Original Article

Development and validation of a deep-learning model for the detection of non-displaced femoral neck fractures with anteroposterior and lateral hip radiographs

Lian-Xin Wang[1#], Zhong-Hang Zhu[2#], Qi-Chang Chen[2], Wei-Bo Jiang[3], Yao-Zong Wang[4], Nai-Kun Sun[1], Bao-Shan Hu[1*], Gang Rui[1], Lian-Sheng Wang[2*]

[1]Department of Orthopedics, The First Affiliated Hospital of Xiamen University, Xiamen, China; [2]Department of Computer Science, Xiamen University, Xiamen, China; [3]Department of Orthopedics, The Second Affiliated Hospital of Jilin University, Changchun, China; [4]Department of Orthopedics, Zhongshan Hospital of Xiamen University, Xiamen, China

Contributions: (I) Conception and design: LX Wang, BS Hu, LS Wang; (II) Administrative support: BS Hu, G Rui; (III) Provision of study materials or patients: LX Wang, NK Sun, WB Jiang, YZ Wang; (IV) Collection and assembly of data: LX Wang, NK Sun, WB Jiang, YZ Wang; ZH Zhu, QC Chen; (V) Data analysis and interpretation: LX Wang, ZH Zhu, QC Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

[*]These authors contributed equally to this work as co-corresponding authors.

Correspondence to: Bao-Shan Hu, MD. Department of Orthopedics, The First Affiliated Hospital of Xiamen University, 55 Zhenhai Rd., Xiamen 361003, China. Email: xmhbs@21cn.com; Lian-Sheng Wang, PhD. Department of Computer Science, Xiamen University, 422 Siming South Road, Xiamen 361005, China. Email: lswang@xmu.edu.cn.

Background: Hip fractures, including femoral neck fractures, are a significant cause of morbidity and mortality in the elderly population and are typically diagnosed using plain radiography. However, diagnosing non-displaced femoral neck fractures can be challenging due to their subtle appearance on hip radiographs. Previous deep-learning models have shown low accuracy in identifying these fractures on anteroposterior (AP) radiographs; however, no studies have used lateral radiographs. This study aimed to evaluate the potential of using deep-learning with both AP and lateral hip radiographs to automatically identify non-displaced femoral neck fractures.

Methods: We conducted a retrospective analysis of patients with femoral neck fractures at The First Affiliated Hospital of Xiamen University. All the hip radiographs were reviewed, and cases of non-displaced femoral neck fractures were included in the study. Additionally, 439 participants with normal hip radiographs were also included in the study. A vision transformer (Vit) model was developed using 1,536 AP and lateral hip radiograph. The model's performance was compared to the performance of two groups of human observers: an expert group comprising orthopedic surgeons and radiologists, and a non-expert group, including emergency physicians and general practice doctors. We also carried out the external validation using two additional data sets to assess the generalizability of the model.

Results: The Vit model showed exceptional performance in detecting non-displaced femoral neck fractures on paired AP and lateral hip radiographs, achieving a binary accuracy of 95.8% [95% confidence interval (CI): 94.9%, 96.8%] and an area under the curve (AUC) of 0.988. Compared to the human observers, the model had a higher accuracy of 96.7% (95% CI: 93.9%, 99.5%) on the paired AP and lateral hip radiographs, while the accuracy of the expert group was 90.5% (95% CI: 85.7%, 95.2%). Further, the model maintained good performance during the external validation, with an AUC of 0.959 on the paired AP and lateral views.

Conclusions: Our Vit model showed expert-level performance in identifying non-displaced femoral neck fractures on paired AP and lateral hip radiographs. This model has the potential to enhance diagnosis accuracy and improve patient outcomes by reducing the need for additional examinations and preoperative time.

Keywords: Deep learning; femoral neck fracture; vision transformer (Vit)

Submitted Jun 07, 2023. Accepted for publication Oct 24, 2023. Published online Nov 13, 2023.
doi: 10.21037/qims-23-814
View this article at: https://dx.doi.org/10.21037/qims-23-814

Introduction

Hip fractures are a significant cause of morbidity and mortality worldwide (1-4). Older adults are particularly vulnerable to hip fractures with as many as 30% dying within one year of sustaining a hip fracture (5,6). Early diagnosis is crucial for good outcomes, as delayed diagnosis can lead to the displacement of the fracture, malunion, and arthritis, which can lead to a poor prognosis (7). Unfortunately, previous studies have reported misdiagnosis rates ranging from 7% to 14% for all types of hip fractures (8,9). Of all hip fractures, non-displaced femoral neck fracture is a major cause of misdiagnosis on hip radiographs (7,10). Displaced femoral neck fractures and intertrochanteric fractures can be easy to identify on radiographs; however, non-displaced femoral neck fractures can present as subtle changes in bone structure that are difficult to distinguish from normal anatomy (*Figure 1*). As a result, additional tests such as computed tomography (CT) scans, bone scans, and magnetic resonance imaging (MRI) are often required for diagnosis, which can increase surgery time and overall care costs (11).

Deep learning is a type of machine learning that uses artificial neural networks to learn from large data sets and make predictions based on new data. In recent years, deep convolutional neural networks (CNNs) have shown promise in medical image analysis (12-16). Many radiographic studies have used CNNs for hip fracture detection (6,17-24). However, the sensitivity of identifying non-displaced femoral neck fractures using CNNs is around 50% (6,18), and previous studies have not used lateral radiographs. Recently, vision transformer (Vit) models have been developed that can completely replace standard convolutions in deep neural networks by operating on a series of image patches. Further, the latest studies indicate that the prediction errors of Vit models are more consistent with those of humans than CNNs (25-27). If an algorithm can achieve expert-level accuracy, the automated detection of non-displaced femoral neck fractures has the potential to reduce missed diagnoses. Thus, Vit models have the capability to minimize delayed management and enhance patient outcomes.

In this study, we evaluated the diagnostic performance of a Vit model for detecting non-displaced femoral neck fractures using plain anteroposterior (AP) and lateral hip radiographs. We also compared the performance of our model to that of 16 clinical physicians with varying levels of experience in musculoskeletal imaging. The overall objective of this study was to develop a Vit model that can help clinical physicians quickly and accurately diagnose non-displaced femoral neck fractures. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-814/rc).

Methods

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by The First Affiliated Hospital of Xiamen University and the requirement of individual consent for this retrospective analysis was waived.

