Journal of Southeast Asian Orthopaedics
ISSN 2821-9848 (Print)
ISSN 2821-9864 (Online)
https://doi.org/10.56929/jseaortho-2025-0263 https://jseaortho.org

Development of an Artificial Intelligence System for Hip Fracture Detection: A YOLOv8 Model Performance Study for Junior Orthopedic Surgeons

Withoone Kittipichai, MD

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

Purpose: Hip fractures represent a critical orthopedic emergency in the geriatric population; diagnostic delays or inaccuracies may result in severe morbidity and mortality. Contemporary artificial intelligence technologies demonstrate potential for precise and rapid radiographic interpretation, particularly in resource-constrained healthcare environments with limited availability of specialists. We aimed to develop and validate the diagnostic performance of a YOLOv8-based deep learning model by junior orthopedic surgeons for the detection of hip fractures, categorizing images into three classifications: normal anatomy, femoral neck fractures, and intertrochanteric fractures.

Methods: This retrospective study analyzed 2,035 anteroposterior hip radiographs from 942 patients. The YOLOv8 architecture was implemented using Google Colab with standardized hyperparameters. The dataset was stratified into training, validation, and testing sets. The performance evaluation utilized mean average precision (mAP@0.5), F1 score, precision, recall, sensitivity, specificity, and confusion matrix analysis.

Results: The YOLOv8 model achieved an mAP@0.5 of 0.879 and a maximum F1 score of 0.86. The model demonstrated a maximum precision, confidence threshold, and maximum recall of 1.00, 0.961, and 0.91, respectively, at a confidence threshold of 0.000. The sensitivity values were 97.7%, 87.0%, and 95.9% for intertrochanteric fractures, femoral neck fractures, and normal anatomy, respectively. The specificity ranged from 97.1% to 99.0% across all classifications, indicating exceptional screening accuracy, particularly for normal anatomy and intertrochanteric fractures.

Conclusions: The YOLOv8 model demonstrated clinical utility as a diagnostic screening tool for hip fractures, particularly in facilities with limited radiological expertise. However, femoral neck fracture classification requires further refinement through dataset augmentation and advanced training methodologies to enhance detection accuracy for this radiologically challenging entity.

Keywords: Hip fracture, YOLOv8, artificial intelligence, radiography, junior orthopedic surgeons, sensitivity, specificity

Article history:
Received June 30, 2025 Revised: August 24, 2025
Accepted: October 6, 2025

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

Thailand is experiencing a profound demographic transition characterized by rapid population aging. The national population of 66,052,615 in 2023 is projected to include 20% of individuals aged 60 years or older by 2024, representing a dramatic increase from 6.8% in 1994. This demographic shift correlates with an escalating hip fracture incidence, particularly among patients with osteoporosis. The United States reports more than 250,000 hip fractures annually, with global projections indicating that the number of cases will increase substantially by 2050[1].

Hip fracture diagnosis traditionally relies on a comprehensive clinical assessment that incurporates patient history, physical examination, and plain radiographic evaluation. However, diagnostic delays or misinterpretations may result in catastrophic complications, including increased mortality rates. Primary care facilities report a misdiagnosis rate of 14%[2], with physician experience in radiographic interpretation serving as a critical determinant[3]. Previous investigations revealed that first-year junior doctors achieve a diagnostic sensitivity of only 73.1–76.9%, whereas specialist orthopedic surgeons attain a sensitivity of 96.2%[3].

Recent advances in artificial intelligence (AI), particularly deep learning architectures and convolutional neural networks (CNNs), have generated considerable interest in automated radiographic diagnosis. International research has demonstrated that deep learning applications for wrist fracture detection achieve 95.2% accuracy[4], whereas CNNs for hip fracture identification yield a sensitivity of 92.7% and specificity of 95%[5]. Cheng et al.[6] developed a DenseNet-121 model for hip-fracture detection, achieving 98% sensitivity and 91% accuracy.

You Only Look Once (YOLO) is a highly regarded computer vision architecture renowned for its superior speed and accuracy in object detection and image segmentation. Since the initial YOLO release in 2015, continuous development has culminated in YOLOv8, the current state-of-the-art version that demonstrates enhanced performance with low-resolution images and partially occluded objects.

