Interdisciplinary Research in Medical Sciences Specialty
ISSN: 3062-4401
2021, Volume 1, Issue 2, Page No: 49-62

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Deep Learning-Based Detection of Femoral Neck Fractures: Superior Diagnostic Performance and Clinical Decision Support Value

Thomas Becker[1], Andreas Vogel[1], Stefan Krüger[1]*

[1]Department of Systems Medicine, Faculty of Medicine, University of Munich, Munich, Germany.

***E-mail** ✉ stefan.krueger.lab@yahoo.com

Received: 29 May 2021; **Revised:** 28 August 2021; **Accepted:** 02 September 2021

ABSTRACT

The substantial incidence of undetected and incorrectly identified femoral neck fractures (FNF) highlights the need for improved diagnostic support. This work evaluates whether artificial intelligence (AI) can reliably identify FNF and compares its diagnostic capacity with that of physicians. Additionally, it examines how clinicians' performance changes when assisted by AI. A total of 4477 hip radiographs—2884 showing FNF and 1593 appearing normal—were gathered from eight leading tertiary hospitals in China (Union Hospital, Tongji Medical College, Huazhong University of Science and Technology; Wuhan Union Hospital; Wuhan Pu'ai Hospital; Tianyou Hospital, Wuhan University of Science and Technology; Hanyang Hospital, Wuhan University of Science and Technology; Northern Jiangsu People's Hospital; Xiangya Changde Hospital; People's Hospital of Tibet Autonomous Region; Second Affiliated Hospital of Soochow University) to form a multicenter dataset. After annotation, the images were divided into 4029 for training and 448 for testing. A Faster R-CNN framework using three backbone networks (VGG16, VGG16-nottop, and ResNet-50) was built and trained. Performance on the test set—accuracy, sensitivity, specificity, missed-diagnosis rate, misdiagnosis rate, PPV, NPV, and diagnostic time—was benchmarked against five clinicians. The top-performing backbone was subsequently provided to physicians as an aid to reassess the test images and determine the additive value of AI.

Among the models, ResNet-50 yielded the strongest results compared with VGG16 (lowest) and VGG16-nottop (intermediate) across accuracy (0.82 vs 0.58 and 0.76), sensitivity (0.93 vs 0.83 and 0.94), specificity (0.62 vs 0.12 and 0.43), missed-diagnosis rate (0.07 vs 0.17 and 0.06), misdiagnosis rate (0.38 vs 0.88 and 0.57), PPV (0.82 vs 0.63 and 0.75), NPV (0.82 vs 0.28 and 0.81), and diagnostic time (0.02 h vs 0.04 h and 0.03 h). Relative to clinicians, the ResNet-50 model showed higher accuracy, sensitivity, lower missed-diagnosis rate, faster interpretation, and superior NPV, though it lagged in specificity and misdiagnosis rate; PPV differences were minimal. With AI support, clinicians improved across every evaluated metric. AI represents a promising tool for detecting FNF and serves as an effective augmentation for physicians, improving diagnostic reliability.

Keywords: Artificial intelligence, Intelligent medicine, Femoral neck fracture, Detection, Diagnosis

How to Cite This Article: Becker T, Vogel A, Krüger S. Deep Learning-Based Detection of Femoral Neck Fractures: Superior Diagnostic Performance and Clinical Decision Support Value. Interdiscip Res Med Sci Spec. 2021;1(1):49-62. https://doi.org/10.51847/zRdOWwMjFe

Introduction

The femoral neck, an essential part of the hip joint, is highly susceptible to breaking because of its anatomical structure [1, 2]. Owing to population aging, projections suggest that global FNF cases will surpass 63 million by 2050 [3]. This injury is frequently seen in older adults and is associated with severe consequences: within one year, mortality reaches 22% in women and 33% in men after a hip fracture [4, 5]. As a major public health challenge, prior research confirms that earlier and more accurate identification, followed by surgical treatment, effectively decreases all-cause mortality [6–8].

Most FNFs stem from traumatic events, and the emergency department (ED) is typically the patient's first point of contact. Although CT and MRI are routinely advised, the frontal hip X-ray remains the most widely used and economically feasible tool for initial evaluation [9, 10]. However, high ED caseloads and limited numbers of experienced physicians contribute to frequent false-negative and false-positive readings, especially in subtle,

occult, or non-displaced fractures. Reports indicate that combined missed and incorrect diagnosis rates may reach 40% [11, 12]. Such diagnostic errors delay treatment and elevate mortality, making it essential to find supportive tools that increase accuracy during ED assessment.

Advances in intelligent medicine provide meaningful avenues for overcoming these issues. This emerging field merges clinical science with modern technologies such as medical AI, mixed-reality systems, navigational assistance, 3D printing, robotic surgery, wearable devices, cloud-based platforms, telemedicine, 5G-enabled care, and blockchain applications. These innovations aim to address gaps in conventional practice [13, 14]. With accelerated progress in big-data analytics, natural-language processing, and image/speech recognition, medical AI has shown remarkable promise in diagnosis, risk estimation, and chronic-disease monitoring [15–17].

AI integrates concepts from mathematics, computer science, and statistics to build algorithms capable of approximating or surpassing human reasoning. In practice, algorithms are coded (commonly in Python), trained on labeled datasets, and optimized to extract and generalize useful patterns, allowing them to analyze new clinical images autonomously. Within medicine, computer-vision-based deep learning is widely applied, enabling automated detection, classification, and prediction through multi-layer neural networks. These AI-powered imaging tools—ranging from enhanced image reconstruction to automated disease recognition—have contributed significantly to diagnostics in pulmonary medicine, ophthalmology, dermatology, and neurology [18–21].

In earlier work, our group had already achieved AI-based automated identification of tibial plateau fractures, femoral intertrochanteric fractures, distal radius fractures, lumbar spondylolisthesis, osteosarcoma, and uterine fibroids. The present research represents our **seventh** successful application of AI-assisted diagnostic technology. In summary, we first compiled a dataset of hip radiographs with and without femoral neck fractures, carried out preprocessing to clean the data, and manually marked the target regions through clinician annotation. The majority of these images were then used for training the AI model, while a smaller portion was reserved for evaluation. After model training, we assessed the system's accuracy in identifying FNF on hip X-rays and compared its diagnostic competence with that of five emergency-department orthopedic specialists. Subsequently, the AI-generated output was used as reference material to examine whether AI could enhance physicians' performance in detecting FNF. The results offer new insights for improving FNF diagnostic precision.

Materials and Methods

Database preparation

This study constructed a broad, multicenter hip X-ray dataset (including FNF and non-fracture images) sourced from eight leading tertiary medical centers in China: Union Hospital, Tongji Medical College, Huazhong University of Science and Technology (Wuhan Union Hospital), Wuhan Pu'ai Hospital, Tianyou Hospital of Wuhan University of Science and Technology, Hanyang Hospital of Wuhan University of Science and Technology, Northern Jiangsu People's Hospital, Xiangya Changde Hospital, the People's Hospital of the Tibet Autonomous Region, and the Second Affiliated Hospital of Soochow University. The inclusion and exclusion rules are summarized in **Table 1**. Altogether, 4477 radiographs from 4477 individuals were obtained, consisting of 2884 FNF images and 1593 normal hip images. To protect patient confidentiality, all identifying information on the images was removed. The original DICOM files were converted into JPG format at 1024 × 1024 resolution using MicroDicom (https://www.microdicom.com). These 4477 JPG images were then stored for subsequent diagnostic review and annotation.