Data set acquisition

We retrospectively searched the radiology reports of The First Affiliated Hospital of Xiamen University for hip or pelvic radiographs using the words "femoral neck fracture" from June 2009 to May 2022. Hip radiographs were obtained from the following four manufacturers of radiologic data sources: GE Healthcare, Philips Medical Systems, Kodak, and Canon. The following images were excluded: (I) images of poor quality (e.g., images with poor

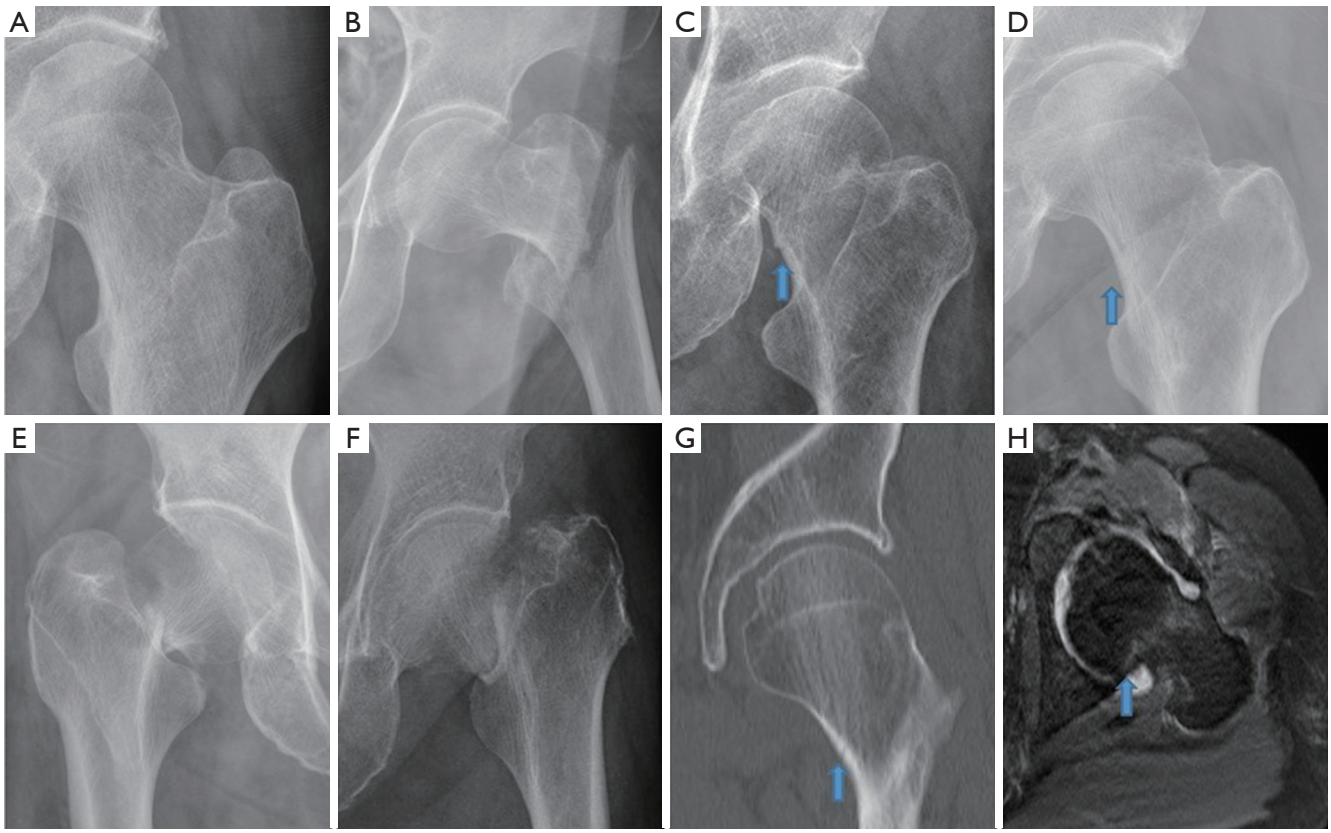

Figure 1 Representative radiographs of different hip fracture cases. (A) A normal hip. (B) A radiograph of a femoral intertrochanteric fracture. (C,D) Non-displaced femoral neck fractures (Garden I/II). (E,F) Displaced femoral neck fractures (Garden III/IV). (G,H) The CT scan and MRI scan of the non-displaced femoral neck fracture in (D), respectively. The blue arrows show the position of the fracture. CT, computed tomography; MRI, magnetic resonance imaging.

detail, contrast, or inappropriate film darkness); (II) images of fractures older than four weeks; (III) images of chronic hip diseases; and (IV) images of patients with hardware (e.g., screws, plates, wires, or pins).

A total of 1,889 patients (aged 18 years or older) were identified as having a femoral neck fracture based on a review of the reports by an orthopedic surgeon with 6 years of experience (Wang LX). Two board-certified orthopedic surgeons (Wang LX and Sun NK, with 6 and 15 years of experience, respectively) excluded 1,264 cases of displaced femoral neck fractures (Garden III/IV) from all the femoral neck fractures on hip radiographs. An additional 296 cases without lateral views were then excluded, resulting in 329 cases of non-displaced femoral neck fractures (Garden I/II). The ground truth for the fracture status was determined by CT scans (148 cases, 45.0%), MRI scans (4 cases, 1.2%), and postoperative radiographs (107 cases, 32.5%). For the remaining 70 cases (21.3%) without CT, MRI, or postoperative images, two board-certified orthopedic surgeons (Wang LX and Sun NK) reached a consensus on the fracture status. Radiographs of normal hips were obtained from patients diagnosed as normal on reports by two board-certified radiologists and reviewed by an orthopedic surgeon (Wang LX) to exclude the presence of a fracture. The data set for this study included 1,536 hip radiographs, consisting of 768 pairs of AP and lateral femoral neck radiographs (439 pairs of radiographs of normal hips, and 329 pairs of radiographs of hip fractures). The workflow for this study, including data inclusion, image pre-processing, training, validation, and testing of the model, is shown in *Figure 2*.