YOLOv8 employs a single-stage detector architecture optimized for real-time object detection. This model processes all images simultaneously to predict bounding boxes and class labels for objects of interest, in contrast to two-stage detectors that require separate region proposal and classification phases. This integration provides YOLOv8 with a superior processing velocity while maintaining exceptional detection accuracy across diverse object categories.

This study aimed to develop and evaluate an AI system utilizing the YOLOv8 architecture to assist in hip fracture diagnosis performed by junior orthopedic surgeons (first- to third-year residents), thereby reducing misdiagnosis rates in healthcare facilities with limited availability of radiological and orthopedic specialists.

MATERIALS AND METHODS

Study Design

This retrospective study utilized a comprehensive database of anteroposterior hip and pelvic radiographs retrieved from the Picture Archiving and Communication System (PACS) at a tertiary care hospital from 2017 to 2023.

Population and Sample

1. Definitions and Classification Criteria:

- Normal Hip:
  - Radiographic appearance of normal anatomical characteristics of the hip.
  - Intact cortical bone continuity.
  - Absence of fracture lines or trabecular pattern disruption.
  - Femoral neck-shaft angle within the normal range (120–135°).
- Femoral Neck Fracture:
  - Fracture line within the femoral neck region.
  - Anatomical location between the femoral head and greater trochanter.
  - Trabecular pattern alterations.
  - Potential cortical disruption or step-off deformity.
  - Classification according to Garden criteria (Types I–IV).
- Intertrochanteric Fracture:
  - Fracture line localized between the greater and lesser trochanters.
  - Cortical bone alignment alterations.
  - Associated trabecular pattern fragmentation.
  - Possible displacement of bony fragments.
  - Classification according to AO/OTA criteria.

2. Inclusion Criteria:

- Patients receiving medical care at the tertiary care hospital (2017–2023).
- Age ≥30 years.
- Radiographic images with adequate resolution for comprehensive anatomical evaluation.

3. Exclusion Criteria:

- Previous surgical intervention with metallic internal fixation devices.
- Radiographs with indeterminate or ambiguous fracture patterns (including images obtained while the patient was on a stretcher or lifting device, where supporting equipment obscured anatomical structures and may alter fracture appearance).
- Patients with concurrent diagnoses of osteoporosis combined with other hip pathologies that significantly altered hip joint anatomy, including:
  - Septic arthritis of the hip.
  - Avascular necrosis of the femoral head.
  - Advanced hip osteoarthritis.

4. Sample Size:

- Total: 2,035 images from 942 patients.
- All radiographs used for model development and testing were obtained from patients who had been definitively diagnosed and treated for hip fracture. Therefore, the ground truth labels were based on confirmed postoperative diagnoses documented in the patients' medical records.
- Distribution: femoral neck fractures (515 images, 25.3%), intertrochanteric fractures (687 images, 33.8%), and normal anatomy (833 images, 40.9%). Demographics: 566 women (60.1%), 376 men (39.9%); age range 40–99 years.
- A formal sample size calculation was not applicable in this study because the objective was to train and validate a deep learning model rather than to test a statistical hypothesis. In computer vision research, model performance typically improves with increasing data volume and diversity up to the point of convergence. Therefore, all eligible radiographs 2,035 images from 942 patients were included to maximize representativeness and minimize sampling bias.

Data Collection Methods

1. Image Data Acquisition:

Anteroposterior view of the hip obtained from the tertiary care hospital.

- Initial hip fracture radiographs were acquired in the **non-traction position**; patients were not placed under traction before imaging.
- All radiographs used in this study were retrieved directly from the hospital's PACS in their **original diagnostic form**, without post-processing of **contrast** or **sharpness**. The only modifications permitted before region-of-interest extraction were **zoom-in or zoom-out adjustments** to optimize visualization during screen capture.
- Images captured the hip area, specifying side and type of hip (normal, femoral neck fracture, and intertrochanteric fracture), using the Windows 11 Snipping Tool application.
- Region of interest limited to hip joint anatomy.
- All patient information was completely de-identified before analysis.

2. Image Data Management:

- Dimensions: 213 × 187 to 672 × 612 pixels.
- File sizes: 4–450 kB.