Table 1. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
(1) Adult patients aged >18 years	(1) Patients aged <18 years (to avoid confusion with low-density epiphyseal lines resembling fracture lines)
(2) Isolated femoral neck fracture (FNF) only; no concomitant fractures (e.g., intertrochanteric, acetabular, or femoral shaft fractures)	(2) Presence of other hip or pelvic fractures in addition to or instead of FNF
(3) Standard, high-quality hip X-ray images taken in correct positioning, without overexposure, motion blur, or obscuring artifacts (e.g., plaster casts, splints, metal fixators, or metal implants/accessories)	(3) Non-standard, poor-quality, or externally performed X-rays; images with incorrect positioning, overexposure, motion blur, or obscuring objects

(4) For positive FNF cases: only preoperative X-ray films were included	(4) Postoperative X-ray films

Label

Three orthopedic specialists, each with over 10 years of clinical experience, independently classified the 4477 radiographs as FNF or normal. Image annotation was then conducted using the Labelme software (https://github.com/wkentaro/labelme). The procedure involved importing the images via the Open Dir function, right-clicking to select Create Rectangle, and entering annotation mode. For FNF cases, the visible fracture line of the femoral neck was enclosed within a rectangular bounding box and saved under the label "fracture." The annotated region was required to be as small and precise as possible to avoid including irrelevant structures, which could negatively impact the training process **(Figure 1a)**. For linear fracture patterns, the rectangle was drawn diagonally for better alignment with the fracture line **(Figure 1b)**. When multiple fracture segments appeared, each region was labeled separately **(Figure 1c)**. For normal images, no special marking was required. If any radiographs were ambiguous or disputed, patient CT scans were used as the reference standard; when CT was unavailable, the final label was assigned based on consensus among the three experts. Each annotated image and its corresponding .json file were incorporated into the final dataset for AI training and testing.

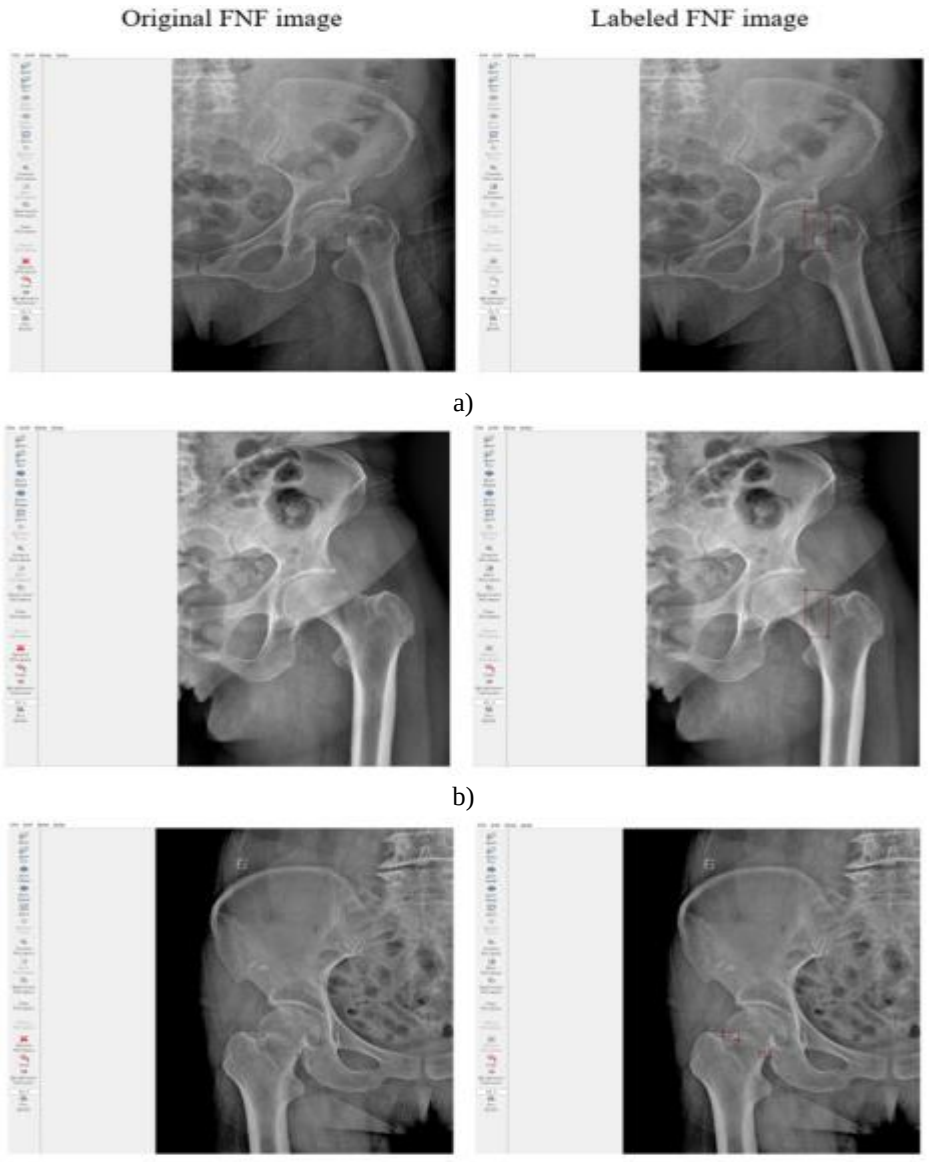


Figure 1. Demonstration of labeling FNF regions. Red rectangles represent annotated fracture zones.

Algorithm preparation

Following dataset construction, this study employed the Faster Region Convolutional Neural Network (Faster R-CNN) to detect FNF on hip X-rays. Faster R-CNN represents a progression from the earlier R-CNN and Fast R-CNN models. The original R-CNN—one of the first deep-learning-based object detectors—operates by extracting region proposals, passing each candidate through a CNN for feature computation, classifying them using a Support Vector Machine (SVM), and refining coordinates through bounding-box regression. However, the repeated feature extraction for each proposal results in substantial, redundant computation and slows the process significantly. Fast R-CNN introduced ROI pooling, which overcomes the inefficiency of processing each proposal separately, thereby markedly speeding up performance. Building upon this, Faster R-CNN integrates a Region Proposal Network (RPN), which further streamlines region generation and enhances overall efficiency.

which enhances detection precision and computational speed, achieving a true "end-to-end" workflow. The schematic layout of Faster R-CNN is presented in **Figure 2**. In summary, Faster R-CNN is composed of five functional components, through which the input image is sequentially processed:

(1) Pre-processing module: receives the raw image, adjusts its scale for consistent parameter handling, and forwards it to the backbone;

(2) Backbone module: serves as the central feature extractor, producing a feature map that is passed to the RPN;

(3) RPN module: applies convolution and softmax layers to refine the feature information and create region proposals, which are then delivered to the ROI pooling stage;

(4) ROI pooling module: merges proposals with feature maps to form proposal–feature maps and sends them to the fully connected layers;

(5) Fully connected module: determines the final classification of each proposal–feature map and outputs bbox_pred and cls_prob, producing the final detection box and diagnostic result. Once these steps are completed, the system outputs the annotated image—showing the suspected FNF region if present—which can also be used as clinical reference material.

