intra-and extracapsular hip fractures in the elderly: Two different pathologies? Rev. Española Cirugía Ortopédica Traumatol. *59*, 227–237.

11. Cannon, J., Silvestri, S., and Munro, M. (2009). Imaging choices in occult hip fracture. The Journal of emergency medicine. J. Emerg. Med. *37*, 144–152.

12. Young, J.W., Burgess, A.R., Brumback, R.J., and Poka, A. (1986). Pelvic fractures: value of plain radiography in early assessment and management. Radiology *160*, 445–451.

13. Dominguez, S., Liu, P., Roberts, C., Mandell, M., and Richman, P.B. (2005). Prevalence of traumatic hip and pelvic fractures in patients with suspected hip fracture and negative initial standard radiographs—a study of emergency department patients. Acad. Emerg. Med. *12*, 366–369.

14. Macri, F., Niu, B.T., Erdelyi, S., Mayo, J.R., Khosa, F., Nicolaou, S., and Brubacher, J.R. (2022). Impact of 24/7 Onsite Emergency Radiology Staff Coverage on Emergency Department Workflow. Assoc Radiol J *73*, 249–258.

15. Paul, P., and Issac, R.T. (2018). Delay in time from fracture to surgery: a potential risk factor for in-hospital mortality in elderly patients with hip fractures. J. Orthop. *15*, 375–378.

16. Vertelis, A., Robertsson, O., Tarasevicius, S., and Wingstrand, H. (2009). Delayed hospitalization increases mortality in displaced femoral neck fracture patients. Acta Orthop. *80*, 683–686.

17. Barahona, M., Barrientos, C., Cavada, G., Brañes, J., Martinez, Á., and Catalan, J. (2020). Survival analysis after hip fracture: higher mortality than the general population and delayed surgery increases the risk at any time. Hip Int. *30*, 54–58.

18. Carrino, J.A., Unkel, P.J., Miller, I.D., Bowser, C.L., Freckleton, M.W., and Johnson, T.G.

(1998). Large-scale PACS implementation. J. Digit. Imag. *11*, 3–7.

19. Bradley, W.G. (2012). Teleradiology. Neuroimaging Clin. *22*, 511–517.

20. Chan, H.P., Samala, R.K., Hadjiiski, L.M., and Zhou, C. (2020). Deep learning in medical image analysis *3–21*. Deep Learning in Medical Image Analysis.

21. Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., and Khan, M.K. (2018). Medical image analysis using convolutional neural networks: a review. J. Med. Syst. *42*, 226–233.

22. Currie, G., Hawk, K.E., Rohren, E., Vial, A., and Klein, R. (2019). Machine learning and deep learning in medical imaging: intelligent imaging. J. Med. Imag. Radiat. Sci. *50*, 477–487.

23. Treder, M., Lauermann, J.L., and Eter, N. (2018). Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. Graefes Arch. Clin. Exp. Ophthalmol. *256*, 259–265.

24. Rahman, T., Chowdhury, M.E.H., Khandakar, A., Islam, K.R., Islam, K.F., Mahbub, Z.B., Kadir, M.A., and Kashem, S. (2020). Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. Appl. Sci. *10*, 3233.

25. Reddy, N.E., Rayan, J.C., Annapragada, A.V., Mahmood, N.F., Scheslinger, A.E., Zhang, W., and Kan, J.H. (2020). Bone age determination using only the index finger: a novel approach using a convolutional neural network compared with human radiologists. Pediatr. Radiol. *50*, 516–523.

26. Ciritsis, A., Rossi, C., Eberhard, M., Marcon, M., Becker, A.S., and Boss, A. (2019). Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. Eur. Radiol. *29*, 5458–5468.

27. Oakden-Rayner, L., Gale, W., Bonham, T.A., Lungren, M.P., Carneiro, G., Bradley, A.P., and Palmer, L.J. (2022). Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet. Digit. Health *4*, e351–e358.

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

29. Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A.P., and Palmer, L.J. (2017). Detecting Hip Fractures with Radiologist-Level Performance Using Deep Neural Networks. https://doi.org/10.48550/ARXIV.1711.06504.