AI Model Development

1. Data Preparation and Annotation:

- Roboflow annotation tools (Smart Polygon) were utilized for precise lesion delineation.
- Annotation was supervised by an experienced orthopedic surgeon.
- Data augmentation techniques were implemented:
  - Auto-orientation correction.
  - Horizontal flip transformation.
  - Bounding box noise addition (0.1% pixel modification).
  - Histogram equalization enhancement.

2. Dataset Partitioning:

- Following augmentation, 4,878 images were systematically divided:
  - Training set: 4,268 images (87.5%).
  - Validation set: 407 images (8.3%).
  - Test set: 203 images (4.2%).

3. Model Training:

- Architecture: YOLOv8.

- Training Parameters:
- Batch size: 16
  - The number of radiographic images processed simultaneously before each parameter update. A moderate batch size was selected to balance computational efficiency and stability of the learning process.
- Epochs: 200
  - The model was exposed to the entire dataset for 200 complete training cycles. This number was selected to ensure sufficient learning of data patterns while monitoring validation loss to minimize overfitting.
- Learning rate: 0.001
  - The step size used for updating model weights during optimization. The selected value is a commonly applied setting for CNN models, providing an appropriate balance between conver-gence speed and training stability.
- Hardware Specifications:
  - Platform: Google Colab.
  - CPU: six cores, 12 logical processors.
  - GPU: NVIDIA A100-SXM4-40 GB.

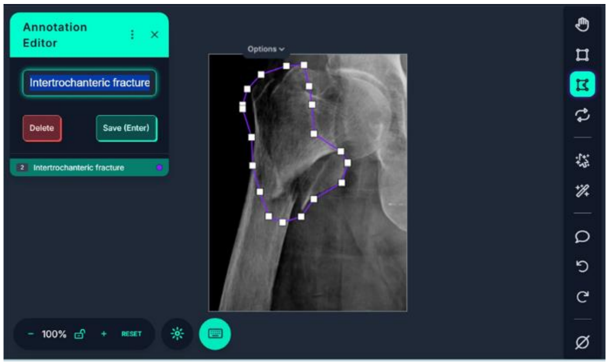


Fig. 1 Lesion localization process using Roboflow annotation tools, showing the precise delineation of fracture boundaries for model training purposes.

Performance Evaluation Metrics

1. Primary Evaluation Parameters

- The model performance was assessed using standard machine learning metrics automatically generated by the YOLOv8 framework during the training and validation phases.

- **Precision (Positive Predictive Value):** Calculated as Precision = TP/(TP + FP), where TP = true positives and FP = false positives. This metric represents the proportion of correctly identified fractures among all positive predictions.

- **Recall (Sensitivity/True Positive Rate):** Calculated as Recall = TP/(TP + FN), where FN = false negatives. This metric measures the ability of the model to identify all actual fracture cases.

- **F1 score:** Calculated as F1 = 2 × (Precision × Recall)/(Precision + Recall). This represents the harmonic mean of precision and recall, providing a balanced performance assessment.

- **Specificity (True Negative Rate):** Calculated as Specificity = TN/(TN + FP), where TN = true negatives. This metric measures the model's ability to correctly identify normal cases.

- **Mean Average Precision (mAP@0.5):** Calculated by averaging precision across different recall levels at the Intersection over Union threshold of 0.5, providing a comprehensive object detection performance assessment.

- **Confusion Matrix:** A 3 × 3 matrix displaying the actual versus predicted classifications for the three categories (normal, femoral neck fracture, and intertrochanteric fracture), enabling a detailed analysis of classification errors and performance across all classes.

2. Analysis of Results:

- The primary focus of the analysis was a comparative evaluation of the model’s accuracy in classifying fracture types.

- The rate of misdiagnosis (error rate) was also analyzed.

Statistical Analysis

Descriptive statistics including frequency, mean, and standard deviation were computed. All proportion-based performance metrics (sensitivity, specificity, PPV, NPV, accuracy, and F1-score) were reported with 95% confidence intervals (95% CI) using the Wilson score method, which provides reliable estimation for binomial data with moderate sample size.

