Journal of Southeast Asian Orthopaedics
ISSN 2821-9848 (Print)
ISSN 2821-9864 (Online)
https://doi.org/10.56929/jseaortho-2026-0292 https://jseaortho.org

Artificial Intelligence-Based Assessment of Hip Fracture Detection from Radiographic Images: Diagnostic Accuracy Compared with that of Orthopedic Surgeons and Radiologists

Withoone Kittipichai, MD

Department of Orthopaedics Surgery, Samut Sakhon Hospital, Samut Sakhon, Thailand

Purpose: Hip fracture is a major global public health concern and one of the leading causes of morbidity and mortality among older adults. Diagnostic inaccuracies often result in delayed treatment and poor outcomes. Artificial intelligence (AI) has shown promise in fracture detection, but studies did not fully reflect real-world clinical practice. We aimed to evaluate the feasibility and capacity of a YOLOv8-based AI model to detect hip fractures from anteroposterior pelvic radiographs as accurately as orthopedic surgeons and radiologists.

Methods: A total of 345 anonymized radiographs were used, comprising 45 images for physician comparison and 300 for extended testing. Various clinicians reviewed 45 images, evenly distributed among normal, femoral neck, and intertrochanteric fractures. Diagnostic accuracy, sensitivity, specificity, and error types were analyzed. The AI model was trained by simulating real-world hospital conditions.

Results: AI achieved an overall accuracy of 0.94, with 0.92 sensitivity and 0.91 specificity, comparable to radiologists and orthopedic surgeons and superior to physicians. Model performance remained stable when tested on the larger dataset ($p > 0.05$). Most errors occurred in minimally displaced femoral neck fractures, though accuracy for this group improved with larger test data. Mean processing time was 1.9–2.3 seconds per image.

Conclusions: The YOLOv8-based AI system demonstrated expert-level diagnostic performance and high processing efficiency without requiring advanced hardware. Our findings highlight its applicability in hospitals. Although occasional misclassifications and mislocalizations occurred, the model shows promise as a clinical decision-support tool for improving diagnostic confidence, reducing delays, and enhancing patient safety.

Keywords: Hip fracture, artificial intelligence, deep learning, YOLOv8, radiograph interpretation, clinical decision support

Article history:
Received: October 13, 2025 Revised: February 10, 2026
Accepted: March 21, 2026

Correspondence to: Withoone Kittipichai, MD
Department of Orthopaedics Surgery, Samut Sakhon Hospital, Samut Sakhon, Thailand
E-mail: withoone@gmail.com

Hip fractures represent a major global public health concern, particularly in aging societies, as they are strongly associated with disability, loss of independence, and increased mortality among older adults. The worldwide incidence of hip fractures continues to rise, with the number of cases projected to increase from ~1.6 million in 2000 to >6.3 million by 2050 [4]. This trend is especially evident in Asian countries undergoing rapid demographic transitions, where the incidence has increased more rapidly than in Western nations. Beyond the clinical implications, hip fractures impose significant economic and healthcare burdens, including prolonged hospitalization, surgical costs, and long-term rehabilitation.

In Thailand, which is approaching a super-aged society, more than 20,000 hip fracture cases occur annually, with a one-year mortality rate of 20–25%. Although surgical intervention is effective, diagnostic delays or misinterpretation of radiographs can lead to postponed surgery, complications, and poorer patient outcomes. Simunovic et al. [10] demonstrated that delays >48 hours after injury are associated with increased postoperative morbidity and mortality, underscoring the importance of rapid and accurate diagnosis in emergency settings.

In recent years, artificial intelligence (AI) has emerged as a promising tool for assisting radiographic interpretation. Deep learning algorithms, particularly convolutional neural networks (CNNs), have shown excellent performance in fracture detection. Studies by Krogue et al. [5] and Cheng et al. [2] demonstrated that AI systems can achieve a diagnostic accuracy comparable to expert radiologists in detecting hip fractures. However, most prior research has been conducted under controlled laboratory conditions, which may not fully represent real-world clinical practice, particularly in resource-limited hospitals. To bridge this gap, the present study evaluated the diagnostic performance of a YOLOv8-based AI model trained on real radiographic data from Thai patients at Tertiary Care Hospital. The model was designed, trained, and tested by the research team and deployed through a web-based application. Diagnostic performance was compared with that of physicians across varying levels of experience, including interns, emergency physicians, radiologists, and orthopedic surgeons, to assess its potential as a practical screening and decision-support tool in everyday clinical practice.

This study aimed to determine whether the AI model developed by the investigators could accurately detect hip fractures from plain anteroposterior radiographs with a diagnostic performance comparable to orthopedic surgeons and radiologists.

METHODS

Study Design and Setting

This diagnostic accuracy study was conducted at a Tertiary Care Hospital, a regional tertiary care center in Thailand, between 2017 and 2023. All patient identifiers were removed prior to analysis, and written informed consent was obtained from participating physicians.

Radiographic Dataset

The study utilized anteroposterior (AP) pelvic radiographs of both hips obtained from patients treated between 2017 and 2023. A total of 942 patients were reviewed, comprising:

- Intertrochanteric fractures: 320 cases.
- Femoral neck fractures: 245 cases.
- Normal hips: 377 cases.