Image pre-processing

All the hip radiographs were extracted as Digital Imaging and Communications in Medicine (DICOM) files from

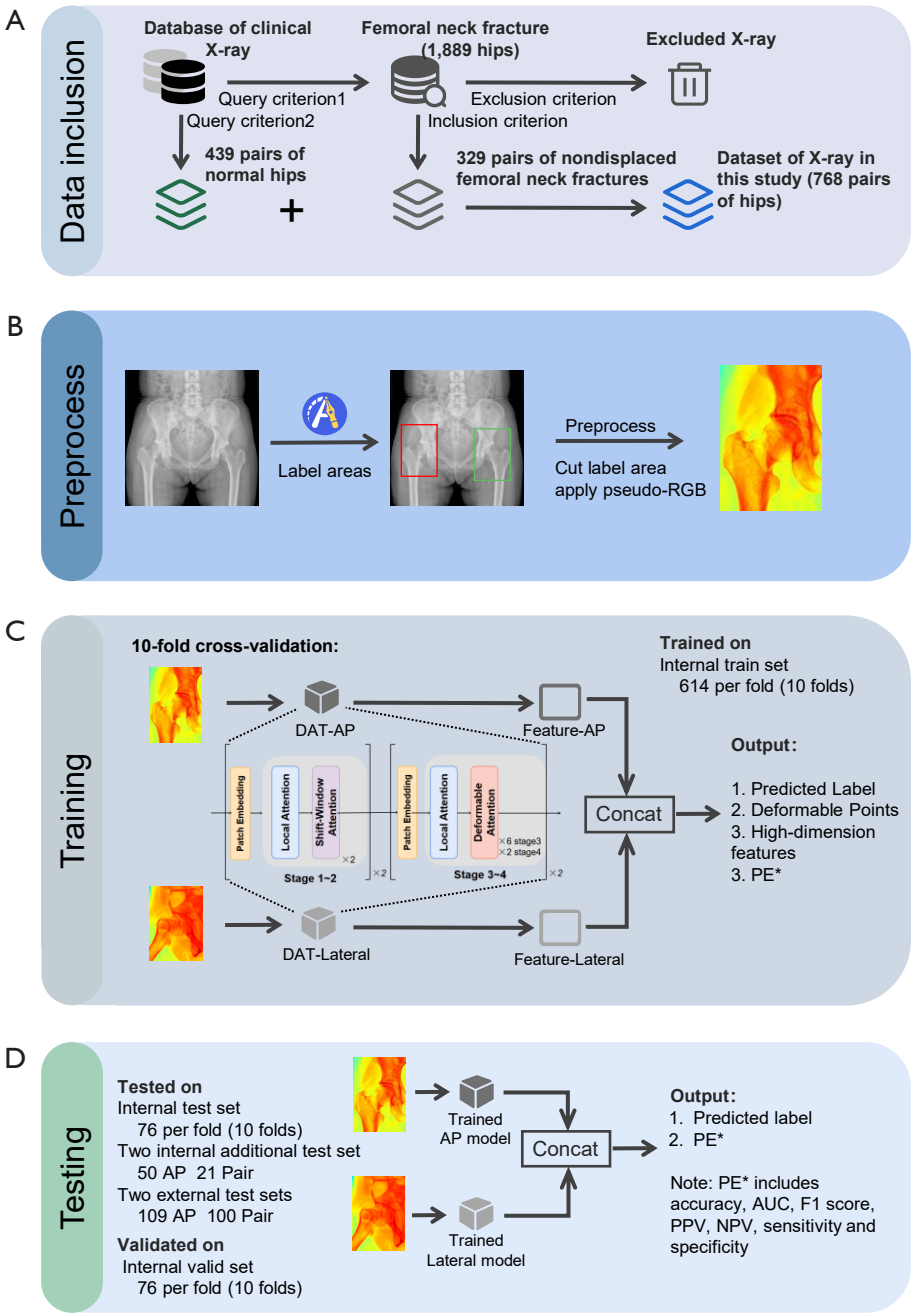

Figure 2 Development and testing for the vision transformer model. (A) A total of 768 pairs of AP and lateral hip radiographs, comprising 329 pairs of non-displaced fractures and 439 normal hips. (B) We cropped the images to a region containing the femoral head and the greater and lesser trochanters. (C) The dual-view DAT model was trained by 10-fold cross-validation. (D) The model was tested on two independent test sets and two external data sets. AUC, area under the curve; PE, Performance evaluation; PPV, positive predicted value; NPV, negative predicted value; Pair, paired AP and lateral views.

the picture archiving and communication system (PACS). An orthopedic surgeon (Wang LX) drew bounding boxes around each femoral neck, encompassing the femoral head and the greater and lesser trochanters in both the AP and lateral hip radiographs. The images were then cropped out of the bounding box area, and truncated normalization, contrast enhancement, and colormap transformation were applied to generate pseudo red, green, and blue (pseudo-RGB) three-channel images for model training. All the images were randomly split into a training set (80%), a validation set (10%), and a test set (10%).

Model architecture

We trained and validated a deep-learning image classification algorithm based on the deformable attention transformer (DAT) framework, which proposes a new deformable self-attention module based on the swim-transformer (28). It allows the self-attention module to focus on relevant areas and capture more information. The DAT framework comprises a feature extraction module with four stages. The first two stages use classic shift-window attention, while the remaining stages use deformable attention. The prediction results were obtained by feeding high-level features extracted from the previous four stages into a fully-connected network. When there were only AP or lateral radiographs, we used a single DAT model to generate the predictions. When there were AP and lateral hip radiographs, we used two individual DAT models to extract features from the AP and lateral hip radiographs separately and then fused these features to produce predictions. During training, the model underwent 100 epochs, with the parameters updated using a stochastic gradient algorithm and a batch size of 32. The network trained a dual-view DAT model and used 10-fold cross-validation to ensure fairness. The final pipeline's accuracy can be adjusted by setting thresholds on the scores returned by the softmax algorithm for each input X-ray image, providing accurate results for different X-ray images. The testing model is available at https://github.com/qc-sw/non-displaced-femoral-neck-fracture.

Model evaluation

The performance of the model was evaluated in relation to the: (I) AP hip radiographs; (II) lateral hip radiographs; (III) paired AP and lateral hip radiographs. The probabilities of fracture were determined for each view, and the final decision was based on the average performance of ten individual test groups. For visualization, we used Gradient-weighted Class Activation Mapping (Grad-CAM) to characterize the model's behavior (29).

In clinical practice, clinicians make the diagnosis of femoral neck fracture based on AP hip radiographs or paired hip radiographs. So we also compared the performance of our model with that of 16 clinicians using two independent testing data sets from June 2022 to October 2022. The first data set comprised AP hip radiographs alone (n=50, of which 25 were normal cases and 25 were fracture cases), while the second data set comprised paired AP and lateral hip radiographs (n=21, of which nine were normal cases and 12 were fracture cases). The expert group included four orthopedic surgeons (with 16–30 years of experience) and four radiologists (with 4–10 years of experience), while the non-expert group included four emergency physicians (with 5–8 years of experience) and four general practice doctors (with 1–8 years of experience). Each physician was shown the images exactly as they were input into the model (model-quality images) and asked to classify each image as "fracture" or "no fracture." To ensure fairness, the comparison was conducted excluding the orthopedic surgeons who determined the ground truth.