30. Cheng, C.-T., Wang, Y., Chen, H.-W., Hsiao, P.-M., Yeh, C.-N., Hsieh, C.-H., Miao, S., Xiao, J., Liao, C.-H., and Lu, L. (2021). A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nat. Commun. *12*, 1066.

31. Yu, J.S., Yu, S.M., Erdal, B.S., Demirer, M., Gupta, V., Bigelow, M., Salvador, A., Rink, T., Lenobel, S.S., Prevedello, L.M., and White, R.D. (2020). Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. Clin. Radiol. *75*, 237.e1–237.e9.

32. Krogue, J.D., Cheng, K.V., Hwang, K.M., Toogood, P., Meinberg, E.G., Geiger, E.J., Zaid, M., McGill, K.C., Patel, R., Sohn, J.H., et al. (2020). Automatic Hip Fracture Identification and Functional Subclassification with Deep Learning. Radiol. Artif. Intell. *2*, e190023.

33. Mawatari, T., Hayashida, Y., Katsuragawa, S., Yoshimatsu, Y., Hamamura, T., Anai, K., Ueno, M., Yamaga, S., Ueda, I., Terasawa, T., et al. (2020). The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs. Eur. J. Radiol. *130*, 109188.

34. Lex, J.R., Di Michele, J., Koucheki, R., Pincus, D., Whyne, C., and Ravi, B. (2023). Artificial Intelligence for Hip Fracture Detection and Outcome Prediction. JAMA Netw. Open *6*, e233391.

35. Cha, Y., Kim, J.-T., Park, C.-H., Kim, J.-W., Lee, S.Y., and Yoo, J.-I. (2022). Artificial intelligence and machine learning on diagnosis and classification of hip fracture: systematic review. J. Orthop. Surg. Res. *17*, 520–523.

36. Kitamura, G. (2020). Deep learning evaluation of pelvic radiographs for position, hardware presence, and fracture detection. Eur. J. Radiol. *130*, 109139.

37. Murphy, E.A., Ehrhardt, B., Gregson, C.L., von Arx, O.A., Hartley, A., Whitehouse, M.R., Thomas, M.S., Stenhouse, G., Chesser, T.J.S., Budd, C.J., and Gill, H.S. (2022). Machine learning outperforms clinical experts in classification of hip fractures. Sci. Rep. *12*, 2058.

38. Bae, J., Yu, S., Oh, J., Kim, T.H., Chung, J.H., Byun, H., Yoon, M.S., Ahn, C., and Lee, D.K. (2021). External Validation of Deep Learning Algorithm for Detecting and Visualizing Femoral Neck Fracture Including Displaced and Non-displaced Fracture on Plain X-ray. J. Digit. Imag. *34*, 1099–1109.

39. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

40. Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980.

41. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., and Kudlur, M. (2016). {TensorFlow}: a system for {Large-Scale} machine learning. 12th USENIX symposium on operating systems design and implementation, 265–283.

42. Mehr, G. (2020). Automating Abnormality Detection in Musculoskeletal Radiographs through Deep Learning. arXiv preprint arXiv:2010.12030.

43. Youden, W.J. (1950). Index for rating diagnostic tests. Cancer *3*, 32–35.

44. Parker, M.J. (1992). Missed hip fractures. Arch. Emerg. Med. *9*, 23–27.

45. Cohen, J.F., and McInnes, M.D.F. (2022). Deep Learning Algorithms to Detect Fractures: Systematic Review Shows Promising Results but Many Limitations. Radiology *304*, 63–64.

46. Khawar, H., Craxford, S., Marson, B.A., Rahman, H.P., and Ollivere, B. (2021). Outcomes after hip fractures sustained in hospital: A propensity-score matched cohort study. Injury *52*, 2356–2360.

47. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.C., et al. (2022). AI recognition of patient race in medical imaging: a modelling study. Lancet. Digit. Health *4*, e406–e414.

48. Selvaraju, R.R., Cogswell, M., Gupta, A., and Parikh, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

49. van der Velden, B.H.M., Kuijf, H.J., Gilhuijs, K.G.A., and Viergever, M.A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med. Image Anal. *79*, 102470.