After quality screening, 629 radiographs met the inclusion criteria and were selected for analysis, consisting of:

- Intertrochanteric fractures 150 images.
- Femoral neck fractures 102 images.
- Normal hips 377 images.

From this dataset, 300 images (100 per category) were randomly selected as the standard test dataset, from which a subset of 45 images (15 per class) were further selected for the physician evaluation subset administered via Google Forms. The subset size was intentionally selected to maintain a balanced class representation while minimizing reader fatigue, thereby allowing participation from multiple physician groups without compromising interpretation quality.

To avoid data leakage, the training and evaluation datasets were completely separated. No images used for model training were included in the testing datasets.

Inclusion Criteria

Radiographs were eligible for inclusion if they met the following conditions:

1. AP projection of both hips.

Clear visualization of the hip joint and pelvic structures, including the superior iliac crest and subtrochanteric region.

3. Fracture classification was confirmed by the gold standard of surgical findings or, in cases where surgery was not performed, by definitive radiological diagnosis. The cases were categorized into three groups:
    - Intertrochanteric fracture.
    - Femoral neck fracture.
    - Normal (no fracture).

Exclusion Criteria

Radiographs were excluded if they:

- Did not include both hips in full view, defined as complete visualization of both femoral heads, necks, acetabula, and extending superiorly to the iliac crests.
- Were underexposed (dark with obscured trabecular and cortical margins) or overexposed (bright with loss of cortical or trabecular detail).
- Showed motion blur or imaging artifacts (e.g., double contour, grid lines, or foreign objects obscuring bone edges).
- Had a mean grayscale intensity outside the range of 70–180 (8-bit scale) on histogram analysis, quantitatively assessed using Python OpenCV.

Physician Participants

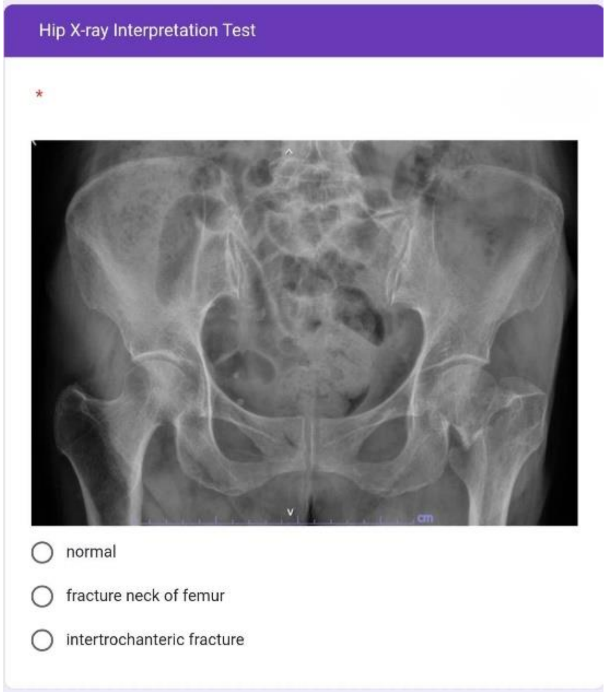
A total of 87 physicians participated, including:

- Interns: 65 (74.7%).
    - Year 1: 27 (31.0%).
    - Year 2: 20 (23.0%).
    - Year 3: 18 (20.7%).
- Emergency physicians: 7 (8.1%).
- Radiologists: 5 (5.8%).
- Orthopedic surgeons: 10 (11.5%).

Each participant interpreted 45 radiographs through a Google Form developed by the investigators. Images were randomly ordered for each respondent to minimize bias. No clinical history or additional contextual information was provided, ensuring evaluation relied solely on image interpretation.

Participants selected one of three options per image:

- Normal.
- Femoral neck fracture.
- Intertrochanteric fracture.



Hip X-ray Interpretation Test

*

○ normal

○ fracture neck of femur

○ intertrochanteric fracture

Fig. 1 Example of a hip radiograph interpretation assessment utilized in this study. The image demonstrates an anteroposterior view of bilateral hips. Participants were required to classify the radiographic findings into one of three diagnostic categories: normal anatomy, femoral neck fracture, or intertrochanteric fracture.

AI Model Development and Evaluation

The AI model used in this study was developed by the investigators using the You Only Look Once version 8 (YOLOv8) architecture, a state-of-the-art object detection framework capable of real-time lesion localization. The model was trained using hip radiographs categorized into three classes (normal, femoral neck fracture, intertrochanteric fracture) and validated through a web-based application deployed on the Hugging Face Spaces platform.

URL: https://huggingface.co/spaces/vithdata/hip_fracture

Hardware specifications (testing environment):

- Central Processing Unit (CPU): 2 vCPUs (Basic)
- Random-Access Memory: 16 gigabytes
- Processing: CPU-only (no Graphics Processing Unit [GPU])

Upon uploading a radiograph, the system automatically performed image inference and displayed:

- Predicted class (Normal / Neck / Intertrochanteric).
- Probability score (confidence level for each class).
- Bounding box highlighting the detected lesion region.
- Processing time (inference time) in seconds.

The model was evaluated on both the 45-image subset (identical to the physician test set) and the 300-image standard test set.