In addition, we used two external data sets to test the performance of our model: one from The Second Hospital of Jilin University (n=109, comprising 55 normal hips and 54 fractured hips), and one from The Zhongshan Hospital of Xiamen University (n=100, comprising 50 pairs of normal hips and 50 pairs of fractured hips). The two external data sets were acquired by experienced orthopedic surgeons following the same protocol as that for the internal data sets. The ground truth for the non-displaced femoral neck fractures was determined by CT scan, MRI scans, postoperative radiographs, and the consensus of four experienced orthopedic surgeons (Wang LX, Sun NK, Wang YZ, and Jiang WB).

To evaluate the performance of our model, we calculated several metrics, including sensitivity, specificity, accuracy, F1 score, receiver operating characteristic (ROC) curve, and area under the curve (AUC). We compared the performance of our model with that of the clinicians using accuracy, sensitivity, specificity, positive predicted value (PPV) and negative predicted value (NPV). All the statistical analyses were performed using the extension packages "scikit-learn", "scipy" and "pandas". The pipeline used to build the Vit model was based on an Ubuntu 18.04 operating system with PyTorch 1.12.1+cu113 open-source library with Python

Table 1 Characteristics of the data sets

Parameters	Development data set (n=768)	Independent test data set (n=30)	External validation data set (n=155)
Age (years)	58.0±18.7	73.9±13.3	70.9±14.0
Sex			
Female	449 (58.5)	19 (63.3)	94 (60.6)
Male	319 (41.5)	11 (36.7)	61 (39.4)
Unilateral hip image	1,536	91	309
No fracture	878 (57.2)	42 (46.2)	155 (50.2)
Fracture	658 (42.8)	49 (53.8)	154 (49.8)

The data are presented as the mean ± standard deviation, number of patients/images with percentages in parentheses.

Table 2 Diagnostic performance of the model on different views

Views	Sensitivity (%)	Specificity (%)	Accuracy (%)	F1 Score (%)
AP	91.3 (88.9, 93.8)	95.1 (92.7, 97.6)	93.4 (92.0, 94.8)	92.1 (90.4, 93.8)
Lateral	78.0 (72.1, 83.9)	89.1 (85.7, 92.5)	84.4 (81.6, 87.1)	80.7 (76.7, 84.7)
Pair	94.0 (91.9, 96.2)	97.5 (95.6, 99.3)	95.8 (94.9, 96.8)	95.0 (93.6, 96.3)

Numbers in parentheses are 95% CIs. AP, anteroposterior view; Pair, paired AP and lateral views; CI, confidence interval.

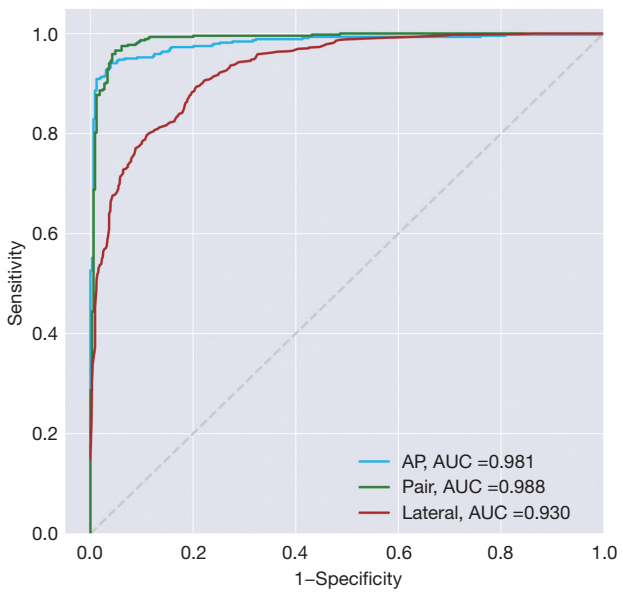

Figure 3 The model's ROC curves for the three different views. The binary classification ROC curves of the model for the AP, lateral and paired views had AUCs of 0.981, 0.930, and 0.988, respectively. ROC, receiver operating characteristic; AP, anteroposterior; AUC, area under the curve; Pair, paired AP and lateral views.

3.9.0 (Python Software Foundation).

Results

Table 1 presents the numbers of patients and images for the development data set, independent test data set, and the external validation data set, along with their respective clinical characteristics.

Performance of the AI model

We evaluated the performance of our AI model using 10-fold cross-validation and found that its performance on the AP and paired views was much better than its performance on the lateral view (*Table 2*). The model's binary accuracy on the AP view was 93.4% [95% confidence interval (CI): 92.0%, 94.8%], with an average F1 score of 0.921. When using paired AP and lateral views, the model achieved a higher accuracy of 95.8% (95% CI: 94.9%, 96.8%), and an average F1 score of 0.950. The binary classification ROC curve of the model on AP and paired views had AUCs of 0.981 and 0.988, respectively (*Figure 3*). These results demonstrate the model's excellent agreement with the ground truth.