To further validate performance and adaptability, three backbone configurations were evaluated: VGG16, VGG16-nottop, and ResNet-50. VGG16, introduced by the Oxford Visual Geometry Group in 2014, includes 13 convolutional and 3 fully connected layers, using uniform 3 × 3 kernels and 2 × 2 max-pooling, and performs effectively in localization and classification tasks [22]. VGG16-nottop removes the final three fully connected layers, making it potentially better suited for single-channel grayscale inputs such as hip X-rays. ResNet-50 is a more recent and advanced architecture with a substantially deeper network (50 layers compared with VGG16's 16). These three variants of Faster R-CNN were prepared for model training.

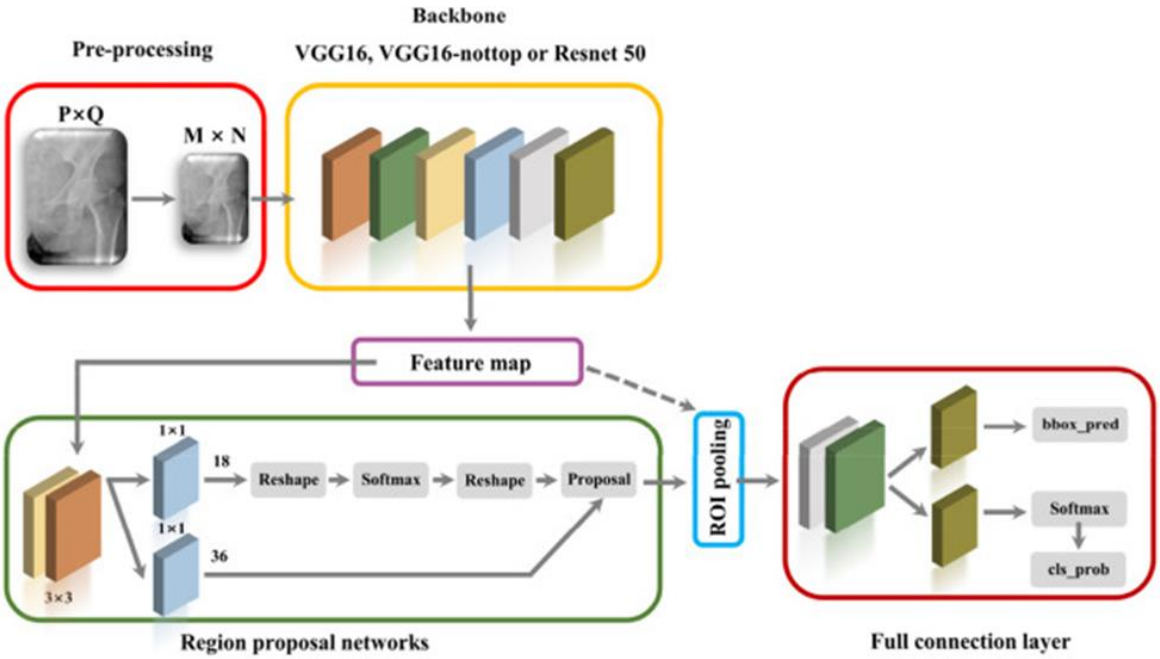


Figure 2. Structure of Faster R-CNN.

Study design

Using the constructed Faster R-CNN models, the 4477 hip X-rays were randomly split into a training set of 4029 images (2595 FNF and 1434 normal) and a testing set of 448 images (289 FNF and 159 normal), maintaining a 9:1 ratio. The allocation of samples is presented in **Table 2**. The training set was provided to the algorithm so it could learn fracture patterns and normal anatomical features. After training, the testing set was used to assess diagnostic ability; the model automatically generated detection boxes and labels indicating suspected FNF areas. The computational setup included Python 3.8, PyTorch 1.6.0, Windows 10, and an NVIDIA GeForce RTX 3080 GPU.

Following model evaluation, additional metrics—F1 score, recall, precision, average precision (AP), precision–recall (P–R) curves, receiver operating characteristic (ROC) curves, and area under the ROC curve (AUROC)—were produced to reflect learning effectiveness. Based on the annotated outputs, the diagnostic qualities of the three backbone models (accuracy, sensitivity, specificity, missed-diagnosis rate, misdiagnosis rate, PPV, NPV, and reading time) were calculated and compared with physician performance from a panel of orthopedic attendings in the ED of Wuhan Union Hospital.

Table 2. Distribution of training and testing datasets.

Dataset	Femoral Neck Fracture (FNF) Images	Normal Hip Images	Total Images
Complete Database	2884	1593	4477
Training Dataset	2595	1434	4029
Testing Dataset	289	159	448

Performance assessment of human level and AI-aided human level

To measure clinician performance in identifying FNF on radiographs, five orthopedic attending physicians from the emergency department of Wuhan Union Hospital were recruited. All participants had at least 3 years of experience interpreting orthopedic X-rays and were not otherwise involved in this study. They independently reviewed the 448 images from the testing dataset and categorized each as FNF or normal, under relaxed conditions without external pressure. Their decisions were compared against the original expert labels, and accuracy, sensitivity, specificity, missed-diagnosis rate, misdiagnosis rate, PPV, NPV, and reading time were calculated and then compared with AI performance.

To evaluate the benefit of AI support, the same five physicians repeated the diagnosis of the 448 X-rays, this time using results from the top-performing Faster R-CNN backbone as reference (including the predicted detection boxes and classifications). They were explicitly informed of the algorithm's limitations to prevent overreliance, and final judgments were to be based on a combination of personal expertise and AI suggestions. The diagnostic metrics of this AI-assisted group (accuracy, sensitivity, specificity, missed-diagnosis rate, misdiagnosis rate, PPV, NPV, and reading time) were computed and compared with unaided human performance.

A schematic overview of the full study process is presented in **Figure 3**.