Definition of Region of Interest

The region of interest (ROI) denotes the proximal femur encompassing the femoral head, femoral neck, intertrochanteric, and subtrochanteric regions, corresponding to the anatomical area relevant for hip fracture diagnosis.

- Bounding boxes generated by the AI model were considered valid if they overlapped the predefined ROI.
- For each radiograph, the detection with the highest confidence score within the ROI was recorded as the final output of the AI model.

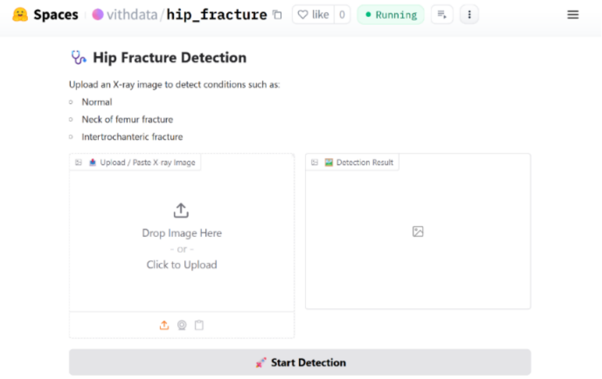
Detections located outside the ROI—such as those involving the sacroiliac joint, pelvic brim, or lumbar spine—were excluded from evaluation.

Gold Standard

All diagnostic labels were verified against the final confirmed diagnosis based on clinical evaluation and surgical findings, serving as the gold standard.

Statistical Analysis

- Diagnostic performance was assessed by calculating sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) from confusion matrices relative to the gold standard.
- McNemar's test and two-proportion Z-tests were used to evaluate differences between AI and physician performances on the same image set.
- One-way analysis of variance (ANOVA) ($p < 0.05$) was employed to compare the image-processing times of the AI model across various radiographic categories.
- Comparison between AI performance on 45 vs. 300 images was analyzed using both two-proportion Z-tests and Fisher's exact tests, with statistical significance defined as $p < 0.05$.

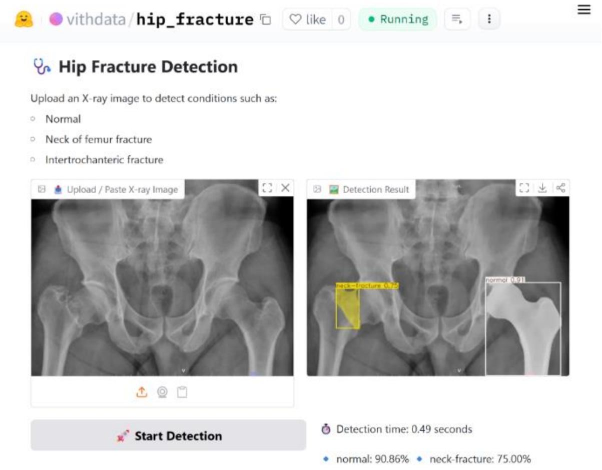


Hip Fracture Detection

Upload an X-ray image to detect conditions such as:

- Normal
- Neck of femur fracture
- Intertrochanteric fracture

Fig. 2 Web-based application interface for automated hip fracture detection. Users can upload radiographic images through either drag-and-drop functionality or direct file selection from the local storage.



Hip Fracture Detection

Upload an X-ray image to detect conditions such as:

- Normal
- Neck of femur fracture
- Intertrochanteric fracture

Detection time: 0.49 seconds

- normal: 90.86%
- neck-fracture: 75.00%

Fig. 3 Workflow of hip fracture detection employing the web-based YOLOv8 application. Following radiograph upload (left), the AI system

generates automated detection results (right) with color-coded bounding boxes and confidence scores: yellow box indicates suspected femoral neck fracture (confidence: 0.75) and white box denotes normal anatomy (confidence: 0.91). Classification probabilities are displayed as normal anatomy = 90.86% and femoral neck fracture = 75.00%. Image inference time was 0.20 seconds.

Ethical Considerations

This study was approved by the Human Research Ethics Committee of the Tertiary Care Hospital. All patient data were fully anonymized prior to analysis, and written informed consent was obtained from all participating physicians (Ethics Approval Number: SKH REC 27/2568/V.1).

RESULTS

Participant Characteristics

A total of 87 physicians participated in the study, comprising 65 interns (74.7%), seven emergency physicians (8.1%), five radiologists (5.8%), and 10 orthopedic surgeons (11.5%).

Among the interns, 27 (31.0%) were first-year, 20 (23.0%) were second-year, and 18 (20.7%) were third-year medical trainees. The participant distribution and characteristics are summarized in Table 1.

Table 1 Characteristics of Physician Participants.

Category	Number (n)	Percentage (%)
Total participants	87	100
Interns	65	74.71
- Year 1	27	31.03
- Year 2	20	22.99
- Year 3	18	20.69
Specialist	22	25.29
- Emergency physicians	7	8.05
- radiologists	5	5.75
- Orthopedic surgeons	10	11.49

Table 2 Diagnostic Performance of Physicians and the AI Model (95% Confidence Intervals).