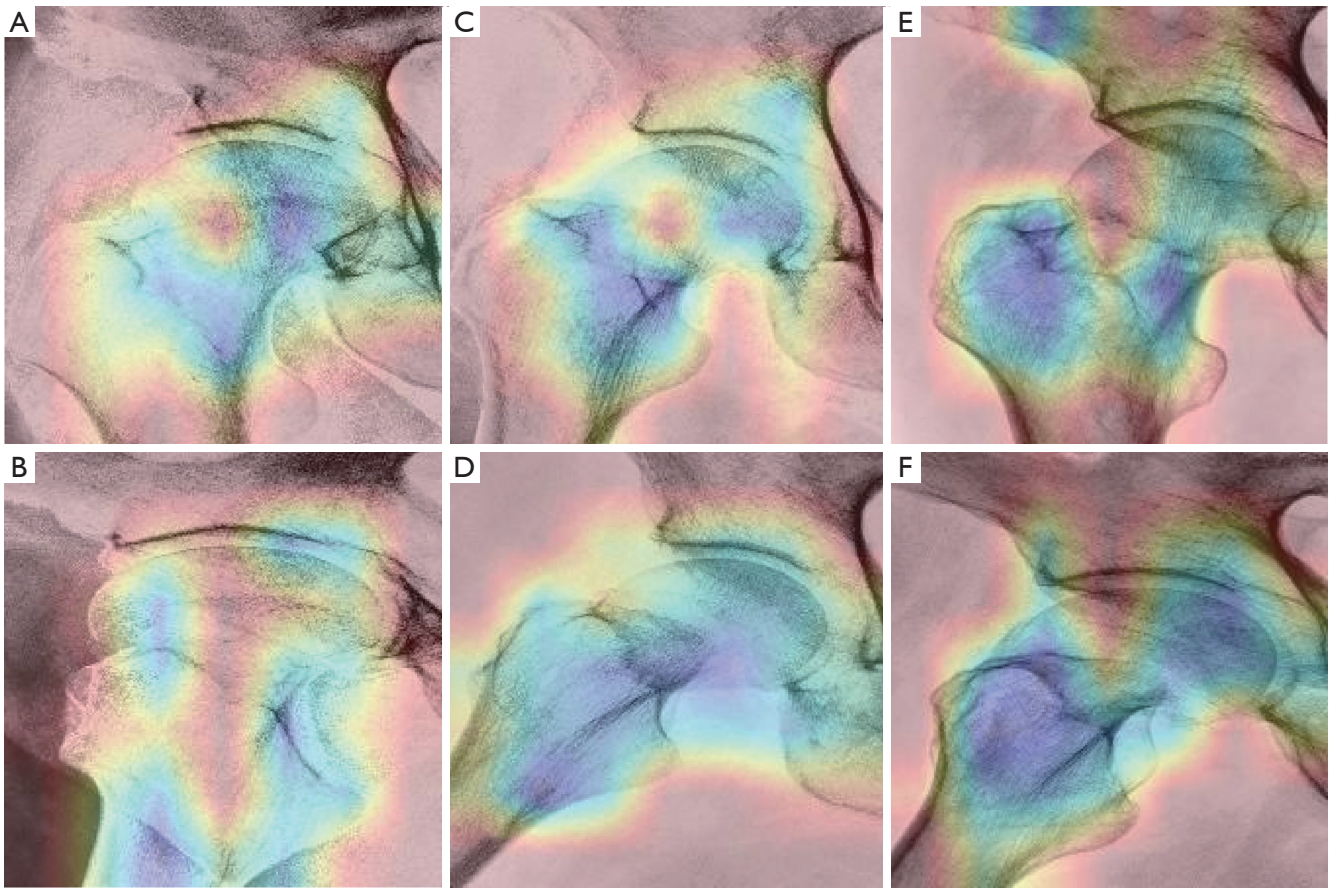

Figure 4 Examples of heatmaps for the model's predictions for three cases of fractures on AP and lateral hip radiographs. AP hip radiograph heatmap (A) and lateral hip radiograph heatmap (B) of a 91-year-old male patient. AP hip radiograph heatmap (C) and lateral hip radiograph heatmap (D) of a 76-year-old male patient. AP hip radiograph heatmap (E) and lateral hip radiograph heatmap (F) of a 71-year-old male patient. AP, anteroposterior.

We also used Grad-CAM maps to further analyze the behavior of our model (*Figure 4*). On the AP view, the model tended to focus on a large area that included the basal part, inner cortex, and outer cortex of the femoral neck, as well as the deviation of the trabecular bone, which is important for diagnosing non-displaced femoral neck fractures. On the lateral hip radiographs, the model tended to focus on the cortex and trabecular bone on both sides of the femoral neck. However, the lucency of the fracture line received little attention, possibly due to the affected and non-displaced morphology of the fracture.

Comparison with clinical physicians on independent data sets

The performance of the model compared with that of all 16 physicians is shown in *Table 3*, and the sensitivity and specificity of each individual physician's performance are plotted on the model's ROC curve in *Figure 5*. On the AP view, the model had a sensitivity of 98.4% (95% CI: 96.9%, 99.9%), a specificity of 72.4% (95% CI: 68.3%, 76.6%), and an accuracy of 85.4% (95% CI: 83.3%, 87.5%). However, the average PPV of the model was 78.1% on the AP view (Appendix 1, Table S1). The sensitivity of the expert group was 95.5% (95% CI: 92.7%, 98.3%), while the sensitivity of the non-expert group was only 73.0% (95% CI: 62.2%, 83.8%). The average NPV was 76.1% for the non-expert group (Appendix 1, Table S1), indicating a high risk of missed diagnosis. The average specificity was 86% for the expert group and 86.5% for the non-expert group. The accuracy of the expert group and non-expert group was 90.8% (95% CI: 84.4%, 97.1%) and 79.8% (95%

Table 3 Diagnostic performance of AI model and 16 physicians on the AP and paired views

AI/physicians	AP view			Paired views		
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
AI	98.4 (96.9, 99.9)	72.4 (68.3, 76.6)	85.4 (83.3, 87.5)	95.0 (90.8, 99.2)	98.9 (96.4, 100.0)	96.7 (93.9, 99.5)
Experts	95.5 (92.7, 98.3) [0.037]	86.0 (74.3, 97.7) [0.012]	90.8 (84.4, 97.1) [0.056]	92.7 (86.9, 98.5) [0.459]	87.5 (74.9, 100.0) [0.033]	90.5 (85.7, 95.2) [0.015]
Non-experts	73.0 (62.2, 83.8) [<0.001]	86.5 (76.2, 96.8) [0.005]	79.8 (76.7, 82.8) [0.002]	80.2 (71.9, 88.5) [0.001]	80.6 (65.9, 95.3) [0.005]	80.4 (75.9, 84.9) [<0.001]

The numbers in the round brackets are the 95% CIs. The numbers in the square brackets are the P values compared to the AI model. AP, anteroposterior; Paired, paired AP and lateral; AI, artificial intelligence; CI, confidence interval.