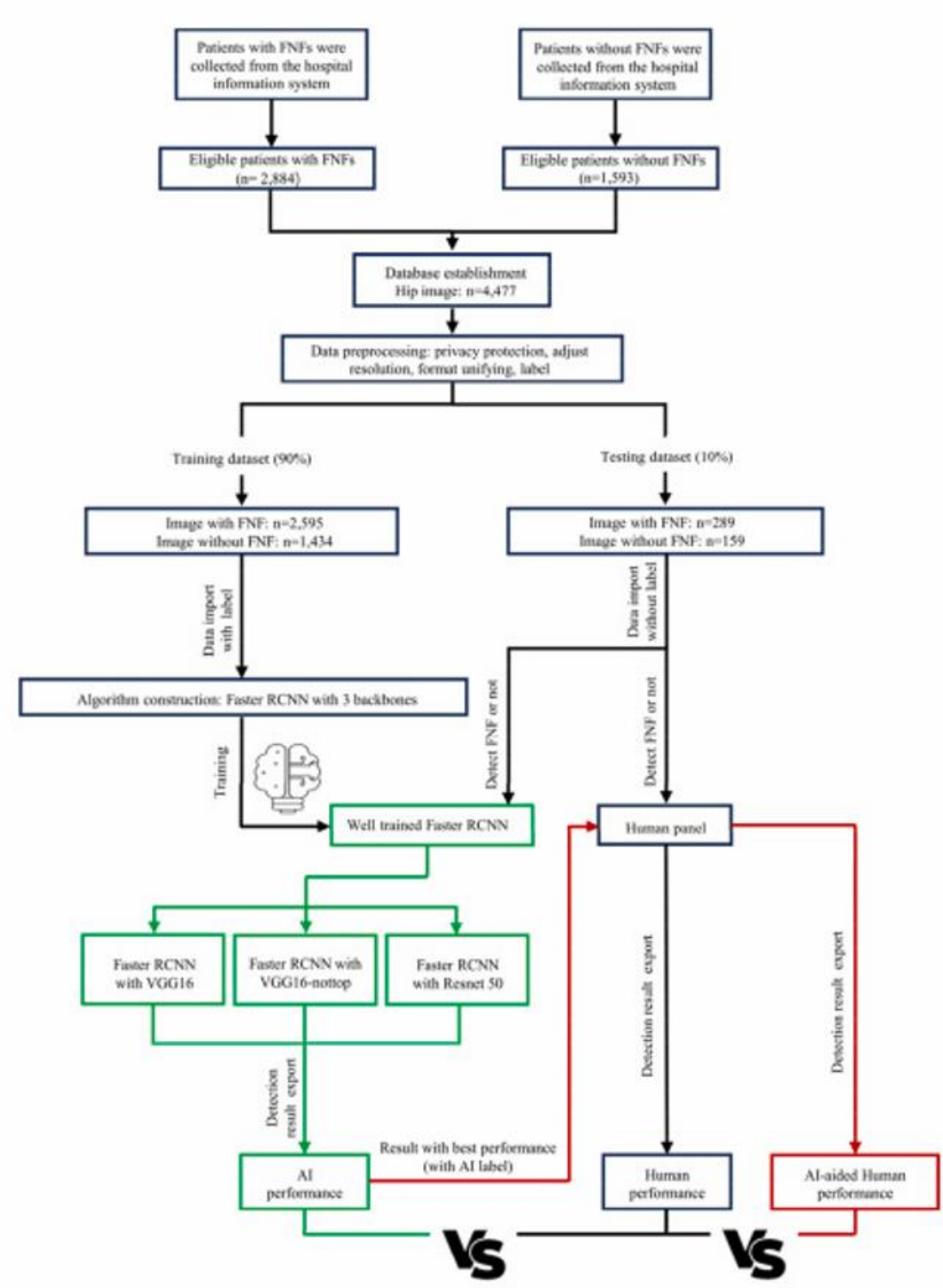


Figure 3. Workflow diagram of the study.

Statistical analysis

All study outcomes were presented as mean ± standard deviation (SD). Statistical processing was performed using GraphPad Prism 8.0.2 (GraphPad Corp., USA). Differences between groups were assessed with Student's t-test. A p value < 0.05 was regarded as statistically meaningful.

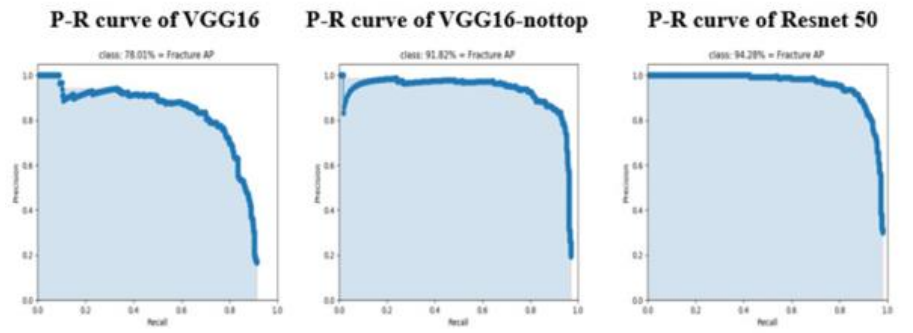
Results and Discussion

Performance of the Faster RCNN model

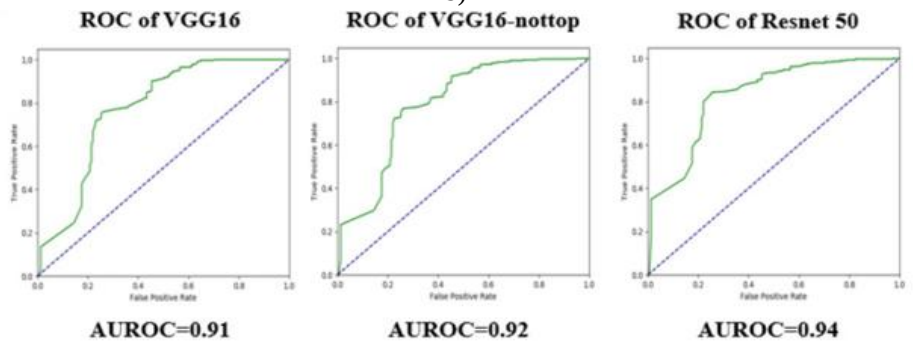
The loaded test images were automatically evaluated to identify femoral neck fractures (FNF). Images recognized as fractured were output with a red bounding box labeled "Fracture" **(Figure 4a)**. For images interpreted as normal

hips, no mark appeared. Using an IOU threshold of 0.5, the F1-scores (precision–recall harmonic mean) reached 0.72 for VGG16, 0.84 for VGG16-nottop, and 0.87 for ResNet-50. The AP values were 0.78, 0.92, and 0.94 for the same sequence of backbones. The P–R curves, ROC curves, and AUROC values of all three versions are presented in **Figures 4b and 4c**. These indicators confirmed that the model trained appropriately on each backbone, suggesting reliable diagnostic outputs. Accuracy, sensitivity, specificity, missed-diagnosis rate, misdiagnosis rate, PPV, NPV, and computation time for each backbone are listed in **Table 3**. Overall, VGG16 showed the lowest performance, VGG16-nottop demonstrated intermediate capability, and ResNet-50 outperformed both. ResNet-50 achieved superior accuracy (0.82 vs 0.58 and 0.76), sensitivity (0.93 vs 0.83 and 0.94), specificity (0.62 vs 0.12 and 0.43), missed-diagnosis rate (0.07 vs 0.17 and 0.06), misdiagnosis rate (0.38 vs 0.88 and 0.57), PPV (0.82 vs 0.63 and 0.75), NPV (0.82 vs 0.28 and 0.81), and required the least time (0.02 h vs 0.04 h and 0.03 h). Therefore, the ResNet-50 model outputs were used as the reference standard for evaluating the AI-supported assessments of the five clinicians.

a)



b)



c)

Figure 4. a. Examples of outputs from the three models, where the red-framed region annotated "Fracture" marks suspected FNF. b. P–R curves. c. ROC and AUROC curves for each backbone.