Group	Class	Sensitivity	Specificity	PPV	NPV	Accuracy
Intern (All)	Neck	0.821 (0.795–0.844)	0.902 (0.887–0.914)	0.806 (0.780–0.831)	0.909 (0.896–0.922)	0.875 (0.862–0.886)
	Intertrochanter	0.855 (0.832–0.877)	0.934 (0.922–0.944)	0.866 (0.843–0.887)	0.928 (0.916–0.939)	0.908 (0.897–0.918)
	Normal	0.823 (0.797–0.846)	0.914 (0.901–0.926)	0.827 (0.801–0.850)	0.912 (0.898–0.924)	0.883 (0.871–0.895)
ER	Neck	0.981 (0.933–0.998)	0.881 (0.829–0.921)	0.805 (0.725–0.869)	0.989 (0.962–0.999)	0.914 (0.878–0.943)
	Intertrochanter	0.876 (0.798–0.932)	0.895 (0.846–0.933)	0.807 (0.723–0.875)	0.935 (0.892–0.965)	0.889 (0.849–0.921)
	Normal	0.667 (0.568–0.756)	0.986 (0.959–0.997)	0.959 (0.885–0.991)	0.855 (0.805–0.897)	0.879 (0.838–0.913)
Radiologist	Neck	0.920 (0.834–0.970)	0.933 (0.881–0.968)	0.873 (0.780–0.938)	0.959 (0.913–0.985)	0.929 (0.887–0.959)
	Intertrochanter	0.933 (0.851–0.978)	0.933 (0.881–0.968)	0.875 (0.782–0.938)	0.966 (0.921–0.989)	0.933 (0.892–0.962)
	Normal	0.827 (0.722–0.904)	0.973 (0.933–0.993)	0.939 (0.852–0.983)	0.918 (0.864–0.956)	0.924 (0.882–0.955)

Table 2 Diagnostic Performance of Physicians and the AI Model (95% Confidence Intervals). (Cont.)

Group	Class	Sensitivity	Specificity	PPV	NPV	Accuracy
Orthopedist	Neck	0.973 (0.933–0.993)	0.960 (0.931–0.979)	0.924 (0.871–0.960)	0.986 (0.965–0.996)	0.964 (0.943–0.980)
	Intertrochanter	0.987 (0.953–0.998)	0.970 (0.944–0.986)	0.943 (0.894–0.973)	0.993 (0.976–0.999)	0.976 (0.957–0.988)
	Normal	0.880 (0.817–0.927)	0.990 (0.971–0.998)	0.978 (0.936–0.995)	0.943 (0.911–0.966)	0.953 (0.930–0.971)
AI (45 images)	Neck	0.800 (0.519–0.957)	1.000 (0.884–1.000)	1.000 (0.735–1.000)	0.909 (0.757–0.981)	0.933 (0.817–0.986)
	Intertrochanter	1.000 (0.782–1.000)	1.000 (0.884–1.000)	1.000 (0.782–1.000)	1.000 (0.884–1.000)	1.000 (0.921–1.000)
	Normal	1.000 (0.782–1.000)	0.900 (0.735–0.979)	0.833 (0.586–0.964)	1.000 (0.872–1.000)	0.933 (0.817–0.986)
AI (300 images)	Neck	0.820 (0.731–0.890)	0.940 (0.898–0.969)	0.872 (0.788–0.932)	0.913 (0.865–0.947)	0.900 (0.860–0.932)
	Intertrochanter	0.910 (0.836–0.958)	0.960 (0.923–0.983)	0.919 (0.847–0.964)	0.955 (0.917–0.979)	0.943 (0.911–0.967)
	Normal	0.910 (0.836–0.958)	0.930 (0.885–0.961)	0.867 (0.786–0.925)	0.954 (0.914–0.979)	0.923 (0.887–0.951)

As shown in Table 2, the diagnostic accuracy of hip radiograph interpretation across various image groups (femoral neck fractures, intertrochanteric fractures, and normal radiographs), varied among physician groups and the AI model.

Orthopedic surgeons demonstrated the highest overall diagnostic accuracy across all categories, with an overall accuracy (Accuracy) ranging from 0.964 to 0.976. In the intertrochanteric fracture group, the model achieved a sensitivity of 0.987 and a specificity of 0.970, indicating superior ability to detect and differentiate between hip fractures.

Radiologists achieved the second-highest performance level, with accuracy values of 0.933 for intertrochanteric fractures and 0.929 for femoral neck fractures. The highest specificity (0.973) was observed in the normal radiograph category, reflecting a strong capability to correctly identify non-fracture cases and minimize false-positive interpretations.

Among interns, the highest diagnostic performance was observed in intertrochanteric fractures (accuracy = 0.908, sensitivity = 0.855, specificity = 0.934). In the normal and femoral neck fracture categories, accuracy values were 0.883 and 0.875, respectively. The NPV for interns was notably high, particularly in the intertrochanteric group (NPV = 0.928), indicating a reliable ability to confirm normal radiographs in the absence of fracture.

Emergency physicians (ER) showed a slightly higher overall accuracy than interns, with the best performance in the intertrochanteric fracture group (accuracy = 0.906, PPV = 0.860, sensitivity = 0.857), suggesting good reliability in confirming positive findings for hip fractures.