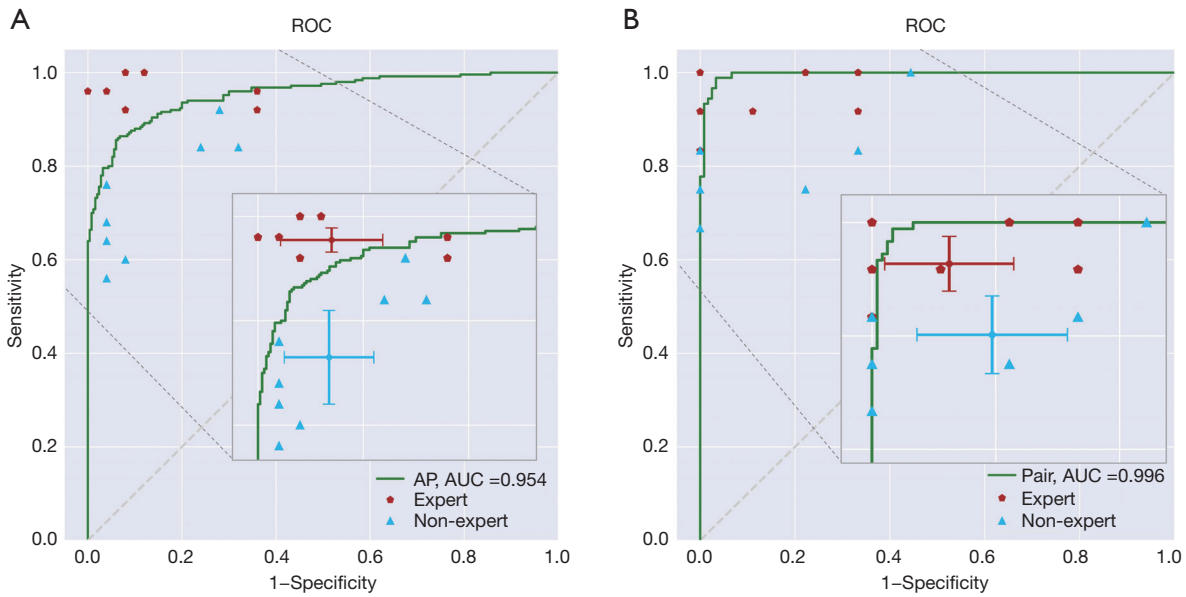

Figure 5 The ROC curve of the model versus that of the experts (orthopedic surgeons and radiologists, in red) and non-experts (emergency physicians and general practice doctors, in blue). (A) The AUC of the model was 0.954 on the AP view. (B) The AUC of the model was 0.996 on the paired views. ROC, receiver operating characteristic; AP, anteroposterior; AUC, area under the curve; Pair, paired AP and lateral views.

CI: 76.7%, 82.8%), respectively. The model achieved an AUC of 0.954 on the AP view. The 95% CIs of both the mean sensitivity and mean specificity of the non-expert group were below the ROC curve of the model, and the mean performance of the expert group was better than that of the model. These results suggest that our AI model outperformed the non-expert physicians but would need to be improved to outperform the expert physicians on the AP radiographs.

On the paired AP and lateral views, the model achieved a sensitivity of 95.0% (95% CI: 90.8%, 99.2%), a specificity of 98.9% (95% CI: 96.4%, 100.0%), and an accuracy of 96.7% (95% CI: 93.9%, 99.5%). The average PPV and NPV of the model were 99.1% and 93.7%, respectively (Appendix 1, Table S2). The sensitivity of the expert group and non-expert group was 92.7% (95% CI: 86.9%, 98.5%) and 80.2% (95% CI: 71.9%, 88.5%), respectively. The specificity of both the expert and non-expert groups were not much improved on the paired views, and was 87.5% (95% CI: 74.9%, 100.0%) and 80.6% (95% CI: 65.9%, 95.3%) for the experts and non-expert groups, respectively. The NPV was 75.3% for the non-expert group on the paired views (Appendix 1, Table S2), which was similar to the NPV on the AP view. The accuracy of the expert group and non-expert group was 90.5% (95% CI: 85.7%, 95.2%) and 80.4% (95% CI: 75.9%, 84.9%), respectively. The 95%

CIs of both the mean sensitivity and mean specificity of the expert group were below the ROC curve of the model. The AUC of the model was 0.996, indicating excellent performance on the paired views. Further, we observed that a few clinicians may have mistakenly classified certain cases of fractures as normal as illustrated in *Figure 6*. However, it is worth noting that all three cases presented in the figure were accurately diagnosed as fractures. These results highlight the potential value of our AI model in improving diagnostic accuracy and reducing the risk of missed diagnosis for femoral neck fractures.

Performance of the AI model in external validation

To further evaluate the generalizability of our AI model, we tested it on two external data sets. *Table 4* shows the performance of the model on these data sets. The model achieved an average binary accuracy above 87%, with an average sensitivity higher than 93%, and a specificity of approximately 80%. The average F1 score of the model was above 0.88, indicating good overall performance. The AUCs were 0.923 and 0.959 on the AP view and paired views, respectively. These results demonstrate that our AI model has good generalizability and can perform well on different data sets, suggesting its potential for use in different clinical institutions.

Discussion

Our study demonstrated the effectiveness of a Vit model in differentiating non-displaced femoral neck fractures from normal hips using both AP and lateral hip radiographs, achieving an accuracy of 95.8% and an AUC of 0.988 on the paired AP and lateral views. To our knowledge, this is the first report of deep learning being used to detect non-displaced hip fractures on both AP and lateral radiographs. The high accuracy and AUC values achieved by our model on different test data sets highlight the potential of this approach to improve diagnostic accuracy.

Non-displaced femoral neck fractures can be difficult to diagnose on hip radiographs, particularly for non-expert physicians, as also evidenced by our results. Several studies have used CNNs to identify non-displaced hip fractures on AP hip radiographs. Krogue *et al.* [2020] presented a CNN model that predicted non-displaced femoral neck fractures with a sensitivity of only 51.2% using 182 cases for training and validation (6). Mutasa *et al.* [2020] also used a CNN model to diagnose non-displaced femoral neck fractures with a sensitivity of 54% using 127 cases for training and validation (18). Their performance was unsatisfactory on single AP hip radiographs, but this was likely due to small data sets and the lack of a lateral view.

Our study demonstrated that the performance of clinicians was not much improved when incorporating lateral views, but the addition of lateral views significantly improved the performance of our model. This highlights the importance of paired AP and lateral radiographs for detecting such fractures, as well as the potential of deep-learning models in improving diagnostic accuracy. The reason for the improved performance of our model on the paired views may be due to the fused features that went through a series of non-linear activation functions, rather than simply combining two predicted probabilities using thresholds. Further, concerns about the routine use of lateral hip radiographs due to discomfort (30-32) can be mitigated by the potential benefits of improved diagnostic accuracy and the reduced need for additional examinations.

Our model has potential applications as a second reader in clinical settings to minimize missed diagnosis and provide an outcome at the expert level for non-displaced femoral neck fractures. By identifying fractures in real time, our model could reduce both harm and costs, potentially improving patient outcomes. Additionally, our model may function as an aid to enhance the performance of all physicians, including the radiologists who give the final report of the radiographs.

