OPEN

A deep learning approach using an ensemble model to autocreate an image-based hip fracture registry

Jacobien H.F. Oosterhoff, MD, PhD[a,b,*], Soomin Jeon, PhD[a,c], Bardiya Akhbari, PhD[a], David Shin, BS[a], Daniel G. Tobert, MD[a], Synho Do, PhD[d], Soheil Ashkani-Esfahani, MD[a,e]

Abstract

Objectives: With more than 300,000 patients per year in the United States alone, hip fractures are one of the most common injuries occurring in the elderly. The incidence is predicted to rise to 6 million cases per annum worldwide by 2050. Many fracture registries have been established, serving as tools for quality surveillance and evaluating patient outcomes. Most registries are based on billing and procedural codes, prone to under-reporting of cases. Deep learning (DL) is able to interpret radiographic images and assist in fracture detection; we propose to conduct a DL-based approach intended to autocreate a fracture registry, specifically for the hip fracture population.

Methods: Conventional radiographs (n = 18,834) from 2919 patients from Massachusetts General Brigham hospitals were extracted (images designated as hip radiographs within the medical record). We designed a cascade model consisting of 3 submodules for image view classification (MI), postoperative implant detection (MII), and proximal femoral fracture detection (MIII), including data augmentation and scaling, and convolutional neural networks for model development. An ensemble model of 10 models (based on ResNet, VGG, DenseNet, and EfficientNet architectures) was created to detect the presence of a fracture.

Results: The accuracy of the developed submodules reached 92%–100%; visual explanations of model predictions were generated through gradient-based methods. Time for the automated model-based fracture–labeling was 0.03 seconds/image, compared with an average of 12 seconds/image for human annotation as calculated in our preprocessing stages.

Conclusion: This semisupervised DL approach labeled hip fractures with high accuracy. This mitigates the burden of annotations in a large data set, which is time-consuming and prone to under-reporting. The DL approach may prove beneficial for future efforts to autocreate construct registries that outperform current diagnosis and procedural codes. Clinicians and researchers can use the developed DL approach for quality improvement, diagnostic and prognostic research purposes, and building clinical decision support tools.

Keywords: hip fracture, fracture registry, image-based registry, machine learning, deep learning

1. Introduction

Hip fractures are one of the most common fractures in the elderly, with more than 300,000 patients per year in the United States alone, predicted to rise to an incidence of 6 million cases annually worldwide in 2050.[1] Many orthopaedic registries have been established, serving as tools for both quality surveillance and evaluating patient outcomes.[2] Data collection is typically conducted manually, through the use of registration forms, or in an automated fashion by filtering diagnosis and procedural codes in the medical record. Data collection is prone to human error, and diagnosis and procedural codes may under-report or misreport clinical conditions.[3]

Machine learning (ML) is rapidly advancing in almost every field within health care. These methods have been shown to outperform methods of data abstraction based on diagnosis and procedural codes in free-text reports, which otherwise could have

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, or patent/licensing arrangements) that might pose a conflict of interest in connection with the submitted article.

[a] Department of Orthopaedic Surgery, Massachusetts General Hospital and Harvard Medical School, Boston, MA, [b] Department of Engineering Systems and Services, Faculty Technology Policy Management, Delft University of Technology, Delft, the Netherlands, [c] Department of Information Sciences and Mathematics, Dong-A University, Busan, South Korea, [d] Laboratory of Medical Imaging and Computation, Massachusetts General Hospital and Harvard Medical School, Boston, MA, [e] Foot & Ankle Research and Innovation Laboratory, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[*] Corresponding author. Address: Delft University of Technology—Faculty TPM, Postbus 5015, 2600 GA Delft, the Netherlands. E-mail address: j.h.f.oosterhoff@tudelft.nl (J.H.F. Oosterhoff).

Source of funding: Nil.

Ethical Review Committee Statement: All study procedures were approved by our institutional review board prior study start-up.

Investigation performed at Massachusetts General Hospital, Harvard Medical School, Boston, MA.

J.H.F. Oosterhoff, S. Jeon, B. Akhbari share first authorship.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.otainternational.org).