In Table 3, we employed the McNemar's test to compare diagnostic outcomes between the AI model and each physician group using the same set of 45 radiographs. The results revealed statistically significant differences between the AI model and interns in detecting intertrochanteric fractures ($p = 0.03$) and normal cases ($p = 0.03$), while no significant difference was observed in femoral neck fractures ($p > 0.05$).

For ER, statistically significant differences were observed in femoral neck fractures ($p = 0.03$) and normal cases ($p = 0.003$), but no significant difference was found in intertrochanteric fractures ($p > 0.05$).

When the performance of radiologists and orthopedic surgeons was compared, no statistically significant differences were observed across all fracture types (*p > 0.05*). These findings describe patterns of agreement and disagreement between the AI model and various physician groups. Given the limited sample size of the physician test subset, these results should be interpreted as exploratory comparisons rather than evidence of equivalence or superiority.

Table 3 McNemar's Test Comparison between the AI Model and Physician Groups.

Group	Class	N readers	% p<0.05	Median p	p-range	Median (b–c)	Median b	Median c
Intern	Neck	65	0%	0.24	0.18–0.65	1	2	1
	Intertrochanteric	65	100%	0.03	0.02–0.04	5	14	2
	Normal	65	100%	0.03	0.02–0.04	4	11	0
ER	Neck	8	100%	0.03	0.02–0.04	4	15	1
	Intertrochanteric	8	0%	0.16	0.15–0.20	3	10	5
	Normal	8	100%	0.003	0.002–0.005	6	9	0
Radiologist	Neck	5	0%	0.56	0.29–1.00	0	2	1
	Intertrochanteric	5	0%	0.56	0.29–1.00	0	1	1
	Normal	5	0%	0.29	0.18–0.54	0	1	0
Orthopedist	Neck	10	0%	0.32	0.29–0.54	1	3	1
	Intertrochanteric	10	0%	0.54	0.32–0.76	0	2	2
	Normal	10	0%	0.29	0.18–0.54	0	1	0

Table 4 Comparison of Diagnostic Accuracy between the AI Model and Physician Groups (Two-Proportion Z-Test).

Group	Class	Accuracy AI vs MD	Z-score	p-value	Interpretation
Intern	Neck	86.7% vs 66.7%	1.18	0.238	Not significant
	Intertrochanteric	93.3% vs 60.0%	2.16	0.031	Significant
	Normal	100% vs 73.3%	2.15	0.032	Significant
ER	Neck	86.7% vs 60.0%	2.12	0.034	Significant
	Intertrochanteric	93.3% vs 66.7%	1.41	0.158	Not significant
	Normal	100% vs 60.0%	3	0.003	Significant
Radiologist	Neck	86.7% vs 86.7%	0	1	Not significant
	Intertrochanteric	93.3% vs 93.3%	0	1	Not significant
	Normal	100% vs 93.3%	1.05	0.293	Not significant
Orthopedist	Neck	86.7% vs 86.7%	0	1	Not significant
	Intertrochanteric	93.3% vs 86.7%	0.61	0.54	Not significant
	Normal	100% vs 93.3%	1.05	0.293	Not significant

As shown in Table 4, a supplementary analysis was carried out by employing, the two-proportion Z-test to further compare the diagnostic accuracy between the AI model and each physician group. Statistically significant differences were observed between the AI model and interns in identifying intertrochanteric fractures (93.3% vs. 60.0%; *p = 0.031*) and normal cases (100% vs. 73.3%; *p = 0.032*). Similarly, statistically significant differences were observed between the AI model and ER in detecting femoral neck fractures (86.7% vs. 60.0%; *p = 0.034*) and normal cases (100% vs. 60.0%; *p = 0.003*).

However, when the performance of radiologists and orthopedic surgeons was compared, no statistically significant differences were observed across all categories (*p > 0.05*).

These findings are consistent with the McNemar's test results and describe differences in diagnostic performance patterns between various groups of caregivers. Given the limited sample size of the physician test subset, these results should be interpreted as exploratory comparisons rather than evidence of superiority or equivalence.

The mean inference time of the YOLOv8 model for image analysis is presented in Table 5. The mean processing time ranged from 1.9 to 2.3 seconds per image, with no statistically significant differences among the three image types (*p = 0.31* by ANOVA).

These findings suggest a consistent computational performance across various radiograph categories under CPU-only computation conditions.

Table 5 Mean AI Processing Time per Image.

Fracture Type	Mean ± SD (s)	Range (min–max)	p-value (ANOVA)
Neck of femur fracture	1.92 ± 1.34	(0.31 – 4.82)	
Intertrochanteric fracture	2.04 ± 1.51	(0.45 – 5.11)	
Normal	2.31 ± 1.72	(0.48 – 5.43)	0.31 (Nonspecific)

Table 6 Comparison of AI Model performance between 45 and 300 radiographs using two-proportion z-test and Fisher's exact test (p-value < 0.05).