Nevertheless, it is important to recognize the limitations of our study. First, the retrospective design introduces potential biases. By selecting cases with both AP and lateral hip radiographs, there is a risk that we overlooked additional non-displaced fractures in our data set. To obtain a more comprehensive understanding, larger prospective studies involving diverse cohorts are necessary to evaluate the performance of our model and verify its generalizability. Second, it is worth noting that fracture diagnosis relies not only on radiological features but also on the patient's history and clinical presentation. Our study focused solely on radiological features, which might have restricted the accuracy of the clinical physicians. Further, our study lacked age-specific comparisons, despite the distinct presentation and management of femoral neck fractures in different age groups. Additionally, variations in the technique used for capturing lateral radiographs by different operators during examinations introduced inconsistencies in our results and limit the reliability of our findings. Moreover, efforts to enhance the decreased specificity observed in our model

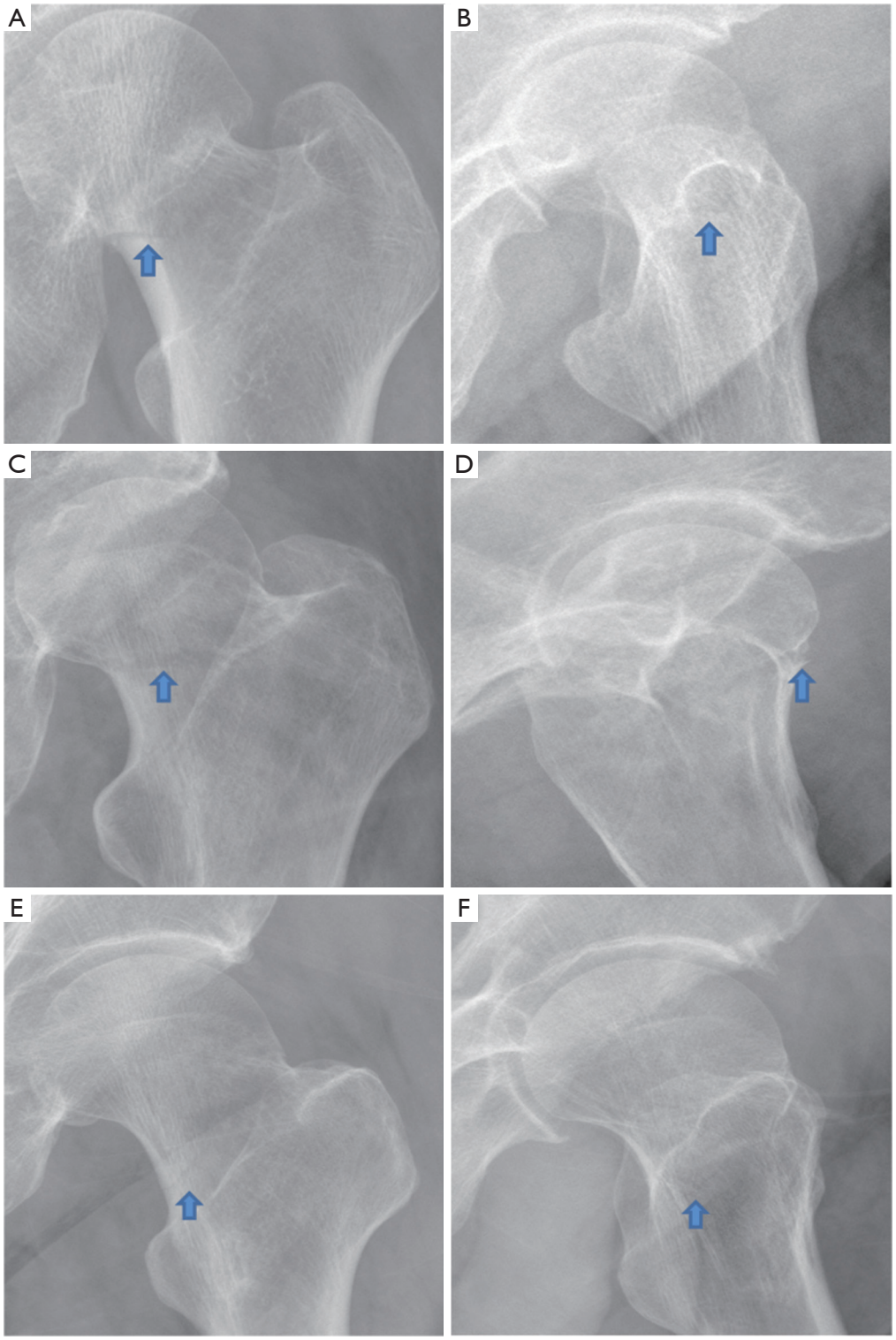

Figure 6 Representative radiographs of three non-displaced femoral neck fracture cases that were diagnosed as normal by some clinicians on paired AP and lateral hip views. The blue arrows show the fractured site. AP (A) and lateral (B) hip radiograph of a 66-year-old female patient. AP (C) and lateral (D) hip radiograph heatmap of an 80-year-old male patient. AP (E) and lateral (F) hip radiograph of a 65-year-old male patient. AP, anteroposterior.

Table 4 Diagnostic performance of AI model on two external data sets

External data sets	Sensitivity (%)	Specificity (%)	Accuracy (%)	F1 Score (%)	AUC (%)
Second Hospital	93.1 (91.4, 94.8)	82.0 (79.7,84.3)	87.6 (86.5, 88.7)	88.4 (87.4,89.3)	92.3 (91.6, 93.0)
Zhongshan Hospital	95.3 (94.1, 96.6)	79.0 (73.1, 84.9)	87.9 (85.6, 90.1)	89.6 (88.0, 91.2)	95.9 (95.4, 96.4)

The numbers in the brackets are the 95% CIs. AUC, area under the curve; Second Hospital, The Second Hospital of Jilin University; Zhongshan Hospital, The Zhongshan Hospital of Xiamen University; CI, confidence interval.

when applied to external data sets could include using larger data sets or refining image acquisition protocols.

Conclusions

In conclusion, our study demonstrates the potential of deep learning to improve diagnostic accuracy for non-displaced femoral neck fractures using paired AP and lateral hip radiographs. Our model has the potential to minimize diagnostic errors, which may have a significant effect on patient recovery and morbidity. Further research is needed to validate the generalizability of our findings and to address the limitations of our study.

Acknowledgments

Funding: This work was supported by funding from the Xiamen Municipal Bureau of Science and Technology (grant No. 3502Z20224ZD1017).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-814/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-814/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by The First Affiliated Hospital of Xiamen University and the requirement for individual consent for this retrospective analysis was waived.