OTAI (2023) e283

Received: 22 May 2023 / Accepted: 17 July 2023
Published online 22 December 2023

http://dx.doi.org/10.1097/OI9.0000000000000283

led to under-reporting.[4] Computer vision using deep learning (DL) algorithms outperforms clinicians at some radiology, pathology, and dermatology diagnoses and helps avoid human bias and limitations in more technical endeavors.[5–7] In orthopaedic trauma, DL algorithms are already on par with humans at recognizing wrist, hand, and ankle fractures on radiographs obtained in the emergency department (83% accuracy) and will likely improve.[8–10] DL has also been used to detect, locate, and classify proximal femoral fractures on conventional radiography demonstrating comparable performance with radiologists and orthopaedic surgeons.[11,12]

Because DL has shown to be successful in interpreting radiographs and assisting in fracture detection, we, therefore, aimed to implement a DL-based approach to create an image-based fracture registry, focusing specifically on the hip fracture population.

2. Methods

2.1. Data Source

This retrospective cohort study was approved and registered with the institutional review board (IRB) prior study start-up. A search in the Research Patient Data Registry (RPDR) by using International Classification of Disease (ICD) codes was performed to identify patients sustaining a proximal femur fracture in our institutions from January 2010 through December 2018. RPDR collects medical records from institutions within the Partners Healthcare System and may be queried after IRB approval. Our institutions accounted for 2 level I trauma centers and 3 community (non-level I trauma) hospitals. Patients who (1) presented with proximal femoral fracture, (2) had available injury radiographs of adequate quality (that included the entire injury and with adequate penetration; rotation and angulation issues typical of initial postinjury radiographs were not a reason for exclusion), and (3) were 18 years or older were included.

2.2. Data Preprocessing

In total, 106,381 images were retrieved. Of those, 82% (87,487 of 106,381 images) were other imaging modalities (eg, computed tomography, ultrasound, magnetic resonance imaging, and DEXA scan) than conventional radiography and, therefore, were excluded for modeling. The remaining images (n = 18,834) were derived from 2919 patients (Fig. 1), of whom 60.2% (1757 patients) were female, and most patients were 60 years or older with 78.3% (2286 of 2919 patients) (Table 1).

All Digital Imaging and Communications in Medicine (DICOM) images were converted into a Portable Network Graphics (PNG) image and anonymized. Images were loaded in Python (Python Software Foundation) with *uint16*. Image enhancement of all images was conducted by applying Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve the visibility level of the images with a clip limit of 5 and a tile grid size 4 by 4.[13]

2.3. Cascade Model

We designed a cascade model (Fig. 2) consisting of 3 submodules for image view classification (MI), postoperative implant detection (MII), and fracture detection (MIII). In total, 3100 radiographs (MI: 1,169, MII: 660, MIII: 1284) were manually labeled by 2 clinical experts, assisted by manual chart review (J.H.F.O., S.A-E.). The images were randomly divided into training, validation, and test sets in a 60/20/20 split.

2.4. Module I: Image View Classification

The first task included classifying the image view into 3 classes: anteroposterior (AP) view of the left proximal femur, AP view of both proximal femurs (bilateral), and AP view of the right proximal femur. In total, 1066 images were manually labeled by physicians: AP view left 47.0% (n = 501), AP view bilateral 29.6% (n = 316), and AP view right 23.4% (n = 249). A convolutional neural network (CNN)-based model architecture (Fig. 2) was used to determine whether the input image is a left, bilateral, or right proximal femoral image. We trained the CNN on the multiclassification task, meaning that the output is a probability value for one of the 3 classes, representing the model's confidence in the prediction, and a rectified linear function (ReLU) was used as the activation function.

2.5. Module II: Postoperative Implant or Internal Fixation Detection

The second task included classifying images in the following 2 categories: images with an implant in the image (ie, hip replacement or internal fixation) and images without in implant