Class	Metric	AI45 (%)	AI300 (%)	z-test	Fisher's	Interpretation
Neck	Sensitivity	80	82	0.8517	1	No statistically significant difference
	Specificity	100	94	0.3298	1	No statistically significant difference
Intertrochanter	Sensitivity	100	91	0.2262	0.6031	No statistically significant difference
	Specificity	100	96	0.4304	1	No statistically significant difference
Normal	Sensitivity	100	91	0.2262	0.6031	No statistically significant difference
	Specificity	90	93	0.9623	1	No statistically significant difference

Specifically, the mean processing times were 1.92 ± 1.34 seconds for femoral neck fractures, 2.04 ± 1.51 seconds for intertrochanteric fractures, and 2.31 ± 1.72 seconds for normal images. The minimal differences across various dataset categories indicate a consistent computational efficiency, confirming that the web-based AI system can rapidly process radiographic images without reliance on high-performance hardware.

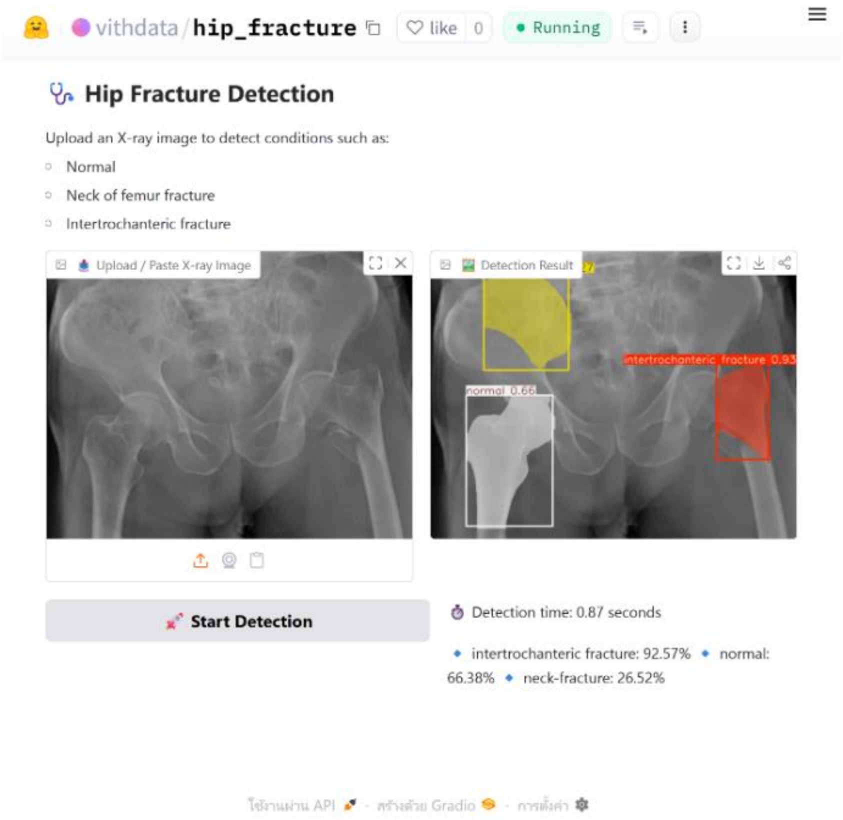
To evaluate the stability of the AI model under various data-volume conditions, performance testing was conducted using the same model applied to two datasets containing 45 and 300 radiographic images, respectively. The clinical performance indicators (sensitivity, specificity, and accuracy) for both datasets are summarized in Table 6. Although minor numerical differences were observed between the two datasets, statistical analyses using both Fisher's exact test and the two-proportion Z-test revealed no significant differences in any parameter (*p > 0.05*).

In the femoral neck fracture group, the sensitivity of the AI model slightly increased from 80% to 82%, whereas specificity was marginally reduced from 100% to 94% when the number of images increased from 45 to 300.

However, both the Z-test (*p = 0.8517 and 0.3298*) and Fisher's exact test (*p = 1.0000 for both comparisons*) indicated no statistically significant differences between the datasets. A similar trend was observed in the intertrochanteric fracture group, where sensitivity decreased from 100% to 91% and specificity was reduced from 100% to 96%.

Both Fisher's exact and Z-test analyses yielded *p*-values above 0.2 (*p = 0.6031–1.0000*), confirming that these variations were not statistically significant.

In the normal radiograph category, sensitivity decreased from 100% to 91%, while specificity increased modestly from 90% to 93%. Despite an opposing direction of change, statistical analysis again demonstrated no significant difference between the two datasets (*p = 0.2262–1.0000*).


vithdata / hip_fracture
like 0
Running
Upload / Paste X-ray Image
Detection Result
intertrochanteric fracture 0.93
normal 0.66
Start Detection
ใช้งานผ่าน API · สร้างด้วย Gradio · การตั้งค่า

Hip Fracture Detection

Upload an X-ray image to detect conditions such as:

- Normal
- Neck of femur fracture
- Intertrochanteric fracture

Detection time: 0.87 seconds

- intertrochanteric fracture: 92.57%
- normal: 66.38%
- neck-fracture: 26.52%

Fig. 4 An example of the output from the web-based hip fracture detection application. The user uploads a hip radiograph (left), and the AI system analyzes and displays the detection results (right). The model identified an intertrochanteric fracture on the right side of the image with a confidence score of 0.93 (red box) and interpreted certain regions as displaying normal anatomy (white box, confidence score = 0.66). A false-positive detection of a femoral neck fracture was also observed (yellow box, confidence score = 0.27). The overall classification probabilities were reported as intertrochanteric fracture = 92.57%, normal = 66.38%, and neck fracture = 26.52%, with a total processing time of 0.16 seconds.