Table 3. Performance metrics of Faster RCNN backbones for FNF detection.

Performance Metric	VGG16	VGG16-nottop	ResNet-50
Total correct / incorrect classifications (out of 448)	259 / 189	341 / 107	367 / 81
Confusion Matrix			
True Positive (FNF correctly identified)	240	273	268
False Negative (FNF missed)	49	16	21
False Positive (Normal misdiagnosed as FNF)	140	91	60
True Negative (Normal correctly identified)	19	68	99
Diagnostic Performance			
Accuracy	0.58	0.76	0.82
Sensitivity (Recall)	0.83	0.94	0.93
Specificity	0.12	0.43	0.62
Missed diagnosis rate	0.17	0.06	0.07
Misdiagnosis rate	0.88	0.57	0.38
Positive Predictive Value (PPV)	0.63	0.75	0.82
Negative Predictive Value (NPV)	0.28	0.81	0.82
Time consumption per test set (hours)	0.04	0.03	0.02

F = image classified as FNF; N = image classified as normal.

Performance of clinicians and clinicians assisted by AI

For all 448 hip radiographs in the test set, the confusion matrices for the five attending physicians—both unaided and assisted by the AI model—were obtained. Accuracy, sensitivity, specificity, missed-diagnosis rate, misdiagnosis rate, PPV, NPV, and diagnostic time were determined for each condition, and the summary is provided in **Table 4**.

Table 4. Diagnostic performance of clinicians alone vs. clinicians supported by AI in identifying FNF.

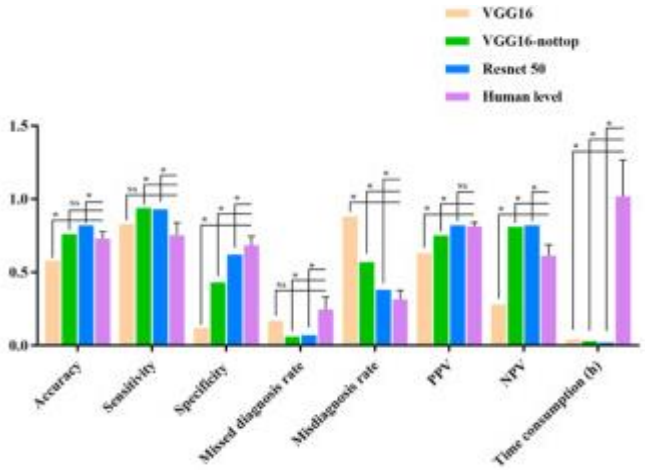
Clinician	Doctor 1 (Unaided)	Doctor 1 (AI-aided)	Doctor 2 (Unaided)	Doctor 2 (AI-aided)	Doctor 3 (Unaided)	Doctor 3 (AI-aided)	Doctor 4 (Unaided)	Doctor 4 (AI-aided)	Doctor 5 (Unaided)	Doctor 5 (AI-aided)
Total correct / incorrect (out of 448)	338 / 110	400 / 48	299 / 149	402 / 46	338 / 110	395 / 53	309 / 139	380 / 68	349 / 99	405 / 43
Confusion Matrix										
True Positive (FNF correctly detected)	221	267	198	258	239	266	188	245	243	278
False Negative (FNF missed)	68	22	91	31	50	23	101	44	46	11
False Positive (Normal → FNF)	42	26	58	15	60	30	38	24	53	32
True Negative (Normal correctly identified)	117	133	101	144	99	129	121	135	106	127
Performance Metrics										
Accuracy	0.75	0.89	0.67	0.90	0.75	0.88	0.69	0.85	0.78	0.90
Sensitivity	0.76	0.92	0.69	0.89	0.83	0.92	0.65	0.85	0.84	0.96
Specificity	0.74	0.84	0.64	0.91	0.62	0.81	0.76	0.85	0.67	0.80
Missed diagnosis rate	0.24	0.08	0.31	0.11	0.17	0.08	0.35	0.15	0.16	0.04
Misdiagnosis rate	0.26	0.16	0.36	0.09	0.38	0.19	0.24	0.15	0.33	0.20
Positive Predictive Value (PPV)	0.84	0.91	0.77	0.95	0.80	0.90	0.83	0.91	0.82	0.90
Negative Predictive Value (NPV)	0.63	0.86	0.53	0.82	0.66	0.85	0.55	0.75	0.70	0.92

Time consumption (hours)	1.24	0.58	1.31	0.72	0.87	0.39	0.94	0.51	0.73	0.37

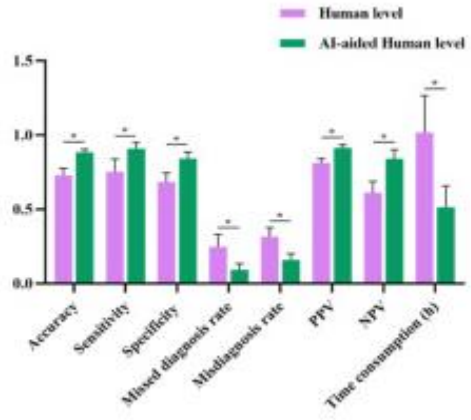
F = diagnosis of FNF; N = diagnosis of normal hip.

Comparative evaluation of algorithm, clinicians, and AI-supported clinicians

Performance metrics of the three Faster RCNN variants were compared against clinicians' baseline performance, and clinicians' results were further compared with their AI-augmented performance **(Table 5)**. ResNet-50 exhibited favorable capability compared with clinicians, showing higher accuracy, sensitivity, lower missed-diagnosis rate, better NPV, and shorter time, while being inferior in specificity and misdiagnosis rate. PPV showed no notable difference. The VGG16-nottop variant showed moderate ability—higher sensitivity, lower missed-diagnosis rate, better NPV, and faster time, but reduced specificity, higher misdiagnosis rate, and a lower PPV; accuracy did not differ significantly. VGG16 demonstrated unsatisfactory performance, improving only time consumption and performing worse in accuracy, specificity, misdiagnosis rate, PPV, and NPV, with no notable difference in sensitivity or missed-diagnosis rate **(Figure 5a)**. When clinicians used outputs from the ResNet-50 model, their diagnostic metrics improved across all categories compared to unaided assessments **(Figure 5b)**.



a)



b)

Figure 5. a) Performance comparison between the three Faster RCNN backbones and clinicians. b) Comparison between clinician-only performance and AI-enhanced performance.

Table 5. Comparison among algorithmic, clinician, and AI-assisted clinician performance.