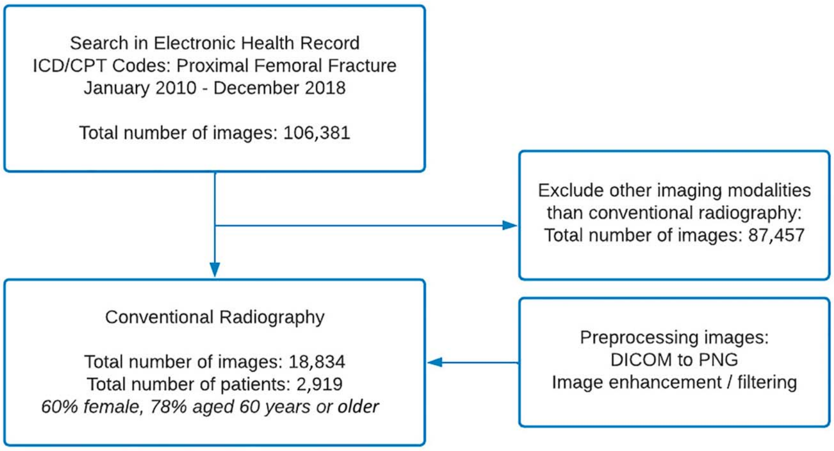
Search in Electronic Health Record
ICD/CPT Codes: Proximal Femoral Fracture
January 2010 - December 2018
Total number of images: 106,381
Exclude other imaging modalities than conventional radiography: Total number of images: 87,457
Conventional Radiography
Total number of images: 18,834
Total number of patients: 2,919
60% female, 78% aged 60 years or older
Preprocessing images: DICOM to PNG
Image enhancement / filtering

Figure 1. Flow diagram for patient inclusion.

TABLE 1

Baseline Characteristics of Patients With Medical Images (Conventional Radiography) for Analysis

	Total Number of Patients and Medical Images		Task 1: Image View–Labeled Images		Task 2: Implant Detection–Labeled Images		Task 3 Teacher: Fx Detection–Labeled Images		Task 3 Student: Fx Detection–Unlabeled Images	
	Patients, N (%)	Images, N (%)	Patients, N (%)	Images, N (%)	Patients, N (%)	Images, N (%)	Patients, N (%)	Images, N (%)	Patients, N (%)	Images, N (%)
Total cohort	2919	18,834	612	1090*	466	820**	924	1284***	595	874
Sex										
Female	1757 (60.2%)	11,526 (61.2%)	377 (61.6%)	681 (62.5%)	296 (63.5%)	537 (65.5%)	347 (37.6%)	485 (37.8%)	237 (39.8%)	365 (41.8%)
Male	1162 (39.8%)	7308 (38.8%)	235 (38.4%)	409 (37.5%)	170 (36.5%)	283 (34.5%)	577 (62.4%)	799 (62.2%)	358 (60.2%)	509 (58.2%)
Age, yr										
<50	347 (11.9%)	1876 (10.0%)	74 (7.8%)	137 (12.6%)	44 (12.3%)	68 (8.3%)	92 (10.0%)	123 (9.6%)	50 (8.4%)	63 (7.2%)
50-59	286 (9.8%)	1888 (10.0%)	47 (5.0%)	85 (7.8%)	36 (8.4%)	66 (8.0%)	84 (9.1%)	117 (9.1%)	65 (10.9%)	94 (10.8%)
60-69	471 (16.1%)	2998 (15.9%)	108 (11.4%)	174 (16.0%)	84 (19.6%)	141 (17.2%)	156 (16.9%)	231 (18.0%)	106 (17.8%)	151 (17.3%)
70-79	584 (20.0%)	4032 (21.4%)	123 (13.0%)	232 (21.3%)	106 (24.8%)	181 (22.1%)	190 (20.6%)	270 (21.0%)	123 (20.7%)	203 (23.2%)
80-89	809 (27.7%)	5254 (27.9%)	184 (19.5%)	323 (29.6%)	140 (32.7%)	266 (32.4%)	275 (29.8%)	375 (29.2%)	173 (29.1%)	250 (28.6%)
90+	422 (14.5%)	2786 (14.8%)	76 (8.1%)	139 (12.8%)	56 (13.1%)	98 (12.0%)	127 (13.7%)	168 (13.1%)	78 (13.1%)	113 (12.9%)

Fx = fracture; N = number.
* AP view L = 277 (25.4%); AP view bilateral = 460 (42.2%); AP view right = 353 (32.4%).
** Implant = 424 (51.7%); no implant = 396 (48.3%).
*** Noted as fracture/control: Teacher = 879 (68.5%)/405 (31.5%).

in the image. In total, we manually labeled 536 images, of which 52.6% (n = 282) contained a postoperative hip implant. The images were randomly divided into a training, validation, and test set in a 60/20/20 split.