Collectively, these findings suggest consistent diagnostic performance patterns across datasets of various sizes, although the results should be interpreted within the context of the study design.

In the AI image interpretation test, the model demonstrated high diagnostic accuracy but exhibited occasional localization errors. In certain cases, the attention map or bounding box focused on regions near the sacroiliac joint or pelvic rim instead of the actual fracture site (Figure 4). Heat map visualization further illustrated the regions where the model concentrated its attention, revealing that some detections outside the predefined ROI corresponded to high-contrast or overlapping bony structures. These detections were excluded from quantitative analysis since they were located outside the ROI.

DISCUSSION

This study evaluated the diagnostic performance of a YOLOv8-based AI system that was developed and trained by the investigators using real-world radiographic data from a regional tertiary care hospital. All fracture cases used in this study were confirmed by surgical findings, ensuring a high-quality reference standard (Gold Standard) for evaluating diagnostic accuracy. Although the model achieved high internal accuracy during the training phase, applying it in real-world clinical practice remains challenging due to variations in image quality, patient positioning, and anatomical differences. Therefore, this study aimed to assess the diagnostic capability of the model under actual clinical conditions, to identify its strengths and limitations, and to determine how AI can complement clinical workflows while guiding future model refinement.

Physicians from various experience levels, including interns, ERs, radiologists, and orthopedic surgeons, participated in the evaluation. The inclusion of ERs and interns was crucial, as they are typically the first responders in emergency settings where diagnostic delays or errors in hip fracture detection most frequently occur. Each physician interpreted 45 radiographs, equally distributed among radiographs depicting normal hips, femoral neck fractures, and intertrochanteric fractures. The actual number (45) of radiographs to be presented to each reader was selected to balance data diversity with the practical workload of clinicians. This design allowed the inclusion of more participants without causing excessive fatigue that could affect diagnostic accuracy. Although no formal sample size calculation was performed, the total of 3,915 individual readings provided a broad exploratory dataset for comparison between physician groups and the AI system.

Performance Comparison Between Physician Groups and the AI System

Our present results revealed clear differences in diagnostic accuracy among physician groups.

Orthopedic surgeons and radiologists achieved the highest accuracy (0.91–0.93), reflecting their expertise in musculoskeletal image interpretation and recognition of subtle cortical changes. In contrast, interns and emergency physicians showed a lower accuracy (0.75–0.82), particularly when distinguishing intertrochanteric fractures or differentiating normal from abnormal radiographs. This variation likely reflects differences in radiologic experience and the inherent complexity of emergency department images, which often involve overlapping soft tissues and inconsistent exposure.

The YOLOv8-based AI system achieved the highest overall accuracy (accuracy = 0.94, sensitivity = 0.92, specificity = 0.91), showing no statistically significant differences compared with the detection accuracy of radiologists and orthopedic surgeons in this study, and with statistically significant differences observed in comparison with less-experienced physician groups in specific categories, particularly in identifying intertrochanteric fractures and correctly classifying normal radiographs.

Our present findings are consistent with previous studies by Krogue et al.[5] and Lex Jr et al.[7], which demonstrated that deep learning models can reach near-expert performance and significantly reduce diagnostic variability among readers.

Model Stability and Processing Efficiency

The AI model showed consistent performance patterns, maintaining consistent accuracy between the 45- and 300-image test sets without statistically significant differences ($p > 0.05$).

This suggests that the model is robust and scalable when exposed to larger and more diverse datasets, a key characteristic for clinical deployment.

Regarding computational performance, the AI model achieved an inference time of 1.9–2.3 seconds per image using only CPU computation, without requiring GPU acceleration. This efficiency aligns with the findings of Kuo et al. [6] and Lex Jr et al. [7], who reported that AI systems with inference times <3 seconds per image can meaningfully reduce turnaround time in emergency care settings.

In clinical practice, physicians typically require several minutes per case, particularly when reviewing multiple projections or consulting with colleagues. Thus, AI integration may support workflow efficiency while maintaining comparable diagnostic performance within the scope of this study.

Error Analysis

Error analysis revealed that femoral neck fractures accounted for most misclassifications in the initial 45-image dataset. The majority of these were minimally or non-displaced fractures, which often lack clear cortical disruption and show only subtle trabecular changes, making them difficult to detect, even for experienced clinicians.

However, when tested with the larger 300-image dataset, the accuracy of the model for femoral neck fracture detection improved slightly, indicating that exposure to more diverse training data enhanced its ability to recognize complex fracture patterns. Nonetheless, this improvement did not reach statistical significance ($p > 0.05$), highlighting the remaining limitations of the model in identifying non-displaced or subtle fractures. This finding is consistent with prior reports by Cheng et al. [2, 3], Krogue et al.[5], and Wang et al. [12], which identified femoral neck fractures as one of the most diagnostically challenging categories for both human and AI readers.