Performance Metric	VGG16	VGG16 -nottop	ResNet -50	Average Unaided Clinicians	Average AI-Assisted Clinicians
Accuracy	0.58	0.76	0.82	0.73 ± 0.04	0.89 ± 0.02
Sensitivity	0.83	0.94	0.93	0.75 ± 0.08	0.91 ± 0.04
Specificity	0.12	0.43	0.62	0.68 ± 0.05	0.84 ± 0.04
Missed diagnosis rate	0.17	0.06	0.07	0.25 ± 0.08	0.09 ± 0.04

Misdiagnosis rate	0.88	0.57	0.38	0.32 ± 0.05	0.16 ± 0.04
Positive Predictive Value (PPV)	0.63	0.75	0.82	0.81 ± 0.02	0.91 ± 0.02
Negative Predictive Value (NPV)	0.28	0.81	0.82	0.61 ± 0.07	0.84 ± 0.05
Time consumption (hours)	0.04	0.03	0.02	1.01 ± 0.22	0.51 ± 0.13

Femoral neck fractures (FNF) represent roughly 13.6% of all fracture cases. As a trauma-related condition, FNF poses substantial risks to patients' overall health [23, 24]. Timely identification, early intervention, and structured rehabilitation are crucial for restoring joint mobility and reducing postoperative complications [25]. In real-world emergency departments, limited medical staffing and the urgency of patient flow often contribute to unsatisfactory diagnostic performance, including frequent missed or incorrect identifications, especially when the displacement is subtle [26, 27]. Because these errors can delay treatment and lead to downstream complications, especially in acute care settings, developing more accurate and efficient diagnostic strategies for FNF is essential.

In this investigation, we built a Faster RCNN framework incorporating three backbone networks and used it to classify FNF on radiographs, subsequently testing its reliability. The metrics—including F1 score, precision, recall, AP, P–R curves, ROC curves, and AUROC—confirmed that the VGG16, VGG16-nottop, and ResNet-50 variants were able to learn relevant fracture features and achieved a good fit on the training data. Their performance was then evaluated using 488 test hip radiographs, where VGG16 showed the weakest results, VGG16-nottop achieved moderate performance, and ResNet-50 achieved the highest scores **(Table 5)**. ResNet-50's advantage is likely linked to its deeper convolutional layers and architectural improvements such as residual connections and BatchNorm operations. In final testing, Faster RCNN with ResNet-50 exceeded the average human performance in accuracy (0.82 vs 0.73 ± 0.04), sensitivity (0.93 vs 0.75 ± 0.08), missed-diagnosis rate (0.07 vs 0.25 ± 0.08), and NPV (0.82 vs 0.61 ± 0.07), while achieving comparable PPV (0.82 vs 0.81 ± 0.02). VGG16-nottop also surpassed clinicians in sensitivity (0.94 vs 0.75 ± 0.08), missed-diagnosis rate (0.06 vs 0.25 ± 0.08), and NPV (0.81 vs 0.61 ± 0.07), and reached similar accuracy to clinicians. In contrast, VGG16 underperformed in most metrics, suggesting it is not ideal for FNF identification.

Some findings deserve additional attention. First, in terms of time consumption, all three backbone models analyzed radiographs dramatically faster than clinicians, approximately 30–50 times faster. However, regarding specificity and misdiagnosis rate, each of the three networks lagged behind human performance. Although ResNet-50 approached the clinician average, it still did not match it. This reflects a tendency of the algorithms to incorrectly flag normal hips as fractures. Structures such as dense intertrochanteric ridges **(Figure 6a)** or a closing epiphyseal line **(Figure 6b)** may resemble basal neck fractures or compression-type injuries, leading to false positives. Nonetheless, this reduced specificity could be interpreted as an intentionally cautious screening behavior that can be corrected by subsequent CT imaging. False positives are often less harmful than false negatives, as missing an FNF may seriously delay essential care. Importantly, when clinicians incorporated the outputs from ResNet-50, their own specificity and misdiagnosis rates improved considerably. Still, improving the model's handling of high-density lines remains a key area for refinement. Most FNF fracture lines are low-density, so enhancing the discrimination of such patterns may reduce erroneous alerts. Despite the shortcomings in specificity, all three backbones achieved strong sensitivity and low missed-diagnosis rates, with even VGG16 reaching human-level performance, indicating that true FNF cases were rarely overlooked.

The comparisons between ResNet-50 and clinicians, and between clinicians alone versus AI-assisted clinicians, consistently demonstrate that AI systems can serve as capable tools for FNF detection and offer substantial advantages:

(1) They help clinicians reduce both missed and incorrect diagnoses, lowering the risk of patient harm.

(2) AI guidance allows clinicians to focus their attention more efficiently, accelerating interpretation and shortening time to treatment.

(3) During periods of heavy patient volume, AI can reduce clinician workload.

(4) AI can act as a safeguard by flagging cases that clinicians might misinterpret.

(5) It supports the development of diagnostic expertise and confidence.

(6) Unlike humans, AI performance does not degrade due to fatigue or emotional strain, helping protect patients from errors under high workload.

(7) With cloud or edge deployment, AI tools can support telemedicine and benefit regions with limited access to specialists, addressing disparities in healthcare resources.

Although the VGG16-nottop and ResNet-50 variants achieved strong results—and in certain metrics outperformed clinicians—some diagnostic inaccuracies remained, as reflected by reduced specificity and higher misdiagnosis rates. VGG16 performed even worse than the clinician baseline. These flaws highlight that algorithmic predictions should complement, not replace, clinical judgment. Findings from the five clinicians in this study further confirm a pattern reported previously: FNF tends to be underdiagnosed, increasing the need for supportive tools like AI [28, 29]. Given their trained performance, VGG16-nottop and ResNet-50 can function as valuable diagnostic assistants and, in some respects, exceed human accuracy. It is also important to note that the doctors in this study worked under more favorable conditions than real emergency departments, with no time pressure or patient-driven stress, and with heightened carefulness, knowing their results were being evaluated against AI. In actual ED settings, AI-assisted diagnosis may prove even more advantageous than demonstrated here.

Previous investigations have demonstrated that artificial intelligence can support clinicians in diagnosing various conditions, and in certain tasks, AI systems may even outperform medical professionals. Moreover, incorporating AI into diagnostic workflows has been shown to decrease the frequency of clinical errors by enhancing diagnostic precision. For example, Wang *et al*. developed a triplet-branch neural network trained on 1151 X-rays of fatigue fractures and 2842 normal radiographs to identify lower-extremity stress fractures. After model training, the system achieved 0.96 accuracy, 0.95 sensitivity, and 0.80 specificity, surpassing the performance of junior radiologists and matching that of senior radiologists [30]. In another study targeting general pelvic trauma, Cheng *et al*. introduced PelviXNet and trained it using 5024 pelvic X-rays. The network ultimately reached 0.92 accuracy, 0.91 sensitivity, and 0.93 specificity and showed diagnostic capability comparable to orthopedic specialists for pelvic and hip injuries [31].