A pretrained ResNet50 architecture was used for training on the training set; and validation and evaluation were performed on the validation and test sets. ResNet50 is a CNN that is 50 layers deep, including one MaxPool and one Average Pool layer (including 26 million parameters).[14]

2.6. Module III: Proximal Femoral Fracture Detection

The third, and most clinically relevant, task included distinguishing images with a proximal femoral fracture from those

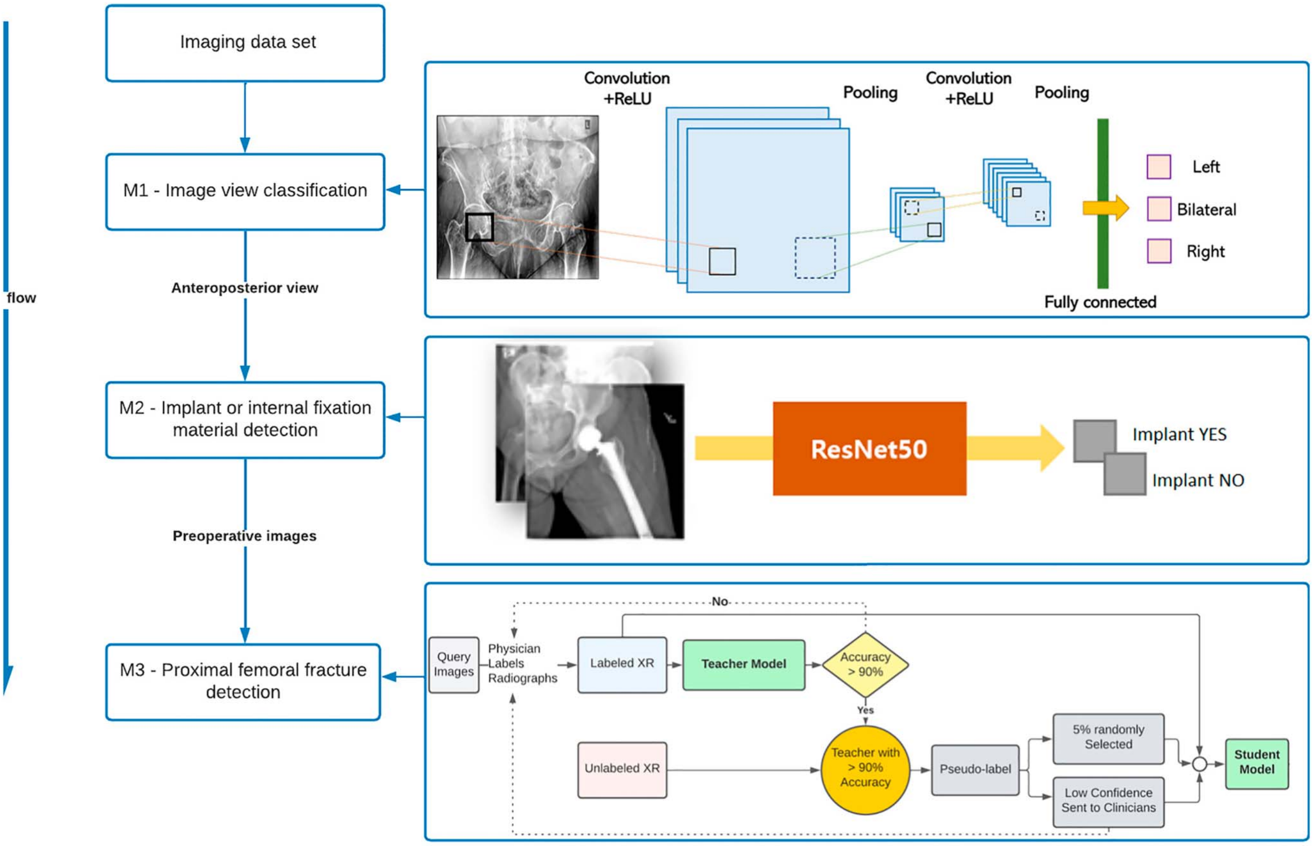

Figure 2. Cascade model flow diagram for autocreating an image-based hip fracture registry.