As this stage represented internal model validation, no physician comparisons were performed. Descriptive statistics, including frequency distributions, percentages, means, and standard deviations, were calculated. The model performance assessment incorporated sensitivity, specificity, and accuracy.

RESULTS

Demographic Characteristics

The study cohort comprised 2,035 hip radiographs from 942 patients, with a mean age of 72.4 ± 8.6 years (range: 40–99 years). The population consisted of 566 women (60.1%) and 376 men (39.9%). Image classification yielded 515 femoral neck fractures (25.3%), 687 intertrochanteric fractures (33.8%), and 833 normal studies (40.9%).

Model Performance Analysis

F1-Confidence Relation

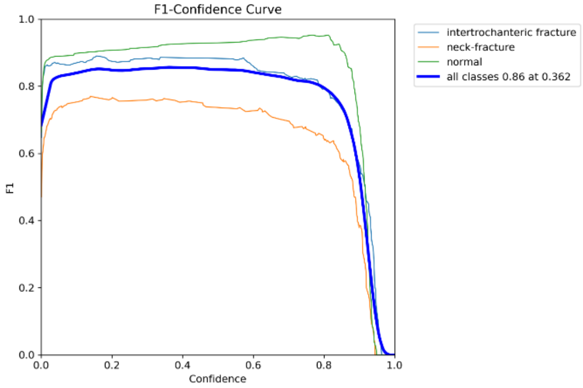


Fig. 2 F1-confidence curve, illustrating the relation between the balanced accuracy metric and confidence levels.

Figure 2 presents the **F1-confidence curve**, illustrating the relation between:

- **Model confidence** in predicting whether a radiograph shows a fracture, and
- **F1-score**, a balanced performance metric that incorporates both sensitivity (ability to correctly identify fractures) and precision (ability to avoid false positives).

The F1-confidence curve demonstrated optimal performance with a maximum aggregate F1 score of 0.86 at a confidence threshold of 0.362. Across categories, normal anatomy achieved the highest F1 scores, followed by intertrochanteric fractures, whereas femoral neck fractures demonstrated the lowest values.

Precision Analysis

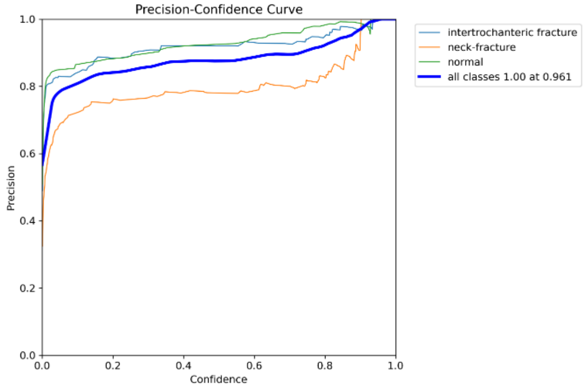


Fig. 3 Precision-confidence curve, demonstrating precision performance across varying confidence thresholds.

Figure 3 displays the precision-confidence curve, which is used to illustrate how the performance of the model (precision mean positive predictive value) changes as the model's confidence threshold is adjusted upward or downward. Maximum overall precision of 1.00 was achieved at a confidence score of 0.961, indicating exceptional accuracy at elevated confidence levels. Normal anatomy and intertrochanteric fractures maintained consistently high precision, whereas femoral neck fractures showed comparatively lower precision.

Precision-Recall Performance

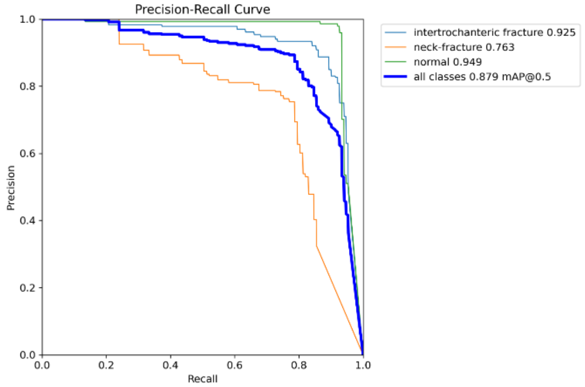


Fig. 4 Precision-Recall Curve, depicting the trade-off between precision and recall.