References

1. Papadimitriou N, Tsilidis KK, Orfanos P, Benetou V, Ntzani EE, Soerjomataram I, et al. Burden of hip fracture using disability-adjusted life-years: a pooled analysis of prospective cohorts in the CHANCES consortium. Lancet Public Health 2017;2:e239-46.
2. Kani KK, Porrino JA, Mulcahy H, Chew FS. Fragility fractures of the proximal femur: review and update for radiologists. Skeletal Radiol 2019;48:29-45.
3. Sözen T, Özışık L, Başaran NÇ. An overview and management of osteoporosis. Eur J Rheumatol 2017;4:46-56.
4. Wáng YXJ, Xiao BH. Estimations of bone mineral density defined osteoporosis prevalence and cutpoint T-score for defining osteoporosis among older Chinese population: a framework based on relative fragility fracture risks. Quant Imaging Med Surg 2022;12:4346-60.
5. Roche JJ, Wenn RT, Sahota O, Moran CG. Effect of comorbidities and postoperative complications on mortality after hip fracture in elderly people: prospective observational cohort study. BMJ 2005;331:1374.
6. 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. Radiol

Artif Intell 2020;2:e190023.
7. Parker MJ. Missed hip fractures. Arch Emerg Med 1992;9:23-7.
8. Hakkarinen DK, Banh KV, Hendey GW. Magnetic resonance imaging identifies occult hip fractures missed by 64-slice computed tomography. J Emerg Med 2012;43:303-7.
9. Chellam WB. Missed subtle fractures on the trauma-meeting digital projector. Injury 2016;47:674-6.
10. Haj-Mirzaian A, Eng J, Khorasani R, Raja AS, Levin AS, Smith SE, Johnson PT, Demehri S. Use of Advanced Imaging for Radiographically Occult Hip Fracture in Elderly Patients: A Systematic Review and Meta-Analysis. Radiology 2020;296:521-31.
11. Cannon J, Silvestri S, Munro M. Imaging choices in occult hip fracture. J Emerg Med 2009;37:144-52.
12. Zhou Z, Zhao G, Kijowski R, Liu F. Deep convolutional neural network for segmentation of knee joint anatomy. Magn Reson Med 2018;80:2759-70.
13. Norman B, Pedoia V, Majumdar S. Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. Radiology 2018;288:177-85.
14. Liu F, Zhou Z, Samsonov A, Blankenbaker D, Larison W, Kanarek A, Lian K, Kambhampati S, Kijowski R. Deep Learning Approach for Evaluating Knee MR Images: Achieving High Diagnostic Performance for Cartilage Lesion Detection. Radiology 2018;289:160-9.
15. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. Sci Rep 2018;8:1727.
16. Olveres J, González G, Torres F, Moreno-Tagle JC, Carbajal-Degante E, Valencia-Rodríguez A, Méndez-Sánchez N, Escalante-Ramírez B. What is new in computer vision and artificial intelligence in medical image analysis applications. Quant Imaging Med Surg 2021;11:3830-53.
17. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, Chung IF, Liao CH. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29:5469-77.
18. Mutasa S, Varada S, Goel A, Wong TT, Rasiej MJ. Advanced Deep Learning Techniques Applied to Automated Femoral Neck Fracture Detection and Classification. J Digit Imaging 2020;33:1209-17.
19. Yamada Y, Maki S, Kishida S, Nagai H, Arima J, Yamakawa N, Iijima Y, Shiko Y, Kawasaki Y, Kotani T, Shiga Y, Inage K, Orita S, Eguchi Y, Takahashi H, Yamashita T, Minami S, Ohtori S. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. Acta Orthop 2020;91:699-704.
20. Bae J, Yu S, Oh J, Kim TH, Chung JH, Byun H, Yoon MS, Ahn C, Lee DK. 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 Imaging 2021;34:1099-109.
21. 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 2019;2:31.
22. 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 2019;48:239-44.
23. Adams M, Chen W, Holcdorf D, McCusker MW, Howe PD, Gaillard F. Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. J Med Imaging Radiat Oncol 2019;63:27-32.
24. Oakden-Rayner L, Gale W, Bonham TA, Lungren MP, Carneiro G, Bradley AP, Palmer LJ. 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 2022;4:e351-8.
25. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. In: Vedaldi A, Bischof H, Brox T, Frahm JM. editors. ECCV 2020: Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol 12346. Springer, Cham; 2020:213-29.
26. Chen M, Radford A, Child R, Wu J, Jun H, Luan D, Sutskever I. Generative pretraining from pixels. In: Hal D, III, Aarti S. editors. ICML'20: Proceedings of the 37th International Conference on Machine Learning; 2020:1691-703.
27. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 2020. [Preprint]. Available online: https://doi.org/10.48550/arXiv.2010.11929

28. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 2021. [Preprint]. Available online: https://doi.org/10.48550/arXiv.2103.14030
29. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradientbased localization. 2017 IEEE International Conference on Computer Vision (ICCV); 22-29 October 2017; Venice, Italy. IEEE; 2017:618-26.
30. Almazedi B, Smith CD, Morgan D, Thomas G, Pereira G. Another fractured neck of femur: do we need a lateral X-ray? Br J Radiol 2011;84:413-7.
31. Collin D, Dunker D, Göthlin JH, Geijer M. Observer variation for radiography, computed tomography, and magnetic resonance imaging of occult hip fractures. Acta Radiol 2011;52:871-4.
32. Kumar DS, Gubbi SD, Abdul B, Bisalahalli M. Lateral Radiograph of the Hip in Fracture Neck of Femur: Is it a Ritual? Eur J Trauma Emerg Surg 2008;34:504-7.

Cite this article as: Wang LX, Zhu ZH, Chen QC, Jiang WB, Wang YZ, Sun NK, Hu BS, Rui G, Wang LS. Development and validation of a deep-learning model for the detection of non-displaced femoral neck fractures with anteroposterior and lateral hip radiographs. Quant Imaging Med Surg 2024;14(1):527-539. doi: 10.21037/qims-23-814

Appendix 1

Confusion matrix demonstrates the average true positive, true negative, false positive, and false negative values for the artificial intelligence (AI) model, the expert group, and the non-expert group on the independent test data set. The average positive predicted value (PPV) and negative predicted value (NPV) of the model were 78.1% (24.6/31.5) and 97.8% (18.1/18.5), respectively on the anteroposterior (AP) view. While on the paired views, the average PPV (11.4/11.5) and NPV (8.9/9.5) of the model were 99.1% and 93.7%, respectively. The average NPV for the non-expert group was 76.1% (21.6/28.4) and 75.3% on the AP and paired views, respectively.

Table S1 Confusion matrix for the AI model, the expert group and the non-expert group on the AP view

Predicted	Actual	
	Normal	Fracture
AI		
Normal	18.1	0.4
Fracture	6.9	24.6
Experts		
Normal	21.5	1.1
Fracture	3.5	23.9
Non-experts		
Normal	21.6	6.8
Fracture	3.4	18.2

AP, anteroposterior; AI, artificial intelligence.

Table S2 Confusion matrix for the AI model, the expert group, and the non-expert group on the paired AP and lateral view

Predicted	Actual	
	Normal	Fracture
AI		
Normal	8.9	0.6
Fracture	0.1	11.4
Experts		
Normal	7.9	0.9
Fracture	1.1	11.1
Non-experts		
Normal	7.3	2.4
Fracture	1.7	9.6

AP, anteroposterior; AI, artificial intelligence.

https://dx.doi.org/10.21037/qims-23-814