without fracture on AP view images. First, 1284 images were manually labeled (derived from 1107 patients; 68.5% of the images contained a proximal femoral fracture (n = 879 for 846 patients (~one image per person)) and 31.5% of the images were control (n = 405 images for 272 patients (~1.5 images per person))). An ensemble model of 10 models (developed based on ResNet, VGG, DenseNet, and EfficientNet architectures) was created to detect the presence of a proximal femoral fracture or not (ie, control vs. fracture). An ensemble method was applied to optimize *MIII*. The following 12 DL algorithms were trained on the training set: 6 ResNet50 (same architectures but with random weight initializations), 1 DenseNet201 (a pretrained CNN using dense connections between layers, 201 layers deep), 1 VGG19 (a pretrained CNN, 19 layers deep), 1 EfficientNetB5 (a pretrained CNN applying a compound coefficient, including 30 million parameters), 1 EfficientNetL2 (a pretrained CNN applying a compound coefficient, including 480 million parameters), and 2 EfficientNetB7 architectures.

To improve the model and automate the annotation for *MIII* in a self-supervised manner, a student-teacher technique (ie, noisy student[15] with EfficientNetB5 as a *Teacher* and EfficientNetB7 as a *Student*) was used. Then, 874 unlabeled images were presented to the *teacher model* and the labels were predicted. Of those, all images with a confidence lower than 80% (n = 29 radiographs) and 5% randomly selected images (n = 40) were manually checked (J.H.F.O.) (Fig. 2).

2.7. Model Performance Metrics

Each model was assessed using the following performance metrics (on both the validation and test sets): area under the receiver-operating characteristic curve (AUC-ROC), accuracy, area under the precision-recall curve (AUC-PRC), sensitivity (recall), specificity, positive predictive value (also known as precision), negative predictive value, F1 score, positive likelihood ratio, and negative likelihood ratio. The primary outcome of interest was the AUC-ROC. The secondary outcomes of interest were the other model performance metrics as mentioned above (Supplementary Table 1, http://links.lww.com/OTAI/A84).

2.8. Model Performance Visual Explanations

Visual explanations of the model's predictions were conducted using gradient-based methods. Gradient-weighted class activation mapping (Grad-CAM) uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting important regions in the image for predicting the concept.[16] A saliency map computes pixel-level importance for a given prediction and gives color to the contribution of the classification.[17]

3. Results

3.1. Module I: Image View Classification

The first model achieved 98.82% accuracy for the unseen test data set (Table 2). However, one bilateral hip image was actually misclassified initially as a right hip image and was later correctly identified by Module I. Therefore, the actual accuracy of the model is 100%. The time per the prediction of an image was on average 0.03 seconds per image.

3.2. Module II—Postoperative Implant or Internal Fixation Detection

The second model had an AUC-ROC of 0.95 and achieved an accuracy of 100% (Table 2). The time per prediction of an image was on average 0.03 seconds per image. An example of a heatmap provided by the Grad-CAM method showed that the model is identifying the implant or control cases accurately (Fig. 3).

3.3. Module III—Proximal Femoral Fracture Detection

The ensemble model yielded an AUC-ROC of 0.96 and accuracy of 91.9% (Table 3). The single-teacher model (based on EfficientNetB5) led to an AUC-ROC of 0.95 and accuracy of 92.3%. The unsupervised method applied in the *student* model (EfficientNetB7) yielded an AUC-ROC of 0.97 and accuracy of 91.9%. Examples of the model's output with saliency mapping showed that the *student* model is identifying the fracture and control cases accurately (Fig. 3).

When the 29 low-confidence unlabeled images were reviewed by a physician, 8 images were flagged as a mismatch. Among those, 4 images were subcapital fractures, 1 image was a severe osteoarthritis case, and 1 image had a pubic rami fracture. The 5% subsampled radiographs were reviewed (Supplementary Table 2, http://links.lww.com/OTAI/A85).

4. Discussion

The aim of this study was to develop a DL-based approach to autocreate an image-based hip fracture registry. The semi-supervised DL approach showed high accuracy, which may mitigate the laborious burden of constructing a data set or registry. The approach developed here may prove beneficial for future efforts to construct broad DL-based orthopaedic registries that outperform current diagnosis and procedural codes.