X-axis = Recall (Sensitivity): Represents the proportion of *true positive cases* correctly identified by the model. High recall indicates fewer missed cases.

Y-axis = Precision (Positive Predictive Value): Represents the proportion of predicted positive cases that were *true positives*. High precision indicates fewer false positives.

As the **confidence threshold** was adjusted from low to high, pairs of values (recall, precision) were generated, forming a curve.

The **area under the curve (average precision**) was calculated for each class, and the mean value across all classes was reported as **mAP**. Figure 4 shows that the precision–recall curve and mAP@0.5 reached 0.879. Class-specific precision values were 0.949 for normal anatomy, 0.925 for intertrochanteric fractures, and 0.763 for femoral neck fractures.

Recall Analysis

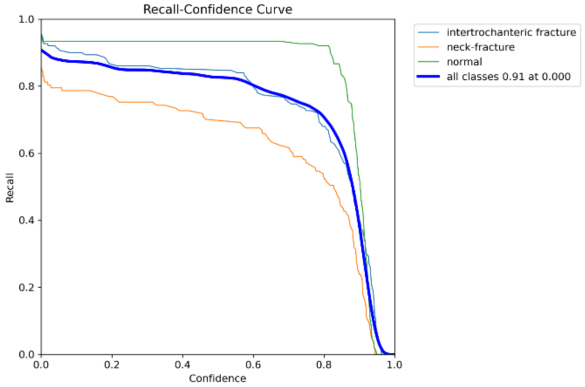


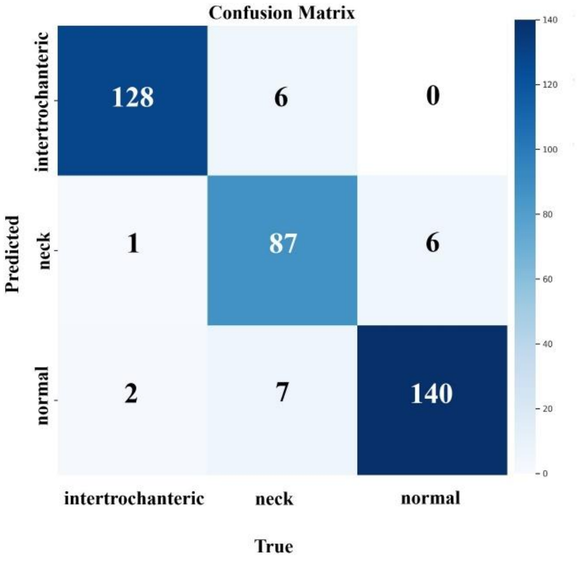
Fig. 5 Recall-confidence curve, which represents recall performance at different confidence levels.

X-axis = Confidence (model confidence level): a value ranging from zero to one indicates how certain the model is before making a final decision.

Y-axis = Recall (Sensitivity): the proportion of true positive cases that the model correctly identified. High recall indicates fewer missed cases.

As the confidence threshold increased from low to high, the model became more stringent in its predictions, causing the recall to gradually decline and then drop sharply near the higher end of the scale (approximately 0.9–1.0). Figure 5 shows that overall recall reached 0.91 at a confidence score of 0.000, demonstrating comprehensive lesion detection capability. Femoral neck fractures exhibited the lowest recall performance among all classifications.

Classification Accuracy Matrix



Confusion Matrix

Predicted \ True	intertrochanteric	neck	normal
intertrochanteric	128	6	0
neck	1	87	6
normal	2	7	140

Fig. 6 Confusion matrix, providing a comprehensive view of the classification performance across all categories.

As shown in Figure 6, the confusion matrix analysis revealed the highest classification accuracy for normal anatomy, followed by intertrochanteric fractures. The highest misclassification rate occurred within the femoral neck fracture category.

Sensitivity and Specificity Analysis

Detailed analysis of diagnostic performance by fracture type revealed, as shown in Table 1:

Sensitivity (True Positive Rate):

- Intertrochanteric fractures: 97.7%
- Normal anatomy: 95.9%
- Femoral neck fractures: 87.0%

Specificity (True Negative Rate):

- All classifications: 97.1–99.0%

These results indicate exceptional screening capability, particularly for normal anatomy and intertrochanteric fracture detection.