50. Chen, H.Y., Hsu, B.W.Y., Yin, Y.K., Lin, F.H., Yang, T.H., Yang, R.S., Lee, C.K., and Tseng, V.S. (2021). Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. PLoS One *16*, e0245992.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
DenseNet-121	Huang et al.[39]	arXiv:1608.06993
Tensorflow	Abadi et al.[41]	arXiv:1605.08695

RESOURCE AVAILABILITY

Lead contact

- Further information and requests for resources should be directed to and will be fulfilled by the lead contact Yet Yen Yan (yan.yet.yen@singhealth.com.sg)

Materials availability

- Iteration of the trained deep learning model is available on request from the lead contact.
- There are restrictions to the availability of patients' medical images used in training of the deep learning model due to institutional and legal regulations over patient confidentiality and imaging.

Data and code availability

- Patients' medical images used in training of the deep learning model in this study cannot be deposited in a public repository due to institutional and legal regulations over patient confidentiality and imaging. To request access, contact the lead contact.
- This paper utilizes publicly available code in development of the deep learning model.[40] Iteration of the trained deep learning model available on request from the lead contact.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

- This paper uses de-identified medical images of patients who had pelvic radiographs performed at a single institution within the study period.
- Description of overall age, gender and ethnicity of the patient population is described in Table 1 of the manuscript.
- Waiver of full ethical deliberation was provided by the host institution's Centralised Institutional Review Board (CIRB), with the study conducted using deidentified data.

METHOD DETAILS

Study cohort and data

Frontal pelvic radiographs (PXRs) performed across ambulatory, inpatient and emergency department settings at a single tertiary teaching hospital between January 2016 and December 2020 were included in this study. All PXRs were acquired using DigitalDiagnost C90 and DigitalDiagnost 4 High Performance digital radiography systems (Philips Healthcare, Netherlands). Waiver of full ethical deliberation was provided by the institution's Centralised Institutional Review Board (CIRB), with the study conducted using only deidentified data.

A total of 40,203 PXRs were extracted from the institution's radiology picture archiving and communications system, of which 18,803 were performed in the emergency department and 21,400 were performed in the ambulatory or inpatient setting. The radiologist's report for each included PXR was also extracted along with radiology reports for any subsequent examinations performed for the patient in the following

6 months. This included reports for subsequent radiographs as well as other modalities such as computed tomography or magnetic resonance imaging where available.

Labeling of PXRs and image pre-processing

All 40,203 PXRs and their accompanying radiology reports were de-identified with any post-contemporaneous labels on these images removed. Presence of a hip fracture was defined as any fracture involving the proximal femora. This included intracapsular (neck of femur) and extracapsular (trochanteric or subtrochanteric) fractures.

To determine ground truth labels for each PXR (i.e., hip fracture present or absent), all PXRs were individually read by 2 board-certified consultant musculoskeletal subspecialty radiologists (YYY and JRT with 10 and 7 years of radiology experience respectively) blinded to the accompanying radiology reports. PXRs with hip fractures were labeled positive, while those without hip fractures were labeled negative. In instances where diagnosis of hip fracture was in doubt, the accompanying PXR report and radiology reports for all imaging performed in the following 6 months were reviewed for any mention of a hip fracture. Final decision on presence or absence of a hip fracture was made by consensus between the 2 musculoskeletal radiologists.

PXRs included in this study were formatted in 16-bit monochrome and ranged from 1512 x 2042 pixels to 2899 x 3254 pixels. To reduce image heterogeneity and computational complexity, all PXRs were uniformly padded to square shapes and resized to 512 x 512 pixels.