The results of this study should be viewed in light of several limitations with the main issue of domain adaptation. Owing to this issue, deep learning–based models need to be externally validated to prevent data inconsistencies. In addition, we convert DICOM images to PNG format in preprocessing (DICOM images typically contain between 4096 and 65,536-pixel values while PNG has only 256-pixel values). In the future, DICOM format can be used to improve the model (a challenge would be the computational power = GPU and CPU memory/time to train the model). Moreover, in this study, we only evaluated the images

TABLE 2

Model Performance of M1 and M2 Tasks on the Hold-out Test Set

	Task 1	Task 2
AUC-ROC	—	95.5%
Accuracy	100%	100%
AUC-PRC	—	95.3%
Sensitivity	100%	100%
Specificity	100%	100%
PPV	100%	100%
NPV	100%	100%
F1 score	100%	100%
LR +	Inf	Inf
LR -	0%	0%

AUC-PRC = area under the precision-recall curve; AUC-ROC = area under the receiver-operating curve; CI = confidence interval; LR- = negative likelihood; LR+ = positive likelihood; NPV = negative predictive value; PPV = positive predictive values.

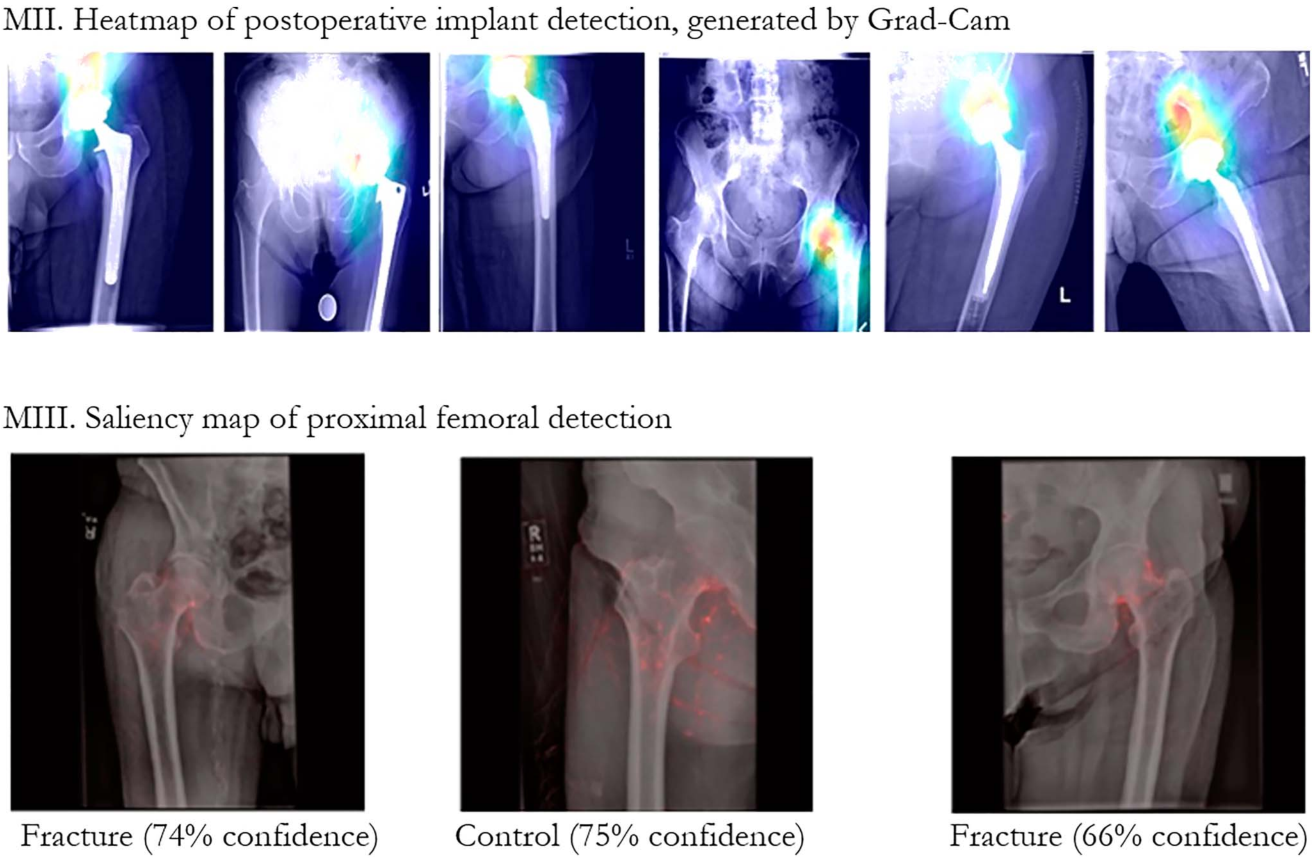
MII. Heatmap of postoperative implant detection, generated by Grad-Cam
L
L
L
MIII. Saliency map of proximal femoral detection
R
Fracture (74% confidence)
Control (75% confidence)
Fracture (66% confidence)