DISCUSSION

Model Performance Comparison

This study employed the YOLOv8 architecture for hip fracture diagnosis using radiographic images obtained from the tertiary care hospital PACS. The dataset of 2,035 images from 942 patients exceeded the sample sizes reported in previous studies, including those of Beyaz et al.[7] (724 images) and Lee et al.[8] (459 images), suggesting adequate statistical power for model training.

The tripartite classification system (normal, femoral neck, and intertrochanteric fractures) aligns with established clinical practice and encompasses the most frequently encountered and diagnostically challenging hip fracture patterns. The achieved sensitivity and specificity values demonstrated high performance, approximating or exceeding those of studies utilizing larger training datasets.

The overall model performance showed an average sensitivity and specificity exceeding 90%, with an F1 score of 0.86, which compared favorably with the existing literature. Analyses by fracture type revealed a lesion detection capability exceeding 85% across all categories. Intertrochanteric fractures achieved a sensitivity of 97%, whereas femoral neck fractures demonstrated a sensitivity of 87%. Specificity consistently exceeded 97% for all classifications, indicating minimal false-positive rates and excellent screening utility.

Table 1 Sensitivity and specificity of the model.

Research Study	Model	Total Train Images	Sensitivity (%)	Specificity (%)	F1-score
This study	YOLOv8	2,035	91	95	0.86
Beyaz et al. 2023	Xception + EfficientNet-B7 + NFNet-F3	724	95.97	91.7	0.917
Krogue et al. 2020	DenseNet (w/ detection module)	3,034	92.7	95	0.9
Lee et al. 2020	Meta-learned DNN	459	87	87	0.867
Yildiz Potter et al. 2024	VarifocalNet FPN	823	95	94	0.98

Table 2 Sensitivity and specificity metrics stratified by disease group.

Class	TP	FN	FP	TN	Sensitivity (%)	Specificity (%)	Precision (PPV) (%)	NPV (%)	Accuracy (%)	F1-score
Intertrochanteric	128	3	6	250	97.7	97.6	95.5	98.8	97.7	0.966
Neck	87	13	3	284	87	99	96.7	95.6	95.6	0.915
NormalHip	140	6	7	234	95.9	97.1	95.2	97.5	96.5	0.956
Overall	355	22	16	768	94.2	98	95.7	97.2	96.5	0.949

TP, true positive; FN, false negative; FP, false positive; TN, true negative; PPV, positive predictive value; NPV, negative predictive value.

AI Architecture Comparison

CNN Models

The traditional CNN architectures (DenseNet, ResNet, VGG16, and Inception-V3) utilized by Krogue[5], Cheng[6], and Lee[8] offer advantages for largescale image training and architectural simplicity. However, these models lack visual lesion localization capabilities, which limits the clinical verification of diagnostic decisions.

Ensemble Model Approaches

Beyaz et al.[7] investigated ensemble methodologies incorporating the Xception, EfficientNet, and NFNet architectures using majority-voting techniques. While individual models demonstrated rapid performance and reduced computational requirements, ensemble implementation necessitated multi-model analysis, increasing developmental complexity and computational resource demands.

Object Detection Models

Object detection architectures (YOLOv5, YOLOv8, Feature Pyramid Networks) provide direct lesion identification and localization capabilities while managing complex compositional elements. Both the study by Potter et al.[9] and current investigation demonstrated robust FPN and YOLO performance, with sensitivity and specificity exceeding 90%. The primary limitation involves time-intensive, resource-demanding, and precise lesion annotation requiring expert supervision.

Clinical Advantages of YOLOv8

In contrast to CNN models that determine fracture presence or absence without lesion localization, the YOLOv8 architecture offers substantial clinical advantages through simultaneous object detection and classification capabilities. This functionality renders YOLOv8 exceptionally suitable for automated diagnostic assistance systems, particularly in resource-constrained environments and for supporting junior medical trainees.

Additionally, YOLOv8 demonstrated superior performance with suboptimal image quality or partially obscured lesions, reflecting real-world clinical scenarios involving variable projection angles, image sharpness variations, and metallic implant interference.