Dataset preparation

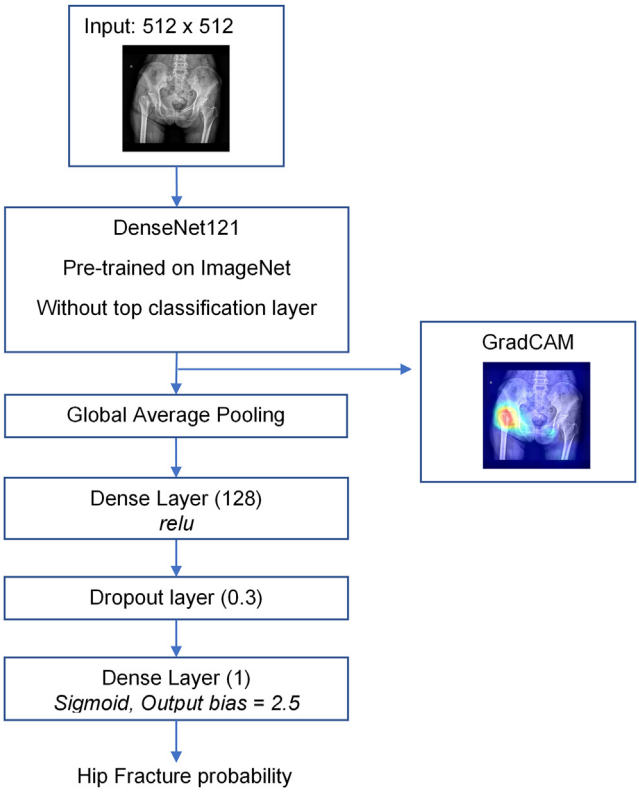
A total of 40,203 PXRs were included in this study. Among the 18,803 PXRs performed in the emergency department, 3,761 (20%) were first randomly selected to form the hold-out test dataset for subsequent evaluation of our hip fracture detection model. A test set consisting only of emergency department PXRs was deliberately chosen to better approximate the potential real-world use scenario, where patients with suspected hip fractures present almost exclusively to emergency departments or urgent care facilities.

Of the remaining 36,442 PXRs in our study, 29,153 (80%) and 7,289 (20%) were randomly allocated for training and validation of our model respectively (Figure 1).

Model training and selection

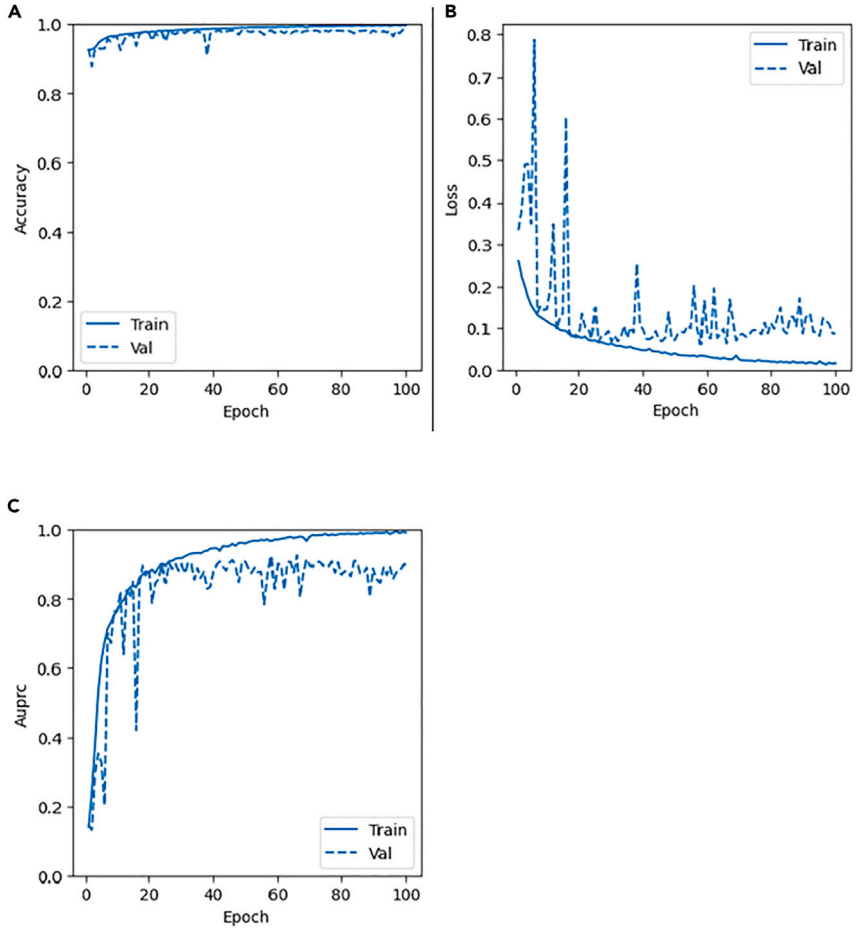
DCNNs are a subset of deep neural networks commonly used in object detection and image classification. Images are received as inputs and multiple convolutional, activation, pooling and fully connected layers are utilized to produce an image classifier. This study employed a DenseNet-121 architecture with pre-trained ImageNet weights,[39] which was then trained on the training set of 29,153 PXRs. DenseNet-121 was selected due to its comparatively fewer parameters and faster training time coupled with similar performance compared to other CNNs.[42]