Figure 3. Gradient-based methods for visual explanations. Oversaturated heatmaps (MII) can occur because of activation magnification, especially when the activations are highly localized and concentrated in the case of an implant.

without considering any additional notes (eg, mechanism of injury) or patients' demographics (eg, sex and age), and using this information may improve the accuracy and generalizability of the model. The mechanism of injury may differ based on age, sex, and race, and having a model incorporating these nonlinear associations (eg, age) might be beneficial in the future. A hip fracture after low-energy trauma (ie, fall from standing height) usually occurs in frail older patients, where younger patients usually sustain a high-energy trauma (eg, motor vehicle accident or fall from height). Incorporating multimodal data integration may eventually contribute to improvement of risk-stratification tools, supporting the decision-making process.[18]

Traditionally, computer vision systems were explicitly programmed. Currently, DL methods are used to automatically

TABLE 3

Model Performance of the M3 Teacher Model

	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	LR+	LR-	AUROC	AUPRC
Voting ensemble	89.3%	90.6%	85.7%	94.5%	77.1%	92.5%	6.35	0.11	50.0%	86.5%
Deep learning ensemble	91.9%	96.5%	79.4%	92.7%	89.3%	94.6%	4.68	0.04	96.2%	98.4%
Res50 voting ensemble	88.9%	88.3%	90.5%	96.2%	74.0%	92.1%	9.27	0.13	—	—
Res50	84.6%	81.9%	92.1%	96.6%	65.2%	88.6%	10.32	0.20	96.2%	98.5%
Res50	88.5%	88.9%	87.3%	95.0%	74.3%	91.8%	7.00	0.13	96.0%	98.3%
Res50	89.7%	91.2%	85.7%	94.5%	78.3%	92.9%	6.39	0.10	96.2%	98.4%
Res50	87.6%	86.5%	90.5%	96.1%	71.2%	91.1%	9.09	0.15	96.4%	98.5%
Res50	88.5%	88.9%	87.3%	95.0%	74.3%	91.8%	7.00	0.13	96.4%	98.5%
Res50	89.3%	88.9%	90.5%	96.2%	75.0%	92.4%	9.33	0.12	96.6%	98.6%
Dense201	87.2%	89.5%	81.0%	92.7%	73.9%	91.1%	4.70	0.13	92.4%	96.1%
VGG19	88.5%	90.1%	84.1%	93.9%	75.7%	91.9%	5.67	0.12	95.6%	98.1%
EfficientB5	91.9%	96.5%	79.4%	92.7%	89.3%	94.6%	4.68	0.04	95.4%	98.1%
EfficientL2	90.2%	95.9%	74.6%	91.1%	87.0%	93.4%	3.78	0.06	95.7%	98.2%
EfficientB7	91.9%	97.1%	77.8%	92.2%	90.7%	94.6%	4.37	0.04	94.3%	97.2%
EfficientB7	90.2%	94.7%	77.8%	92.0%	84.5%	93.4%	4.26	0.07	93.8%	96.7%

LR = likelihood ratio; NPV = negative predictive value; PPV = positive predictive value; Res50 voting ensemble is a voting classifier based on only Res50 models.