Potential Causes of AI Misclassification

In this study, heatmap-based visualization, such as Grad-CAM, was not incorporated, and therefore the exact sites where the AI failed to detect fractures could not be localized. Nevertheless, previous studies have provided insights into common sources of error. Cheng et al.[6] demonstrated that AI often misinterprets subtle trabecular changes in heatmaps, while Krogue et al.[5] reported particularly low sensitivity for nondisplaced femoral neck fractures, consistent with our results, in which femoral neck fractures had lower sensitivity than those of intertrochanteric fractures. Similarly, Pinto et al.[3] highlighted that subtle or occult fractures on plain radiographs are challenging even for radiologists and thus remain a limitation for AI. Beyaz et al.[7] showed that using ensemble CNN models and multicenter data improved generalizability and reduced false positives, supporting the notion that broader and more diverse datasets may mitigate some of the failure modes observed in our model. While our study excluded postoperative images with metal implants to avoid confounding artifacts, prior studies (Shi et al.[2]) emphasized that variability in radiographic exposure and image quality remains a major source of diagnostic error. Collectively, these comparisons suggest that misclassification in our model likely arose from subtle nondisplaced fractures, limited dataset size for certain subgroups, and the inherent limitations of plain radiography.

Annotation Bias

Previous studies have highlighted that data labeling can be prone to errors, particularly in subtle or borderline fractures that may be interpreted as “normal.” Pinto et al.[3] reported that missed diagnoses on plain radiographs in emergency settings are relatively common and can directly translate into annotation bias when training AI models. Similarly, Lindsey et al.[4] demonstrated that although deep neural networks improve fracture detection by clinicians, subtle fractures remain a significant challenge. To minimize this issue, all images in our study were reviewed and grouped by the treating orthopedic surgeon before training.

Overfitting and Underfitting

The risks of overfitting and underfitting have been well documented in prior studies. Krogue et al.[5] noted that models trained on single-center datasets may overfit specific image characteristics and perform poorly in external settings. Cheng et al.[6] emphasized the importance of dataset size and diversity and noted that insufficient variability can lead to underfitting and reduced generalizability. By contrast, Beyaz et al.[7] demonstrated that training on multicenter datasets with ensemble models mitigated overfitting and improved diagnostic robustness. Our study attempted to address these limitations through careful annotation review, but the relatively small sample size of femoral neck fractures may have contributed to underrepresentation and low sensitivity in this subgroup.

Potential Implementation Barriers and Solutions

One of the major barriers to applying AI models in real-world clinical practice is concern regarding diagnostic accuracy and reliability. For this reason, we consider the proposed model to be the most valuable *decision-support tool* to assist clinicians, particularly junior doctors, in confirming or validating their initial interpretation rather than fully replacing human judgment. This approach can help improve confidence in diagnosis while minimizing the risk of overreliance on AI.

Regarding cost and feasibility, because the YOLOv8 model has already been developed and trained, it can be deployed on the intranets of healthcare facilities without requiring expensive infrastructure. Moreover, the model can also be implemented through free hosting platforms, such as Hugging Face, which allows the tool to be accessed by multiple centers at no additional cost. This flexibility supports practical adoption, particularly in resource-limited hospitals.

Limitations Regarding the Single-Center Design and Exclusion Criteria

A key limitation of this study is that all data were collected from a single hospital, which may reduce the generalizability of the results to other populations or imaging environments. However, the choice of YOLOv8 as the core architecture provides advantages because it is designed to handle images of varying quality, including lower-resolution or partially obscured images, making it more adaptable to real-world radiographs from different institutions. Future research should expand to include multicenter datasets to validate the external applicability of the model.

Another important limitation of this study is the exclusion of patients who had osteoporosis combined with other hip pathologies, such as septic arthritis of the hip, avascular necrosis of the femoral head, and advanced hip osteoarthritis. These conditions were excluded because they often cause significant anatomical distortion or cortical bone irregularity, making it difficult for the model to accurately learn and classify normal versus fractured anatomy during the initial training phase. Nevertheless, the ability to recognize fractures in atypical or deformed hip anatomy represents an important opportunity for future model improvement.