Localization errors (mislocalization) were frequently observed near the iliac crest, acetabular rim, and sacroiliac joint, where overlapping bone structures and high-contrast transitions often misdirect the focus of the model. These results agree with Krogue et al. [5] and Cheng et al. [2, 3], who reported that deep-learning models tend to attend to visually prominent areas rather than diagnostically relevant fracture sites. These regions are commonly associated with texture heterogeneity and shadowing effects, which have been recognized as recurrent sources of false detections in musculoskeletal image analysis.

To mitigate classification and localization errors, several promising refinement approaches have been proposed. Data curation and augmentation, including contrast-aware adjustments, exposure jittering, and synthetic sample generation, can broaden image diversity and improve model generalization. A two-stage ROI-constrained pipeline can minimize off-target detections by localizing the femoral region before fracture classification, a design successfully applied in distal-radius fracture models (Min et al.[9]). Multi-view integration and test-time augmentation, such as flipping, rotation, and multi-view voting, can stabilize predictions under variable projection angles. Class-imbalance handling using focal loss, unified focal loss, or adaptive re-weighting improves sensitivity for rare or subtle fracture cases (Lin et al. [8], Kuo et al. [6] , Cheng et al. [2, 3]). Finally, enhancing generalizability through multicenter training and active learning with a reader-in-loop feedback framework supports continual model adaptation and reduces site-specific bias (Wang et al.[11], Sheller et al.[12]).

Clinical Implications

The findings of this study show that an AI model demonstrated promising potential, exhibiting both strong diagnostic accuracy and rapid processing speed, even when operating solely on CPU without the need for high-performance computing resources. This highlights its feasibility for use in general or community hospitals with limited technological infrastructure.

However, given the occasional occurrence of misclassification and mislocalization, AI should function primarily as a clinical decision-support tool, assisting rather than replacing physicians. When used alongside clinical assessment and physician judgment, AI may assist in enhancing diagnostic confidence and reducing potential delays in interpretation and ultimately improve diagnostic accuracy and patient safety.

Limitations

Several limitations of the present study should be acknowledged. First, each physician evaluated only 45 radiographs, which may not fully represent real-world clinical performance. Second, image evaluation was performed through an online platform rather than within a Picture Archiving and Communication System (PACS) environment of a hospital. Third, the AI model was trained on data from a single institution, which may limit generalizability to external datasets. Finally, only AP radiographs were analyzed; incorporating lateral or multi-view radiographs may potentially enhance diagnostic accuracy.

Future Directions

Future research should focus on multicenter studies to validate model generalizability across institutions, and on multi-view learning frameworks that integrate AP and lateral projections for improved fracture localization. The application of explainable AI techniques, such as Grad-CAM or attention mapping, could also enhance transparency and physician confidence by visualizing the reasoning process of the AI model. In addition, seamless and secure integration of AI into existing PACS or Hospital Information System infrastructures will be essential for real-world adoption, especially in hospitals with limited technical and human resources.

CONCLUSIONS

The YOLOv8-based AI model was trained on real radiographs from a regional tertiary care hospital and achieved high diagnostic accuracy, with no statistically significant differences observed between radiologists and orthopedic surgeons in this study. The model maintained stability when tested on larger datasets and demonstrated near real-time inference using only CPU processing.

Although its accuracy was slightly reduced for non-displaced or subtle fractures, the results support the integration of AI as an AI-assisted diagnostic tool under physician supervision to improve diagnostic speed, consistency, and quality, particularly in healthcare systems constrained by workforce and resource limitations.

REFERENCES

1. Cooper C, Campion G, Melton LJ. Hip fractures in the elderly: a world-wide projection. Osteoporos Int 1992;2:285-9.
2. Simunovic N, Devereaux PJ, Sprague S, et al. Effect of early surgery after hip fracture on mortality and complications: systematic review and meta-analysis. CMAJ 2010;182:1609-16.
3. Krogue JD, Cheng KV, Toogood P, et al. Automatic hip fracture identification and classification using deep learning. Radiol Artif Intell 2020;2:e190023.
4. Cheng CT, Chen CC, Cheng FJ, et al. A human-algorithm integration system for hip fracture detection on plain radiography: system development and validation study. JMIR Med Inform 2020;8:e19416.
5. Lex JR, Michele JD, Koucheki R, et al. Diagnostic accuracy of deep learning in detecting hip fractures: a systematic review and meta-analysis. JAMA Netw Open 2023;6:e233391.
6. Kuo RYL, MacKinnon T, Yeom K. Artificial intelligence and deep learning for fracture detection: a systematic review and meta-analysis. Radiology 2022;304:50-62.
7. Cheng CT, Ho TY, Lee TY, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29:5469-77.
8. Wang LX, Zhu ZH, Chen QC, et al. 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;14:1150-62.
9. Min H, Rabi Y, Wadhawan A, et al. Automatic classification of distal radius fracture using a two-stage ensemble deep learning framework. Phys Eng Sci Med 2023;46:877-86.
10. Lin TY, Goyal P, Girshick R, He K, Dollár P, editors. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017. p.2980-8.
11. Guan H, Yap PT, Bozoki A. Federated learning for medical image analysis: A survey. Pattern Recognit 2024;151:110424.
12. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medical imaging and genomics: Multi-institutional collaboration without sharing patient data. Sci Rep 2020;10:12598.