extract features from images. Promising computer vision models using DL are increasingly incorporated into clinical practice. This includes predicting the risk of breast cancer on mammographs[19] and autolabeling of chest X-rays[20] to interpret pathology on whole-slide histopathology images.[21] To the best of our knowledge, this is the first study conducting a DL approach to create an image-based medical registry. The results of this study show that the developed approach may outperform current data collection strategies for developing a hip fracture registry. Data quality audits from current hip fracture registries worldwide reveal a high rate of errors in data collection.[22] In the literature, the data accuracy of hip fracture registries (actual cases that are found in the registry) ranged from 58% to 90%.[23–25] In this study, the cascade model was built to accurately capture hip fracture cases for registry purposes. Future efforts can focus on adding location-specific classifications to the image-based registry for further evaluation of quality reporting and patient outcomes. The current student-teacher technique framework used for proximal femoral fracture detection (MIII) showed to improve the model performance. The teacher network is first trained on the task, and the student will then learn from the teacher and step-by-step feedback.

In summary, clinicians and researchers can use the developed DL model approach for quality improvement, diagnostic and prognostic research purposes, and building clinical decision support tools.

REFERENCES

1. Arshi A, Lai WC, Chen JB, et al. Predictors and sequelae of postoperative delirium in geriatric hip fracture patients. *Geriatr Orthop Surg Rehabil*. 2018;9:2151459318814823.
2. Delaunay C. Registries in orthopaedics. *Orthop Traumatol Surg Res*. 2015;101(1 suppl):S69–S75.
3. Quan H, Li B, Saunders LD, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Serv Res*. 2008;43:1424–1441.
4. Karhade AV, Bongers MER, Groot OQ, et al. Development of machine learning and natural language processing algorithms for preoperative prediction and automated identification of intraoperative vascular injury in anterior lumbar spine surgery. *Spine J*. 2021;21:1635–1642.
5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
6. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
7. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal*. 2017;35:303–312.
8. Langerhuizen D, Janssen S, Mallee W, et al. What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. *Clin Orthop Relat Res*. 2019;Nov:2482–2491.
9. 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*. 2019;32:672–677.
10. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A*. 2018;115: 11591–11596.
11. Murphy EA, Ehrhardt B, Gregson CL, et al. Machine learning outperforms clinical experts in classification of hip fractures. *Sci Rep*. 2022;12:2058.
12. Cheng C-T, Wang Y, Chen H-W, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. *Nat Commun*. 2021;12:1066.
13. Campos GFC, Mastelini SM, Aguiar GJ, et al. Machine learning hyperparameter selection for Contrast limited adaptive Histogram equalization. *EURASIP J Image Video Process*. 2019;2019:59.
14. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016:770–778.
15. Xie Q, Hovy EH, Luong M-T, et al. *Self-training with Noisy Student Improves ImageNet Classification*; 2019. *CoRR*. abs/1911.0. http://arxiv.org/abs/1911.04252
16. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2019;128:336–359.
17. Simonyan K, Vedaldi A, Zisserman A. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*; 2013. doi. 10.48550/ARXIV.1312.6034
18. Boehm KM, Aherne EA, Ellenson L, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer*. 2022;3:723–733.
19. Yala A, Mikhael PG, Strand F, et al. Toward robust mammography-based models for breast cancer risk. *Sci Transl Med*. 2021;13:eaba4373.
20. Kim D, Chung J, Choi J, et al. Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. *Nat Commun*. 2022;13:1867.
21. Diao JA, Wang JK, Chui WF, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat Commun*. 2021;12:1613.
22. Werner M, Macke C, Gogol M, et al. Differences in hip fracture care in Europe: a systematic review of recent annual reports of hip fracture registries. *Eur J Trauma Emerg Surg*. 2022;48:1625–1638.
23. Tan AC, Armstrong E, Close J, et al. Data quality audit of a clinical quality registry: a generic framework and case study of the Australian and New Zealand Hip Fracture Registry. *BMJ Open Qual*. 2019;8:e000490.
24. Voeten SC, Arends AJ, Wouters MWJM, et al. The Dutch Hip Fracture Audit: evaluation of the quality of multidisciplinary hip fracture care in the Netherlands. *Arch Osteoporos*. 2019;14:28.
25. Cundall-Curry DJ, Lawrence JE, Fountain DM, et al. Data errors in the National Hip Fracture Database: a local validation study. *Bone Joint J*. 2016;98-B:1406–1409.