Clinical Implementation Potential

Based on the present findings, the YOLOv8 model demonstrates strong potential for real-world clinical integration. With sensitivity, specificity, and F1 scores consistently above 90%, the model provided sufficient diagnostic reliability for application as a supportive screening tool. Its real-time processing speed allows for rapid decision-making in emergency departments, which is crucial for minimizing delays in hip fracture management. Importantly, the model can serve as a decision-support mechanism for junior doctors in settings with limited radiological coverage, thereby enhancing diagnostic safety. Furthermore, because the system can be deployed on hospital intranets or secure web platforms without extensive infrastructure investment, it is highly scalable and accessible, even in resource-limited hospitals. This scalability extends to telemedicine networks, where peripheral clinics may benefit from AI-assisted preliminary interpretations before confirmation by orthopedic specialists. These strengths suggest that YOLOv8 is not only technically robust but also clinically feasible and cost-effective, making it a promising candidate for widespread implementation in diverse healthcare settings.

Recommendations for Future Development

Based on the sensitivity and specificity analyses, opportunities for improvement include femoral neck fracture dataset augmentation and advanced data augmentation techniques during model training. Potential enhancements include controlled brightness and sharpness adjustments, minor rotational modifications, and controlled noise introduction to facilitate diverse image learning and reduce overfitting, which are particularly relevant for morphologically complex femoral neck fractures.

CONCLUSIONS

The YOLOv8-based model demonstrated significant clinical potential, with performance metrics closely aligned with established research findings from Krogue[5], Cheng[6], and Lee[8], while approaching the results reported by Beyaz et al.[7]. The hip fracture diagnostic efficiency of the model consistently exceeded 90%, indicating its robust capability for real-world clinical applications.

Importantly, although the model showed high overall sensitivity and specificity, certain limitations were noted in failure cases. For example, subtle or nondisplaced fractures, borderline cases between normal and fractured, and images with lower quality or anatomical variations occasionally led to misclassifications. These findings are consistent with challenges reported in previous studies[5–8].

Such failure cases highlight the necessity for larger and more diverse training datasets, improved annotation consistency, and possible integration with multiview or multimodality imaging to enhance detection performance. Additionally, interpretability tools, such as heat maps or attention maps, could be applied in future work to precisely identify where AI may misread fracture signals.

In summary, this study established YOLOv8 as a highly appropriate architecture for hip fracture diagnostic assistance from radiographic images, supported by superior performance metrics, processing velocity, localization capabilities, and alignment with practical healthcare delivery requirements. Nevertheless, continued refinement is essential to minimize missed diagnoses and strengthen confidence in its clinical adoption.

ACKNOWLEDGMENTS

The authors acknowledge the support of the tertiary care hospital and Human Research Ethics Committee in facilitating this study.

APPENDIX

Source code for training the YOLOv8 model:
https://drive.google.com/file/d/1sb4gwajjgMHLF1kEf66YXjjSuP7W3eWK/view?usp=sharing

REFERENCES

1. Sing CW, Lin TC, Bartholomew S, et al. Global epidemiology of hip fractures: secular trends in incidence rate, post-fracture treatment, and all-cause mortality. J Bone Miner Res 2023;38:1064-75.

2. Shi BY, Hannan CV, Jang JM, et al. Association between delays in radiography and surgery with hip fracture outcomes in elderly patients. Orthopaedics 2020;43:e609-e15.

3. Pinto A, Berritto D, Russo A, et al. Traumatic fractures in adults: missed diagnosis on plain radiographs in the Emergency Department. Acta Biomed 2018;89:111-23.

4. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018;115:11591-6.

5. Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and

functional subclassification with deep learning. Radiol Artif Intell 2020;2:e190023.

6. 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.

7. Beyaz S, Yayli SB, Kilic E, et al. The ensemble artificial intelligence (AI) method: Detection of hip fractures in AP pelvis plain radiographs by majority voting using a multi-center dataset. Digit Health 2023;9:20552076231216549.

8. Lee C, Jang J, Lee S, et al. Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. Sci Rep 2020;10:13694.

9. Beyaz S, Yayli SB, Kilic E, et al. Comparison of artificial intelligence algorithm for the diagnosis of hip fracture on plain radiography with decision-making physicians: a validation study. Acta Orthop Traumatol Turc 2024;58:4-9.