In the context of femoral intertrochanteric fractures, Liu *et al*. built a Faster RCNN model using 700 hip radiographs. After optimization, it attained 0.88 accuracy, 0.89 sensitivity, and 0.87 specificity—outperforming human readers and delivering results more quickly [11]. Regarding rib fractures, Weikert *et al*. trained a CNN using 511 CT examinations and achieved 0.99 accuracy, 0.87 sensitivity, and 0.91 specificity [32]. Scaphoid fractures are particularly challenging because roughly 20% are not visible on initial X-rays; to address this, Yoon *et al*. trained a deep CNN with 11,838 wrist images and obtained strong performance (0.95 accuracy, 0.87 sensitivity, 0.92 specificity), enabling better detection of occult cases [33]. Seo *et al*. designed a CNN to identify osteoporotic vertebral compression fractures and evaluate vertebral height loss; trained on 387 lateral spine radiographs, the model achieved 0.98 accuracy, 0.93 sensitivity, and 0.99 specificity [34]. Beyond fractures, AI applications have expanded to musculoskeletal conditions such as meniscal tears [35], anterior cruciate ligament injuries [36], scoliosis [37], and osteoarthritis [38]. Collectively, past literature establishes that AI performs effectively across orthopedic diagnostic tasks, and a consolidated summary is presented in **Table 6**.

Compared with the above studies, the main contributions and observations of the present work include:

(1) Construction of a Faster RCNN framework dedicated to detecting FNF on hip radiographs, with its diagnostic reliability confirmed through performance assessment.

(2) Integration of three distinct backbone architectures—VGG16, VGG16-nottop, and ResNet-50—demonstrating that backbone choice substantially influences model outcomes.

(3) Development of a large-scale, multi-center hip X-ray database sourced from **eight** leading Chinese hospitals. Such an extensive dataset enables robust model training and supports algorithmic stability, while the multi-center design enhances generalizability and ensures consistent detection performance across institutions.

The study also contains several limitations:

(1) The model was trained solely on anteroposterior views and therefore is not applicable to lateral radiographs.

(2) The work focused exclusively on identifying FNF and did not address fracture classification. Classification is more complex yet clinically crucial for determining treatment strategies (e.g., conservative vs. operative management and specific surgical choices).

(3) An external validation set was not established. Although the dataset included radiographs from eight hospitals, carving out an independent external subset was not feasible due to the limited total sample size and the need to prioritize effective model training. Future efforts should aim to expand the dataset and incorporate a separate external cohort to better evaluate generalizability and model transportability.

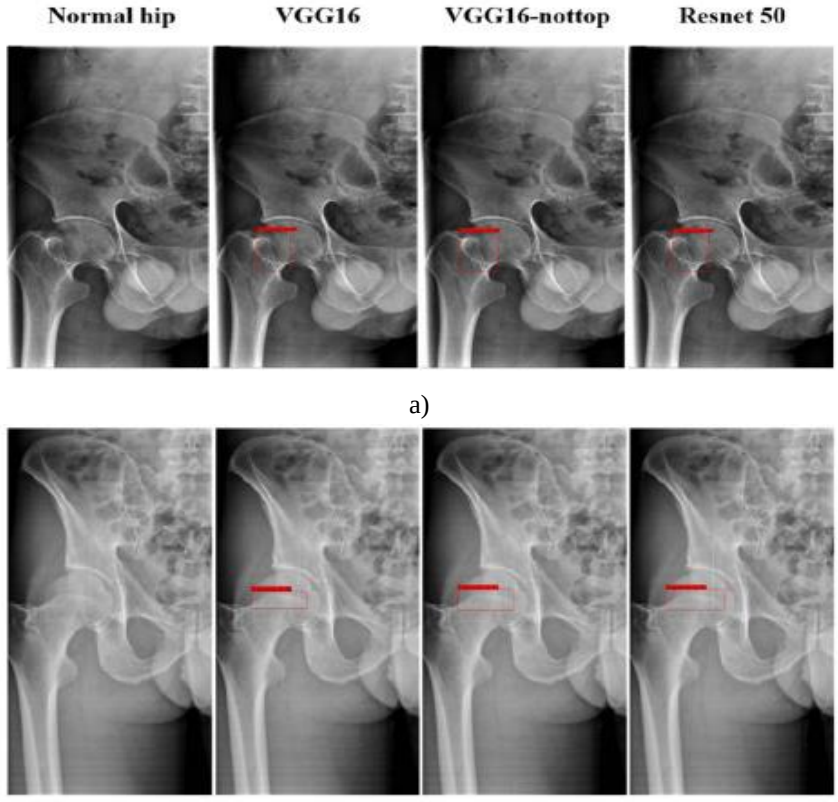


Figure 6. illustrates examples of normal hips incorrectly labeled as fractures by Faster RCNN. In panel A, a dense intertrochanteric crest is mistaken for an FNF, while panel B shows a closed epiphyseal line being misidentified. In both cases, the highlighted red box indicates the algorithm-predicted fracture region.

Table 6. A consolidated overview of AI performance across orthopedic diagnostic applications.

Fracture Type	Database Size (images)	Accuracy	Sensitivity	Specificity	Reference
Lower-extremity fatigue fracture	3993	0.96	0.95	0.80	[30]
Pelvic and hip fracture	5024	0.92	0.91	0.93	[31]
Femoral intertrochanteric fracture	700	0.88	0.89	0.87	[11]
Rib fracture	511	0.99	0.87	0.91	[32]
Scaphoid fracture	11,838	0.95	0.87	0.92	[33]
Spinal osteoporotic compression fracture	387	0.98	0.93	0.99	[34]
Femoral neck fracture (present study)	**4477**				
─ VGG16	4477	0.58	0.83	0.12	–
─ VGG16-nottop	4477	0.76	0.94	0.43	–
─ ResNet-50	4477	0.82	0.93	0.62	–

Conclusion

Artificial intelligence, as an emerging tool in modern clinical practice, is capable of assisting with FNF detection and can serve as a highly effective aid to clinicians, helping strengthen diagnostic accuracy and support clinical decision-making.

Acknowledgments: Thanks for the support of department of Orthopedics, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology.

Conflict of Interest: None

Financial Support: This study was supported by the National Natural Science Foundation of China (No.81974355 and No.82172524) and the Free Innovation Pre-research Fund of Wuhan Union Hospital (No.2024XHYN047).

Ethics Statement: The study was reviewed and approved by the ethics committee of Union Hospital, Tongji Medical College, Huazhong University of Science and Technology (No. 20220840), and Chinese Clinical Trial Registry (No. ChiCTR2300070658, https://www.chictr.org.cn/showproj.html?proj=193847).