Model training was performed via stochastic gradient descent using the adaptive moment estimation (Adam) optimiser, for 100 epochs at a learning rate of 0.0001.[40] The loss function for optimization was binary cross-entropy loss: $L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p(y_i)) + (1 - y_i)\log(1 - p(y_i))$, where i is the number of images in the training set and $y_i = 1$ if it is a fracture and 0 otherwise. $p(y_i)$ is given by the sigmoid activation function with proper initial bias. Due to the highly unbalanced positive and negative classes, initial bias was set to the log of ratio between number of negative images and number of positive images in the training set, which is approximately 2.5. A larger batch size of 32 was employed to ensure that each batch included positive images. A network diagram is shown in Figure. During training of the model, augmentation of PXRs was performed with random operations of zoom (10%), horizontal and vertical translations (100, −100 pixels), horizontal flips, rotations (15°, −15°) and brightness jittering (25%). As per Oakden-Rayner et al., it was shown that each augmentation technique contributed to absolute AUC improvement of around 0.01.[27] On our dataset, augmentation improves the overall area under precision-recall curve (AUPRC) by about 0.04.



Architecture diagram of the DCNN

Model validation was performed using the validation set of 7,289 PXRs, with model parameters selected from the epoch (epoch 66) with highest AUPRC. Changes in accuracy, loss, and AUPRC during the training process are shown in Figure. The network was trained using TensorFlow on a workstation with NVIDIA Quadro TRX 6000 GPU with 32 GB DDR4.[40] Training time was approximately 40 h for 100 epochs.



Performance in the training and validation datasets

(A) Accuracy change during training.

(B) Loss change during training.

(C) AUPRC change during training.

Model evaluation and statistical analysis

The trained model was evaluated using the test set of 3,761 PXRs, (with 463 or 12.3% positive for hip fracture), using parameters of accuracy, sensitivity (recall), specificity, positive predictive value (precision) and negative predictive value, with their respective formulae as listed below: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ $Sensitivity = \frac{TP}{TP+FN}$ $Specificity = \frac{TN}{TN+FP}$ $PPV = \frac{TP}{TP+FP}$ $NPV = \frac{TN}{TN+FN}$. Where TP = true positive, TN = true Negative, FP = false positive and FN = false negative. These together with area under receiver operating characteristics (AUROC) and area under precision-recall curve (AUPRC) were used to compare the model against prior published work on automated hip fracture detection.

In instances where the model predicted a hip fracture, Grad-CAM heatmap was produced and overlaid on the associated PXR. Fused images were individually reviewed by the 2 board certified musculoskeletal radiologist to assess if the fractures had been correctly localized by the model.

QUANTIFICATION AND STATISTICAL ANALYSIS

- The trained model was evaluated using the test set of 3,761 PXRs, (with 463 or 12.3% positive for hip fracture), using parameters of accuracy, sensitivity (recall), specificity, positive predictive value (precision) and negative predictive value. These together with area under receiver operating characteristics (AUROC) and area under precision-recall curve (AUPRC) were used to compare the model against prior published work on automated hip fracture detection.

- In instances where the model predicted a hip fracture, Grad-CAM heatmap was produced and overlaid on the associated PXR.
- Descriptive statistics were used to summarize population characteristics. Categorical variables were summarized using frequencies (percentages), and continuous variables using mean (standard deviation) after assessing for normality.