References

1. Hackl S, von Ruden C, Weisemann F, Klopfer-Kramer I, Stuby FM, Hogel F. Internal fixation of garden type III femoral neck fractures with sliding hip screw and anti-rotation screw: does increased valgus improve healing? Medicina (Kaunas). 2022;58(11).
2. Florschutz AV, Langford JR, Haidukewych GJ, Koval KJ, Gardner MJ, Ricci WM. Femoral neck fractures: current management. J Orthop Trauma. 2015;29(3):121-9.
3. Cooper C, Campion G, Melton LJ 3rd, O'Neill TW, Kanis JA, Johnston CC. Hip fractures in the elderly: a world-wide projection. Osteoporos Int. 1992;2(6):285-9.
4. Brauer CA, Coca-Perraillon M, Cutler DM, Rosen AB, Wilson DR, Katz JN. Incidence and mortality of hip fractures in the United States. JAMA. 2009;302(14):1573-9.
5. Klop C, Welsing PM, Cooper C, Harvey NC, Kanis JA, Rizzoli R. Mortality in British hip fracture patients, 2000-2010: a population-based retrospective cohort study. Bone. 2014;66:171-7.
6. Fu MC, Boddapati V, Gausden EB, Samuel AM, Russell LA, Lane JM. Surgery for a fracture of the hip within 24 hours of admission is independently associated with reduced short-term post-operative complications. Bone Joint J. 2017;99-B(9):1216-22.
7. Maheshwari K, Planchard J, You J, Bosco JA, Egol KA, Zuckerman JD. Early surgery confers 1-year mortality benefit in hip-fracture patients. J Orthop Trauma. 2018;32(3):105-10.
8. Leer-Salvesen S, Engesaeter LB, Dybvik E, Furnes O, Kristensen TB, Gjertsen JE. Does time from fracture to surgery affect mortality and intraoperative medical complications for hip fracture patients? Bone Joint J. 2019;101-B(9):1129-37.
9. Hakkarinen DK, Banh KV, Hendey GW, Miller JB, Ferguson B, Holmes JF. Magnetic resonance imaging identifies occult hip fractures missed by 64-slice computed tomography. J Emerg Med. 2012;43(2):303-7.
10. Rehman H, Clement RG, Perks F, White TO, Court-Brown CM, Duckworth AD. Imaging of occult hip fractures: CT or MRI? Injury. 2016;47(6):1297-301.
11. Liu P, Lu L, Chen Y, Zhang X, Wang H, Li Z. Artificial intelligence to detect the femoral intertrochanteric fracture: the arrival of the intelligent-medicine era. Front Bioeng Biotechnol. 2022;10:927926.
12. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of artificial intelligence in medicine: an overview. Curr Med Sci. 2021;41(6):1105-15.
13. Liu PR, Zhang JY, Xue MD, Wang YL, Chen H, Zhao Q. Artificial intelligence to diagnose tibial plateau fractures: an intelligent assistant for orthopedic physicians. Curr Med Sci. 2021;41(6):1158-64.
14. Liu P, Lu L, Liu S, Wang H, Zhang X, Chen Y. Mixed reality assists the fight against COVID-19. Intell Med. 2021;1(1):16-8.
15. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-97.
16. Yang Y, Zhao Y, Liu X, Huang J, Wang Z, Chen L. Artificial intelligence for prediction of response to cancer immunotherapy. Semin Cancer Biol. 2022;87:137-47.
17. Contreras I, Vehi J, Oviedo S, Gomez EJ, Hernando ME, Brugués E. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res. 2018;20(5):e10775.
18. Handa T, Tanizawa K, Oguma T, Uozumi R, Watanabe K, Ikezoe K. Novel artificial intelligence-based technology for chest computed tomography analysis of idiopathic pulmonary fibrosis. Ann Am Thorac Soc. 2022;19(3):399-406.

19. Li JO, Liu H, Ting DSJ, Jeon S, Chan RVP, Kim JE. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. Prog Retin Eye Res. 2021;82:100900.
20. Stiff KM, Franklin MJ, Zhou Y, Madabhushi A, Knackstedt TJ, Marchetti MA. Artificial intelligence and melanoma: a comprehensive review of clinical, dermoscopic, and histologic applications. Pigment Cell Melanoma Res. 2022;35(2):203-11.
21. Hillis JM, Bizzo BC, Mehta NR, Gupta A, Patel SK, Singh R. Use of artificial intelligence in clinical neurology. Semin Neurol. 2022;42(1):39-47.
22. Singh M, Bansal S, Ahuja S, Dubey RK, Panigrahi BK, Dey N. Transfer learning-based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data. Med Biol Eng Comput. 2021;59(4):825-39.
23. Dousa P, Cech O, Weissinger M, Dzupa V, Bartonicek J, Skala-Rosenbaum J. Trochanteric femoral fractures. Acta Chir Orthop Traumatol Cech. 2013;80(1):15-26.
24. Fletcher JWA, Sommer C, Eckardt H, Knobe M, Gueorguiev B, Stoffel K. Intracapsular femoral neck fractures-A surgical management algorithm. Medicina (Kaunas). 2021;57(8).
25. Moran CG, Wenn RT, Sikand M, Taylor AM, Costa ML, Achten J. Early mortality after hip fracture: is delay before surgery important? J Bone Joint Surg Am. 2005;87(3):483-9.
26. Dominguez S, Liu P, Roberts C, Mandell M, Richman PB, Shapiro NI. Prevalence of traumatic hip and pelvic fractures in patients with suspected hip fracture and negative initial standard radiographs. Acad Emerg Med. 2005;12(4):366-9.
27. Perron AD, Miller MD, Brady WJ, Keats TE, Daffner RH, De Smet AA. Orthopedic pitfalls in the ED: radiographically occult hip fracture. Am J Emerg Med. 2002;20(3):234-7.
28. Bernstein EM, Kelsey TJ, Cochran GK, Deafenbaugh BK, Kuhn KM, McGrory BJ. Femoral neck stress fractures: an updated review. J Am Acad Orthop Surg. 2022;30(7):302-11.
29. Ring J, Talbot C, Cross C, Hinduja K, Costa ML, Griffin XL. NHSLA litigation in hip fractures: lessons learnt from NHSLA data. Injury. 2017;48(8):1853-7.
30. Wang Y, Li Y, Lin G, Zhang H, Chen X, Zhao W. Lower-extremity fatigue fracture detection and grading based on deep learning models of radiographs. Eur Radiol. 2023;33(1):555-65.
31. Cheng CT, Wang Y, Chen HW, Wu YC, Hsu CY, Chen JH. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nat Commun. 2021;12(1):1066.
32. Weikert T, Noordtzij LA, Bremerich J, Stieltjes B, Heye T, Boll DT. Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography. Korean J Radiol. 2020;21(7):891-9.
33. Yoon AP, Lee YL, Kane RL, Kuo CF, Lin C, Chung KC. Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. JAMA Netw Open. 2021;4(5):e216096.
34. Seo JW, Lim SH, Jeong JG, Kim YJ, Kim KG, Jeon JY. A deep learning algorithm for automated measurement of vertebral body compression from X-ray images. Sci Rep. 2021;11(1):13732.
35. Li J, Qian K, Liu J, Zhang Y, Wang L, Chen Z. Identification and diagnosis of meniscus tear by magnetic resonance imaging using a deep learning model. J Orthop Translat. 2022;34:91-101.
36. Zhang L, Li M, Zhou Y, Lu G, Zhou Q, Wang X. Deep learning approach for anterior cruciate ligament lesion detection: evaluation of diagnostic performance using arthroscopy as the reference standard. J Magn Reson Imaging. 2020;52(6):1745-52.
37. Yang J, Zhang K, Fan H, Liu X, Wang Y, Chen Z. Development and validation of deep learning algorithms for scoliosis screening using back images. Commun Biol. 2019;2:390.
38. Pedoia V, Lee J, Norman B, Link TM, Majumdar S, Souza RB. Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire Osteoarthritis Initiative baseline cohort. Osteoarthr Cartil. 2019;27(7):1002-10.