Heliyon 8 (2022) e11266

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Artificial intelligence (AI) vs. human in hip fracture detection

Nattaphon Twinprai [a], Artit Boonrod [b], Arunnit Boonrod [c], Jarin Chindaprasirt [d], Wichien Sirithanaphol [e], Prinya Chindaprasirt [f], Prin Twinprai [g,*]

[a] Trauma Unit, Department of Orthopedics, Srinagarind Hospital, Khon Kaen University, Thailand
[b] Sport Unit, Department of Orthopedics, Srinagarind Hospital, Khon Kaen University, Thailand
[c] Neurology Unit, Department of Radiology, Srinagarind Hospital, Khon Kaen University, Thailand
[d] Department of Internal Medicine, Srinagarind Hospital, Khon Kaen University, Thailand
[e] Department of Surgery, Srinagarind Hospital, Khon Kaen University, Thailand
[f] Sustainable Infrastructure Research and Development Center, Department of Civil Engineering, Faculty of Engineering, Khon Kaen University, Thailand
[g] Musculoskeletal Unit, Department of Radiology, Srinagarind Hospital, Khon Kaen University, Thailand

ARTICLE INFO

Keywords:
Hip fracture
Computer vision
Artificial intelligence
Deep learning
Trauma

ABSTRACT

Objective: This study aimed to assess the diagnostic accuracy and sensitivity of a YOLOv4-tiny AI model for detecting and classifying hip fractures types.

Materials and methods: In this retrospective study, a dataset of 1000 hip and pelvic radiographs was divided into a training set consisting of 450 fracture and 450 normal images (900 images total) and a testing set consisting of 50 fracture and 50 normal images (100 images total). The training set images were each manually augmented with a bounding box drawn around each hip, and each bounding box was manually labeled either (1) normal, (2) femoral neck fracture, (3) intertrochanteric fracture, or (4) subtrochanteric fracture. Next, a deep convolutional neural network YOLOv4-tiny AI model was trained using the augmented training set images, and then model performance was evaluated with the testing set images. Human doctors then evaluated the same testing set images, and the performances of the model and doctors were compared. The testing set contained no crossover data.

Results: The resulting output images revealed that the AI model produced bounding boxes around each hip region and classified the fracture and normal hip regions with a sensitivity of 96.2%, specificity of 94.6%, and an accuracy of 95%. The human doctors performed with a sensitivity ranging from 69.2 to 96.2%. Compared with human doctors, the detection rate sensitivity of the model was significantly better than a general practitioner and first-year residents and equivalent to specialist doctors.

Conclusions: This model showed hip fracture detection sensitivity comparable to well-trained radiologists and orthopedists and classified hip fractures highly accurately.

1. Introduction

Hip fractures are one of the most severe public health issues, particularly among the elderly. In the United States, more than 250,000 hip fractures occur each year [1]. The world population is aging. It is predicted that the number of people aged 60 and older will reach 2 billion by 2050, up from 900 million in 2015 [2]. A missed hip fracture diagnosis is so devastating and causes such morbidity and mortality that even a delayed diagnosis can worsen the outcome.

Clinical history, physical exams, and most importantly, hip or pelvic radiographs are used to diagnose a hip fracture. Accurate and immediate interpretation of the film requires specialized doctor experience and knowledge.

Missed diagnoses are common in rural areas, particularly in primary care settings, due to a lack of resources and consultation with specialists. The rate of misdiagnosis has been reported to be as high as 14% [1].

Artificial intelligence (AI) technology, especially computer vision, has been used successfully in medical imaging. In several previous studies [1, 3, 4], fracture detection accuracy was promising, with some AI models performing as well as a specialist doctor.

The AI model You-Only-Look-Once (YOLO) is a deep convolutional neural network (DCNN) that can perform image detection tasks (e.g., draw a bounding box around a fracture) and classification tasks (e.g., identify normal, femoral neck, intertrochanteric, and subtrochanteric fracture class types). It employs multi-layer image detection and applies single neural network algorithms to an entire image faster while

* Corresponding author.
E-mail address: princh@kku.ac.th (P. Twinprai).

https://doi.org/10.1016/j.heliyon.2022.e11266
Received 4 February 2022; Received in revised form 26 April 2022; Accepted 20 October 2022

consuming fewer resources than the former regional convolutional neural network (R–CNN). YOLO is a supervised learning model that produces high-probability outcomes, is superior to unsupervised machine learning, and uses fewer training images.

The focus of this research is to assess the diagnostic accuracy and sensitivity of a YOLO-based AI model for detecting hip fractures and distinguishing hip fracture types in hip and pelvic radiographs. According to our hypothesis, the performance of the model could be comparable to that of a well-trained radiologist or orthopedist.

This retrospective study was approved by the Khon Kaen University Ethics Committee for Human Research; approval No. HE641136.

2. Materials and Methods

2.1. Dataset

Hip and pelvic anterior-posterior (AP) radiographs of patients at least 18 years old between January 2015 to December 2020 were assembled from the picture archiving and communication system (PACS) of the Srinagarind Hospital. Srinagarind Hospital is a university teaching hospital, a Level 1 trauma center, capable of tertiary care, and affiliated with the Faculty of Medicine, Khon Kaen University, Thailand.

Radiographs with poor image quality, previous surgical fixation, or non-hip fracture diagnoses, such as bone metastasis or osteomyelitis, were removed by consensus of the P.T. (Radiologist) and N.T. (Orthopedist). Then, using the tools in the PACS system, each radiograph was carefully de-identified (patient name, age, sex, and radiograph date). Next, we converted each radiograph to a JPEG format with an image size of 1024 × 1024 pixels and a file size of around 90 kb while ensuring adequate windowing, contrast, and exposure.

Pelvic and hip x-ray images were reviewed from the recent year information backward in order to obtain the best quality image. Searching was performed until the desired dataset of 500 per classification were obtained. The normal class was obtained from images between December 2020 and April 2017 and the fracture class between December 2020 and January 2015. The difference in the length of time interval was due to the larger number of normal images than the fracture ones.

The dataset contained 1000 images: 500 images with a fracture and 500 images without a fracture (normal). All images shared the same proportions to ensure a consistent fit. Of the 500 fracture images, 235 were of a femoral neck fracture, 235 were of an intertrochanteric fracture, and 30 were of a subtrochanteric fracture. The images were divided into a training set consisting of 450 fracture and 450 normal images (900 images total) and a testing set consisting of 50 fracture and 50 normal images (100 images total). To avoid dataset crossover, images from the same patient were renamed and saved in the same folder.

2.2. Ground truth

Each of the 900 training set images was reviewed together by the consensus of a trauma orthopedist (NT) and an advanced diagnostic body imaging radiologist (PT). Complex cases having a subtle X-ray finding and where further CT or MRI was required (standard care in our hospital) were also reviewed. For cases where surgery was performed, intraoperative records and postoperative radiographs were evaluated for diagnosis confirmation.

Onto each image, for homogeneity, a bounding box was carefully, manually drawn in a rectangular shape with an annotation tool (Figure 1) by only one trauma orthopedist (NT) around each hip, extended to the subtrochanteric area, and labeled either [1] normal [2], femoral neck

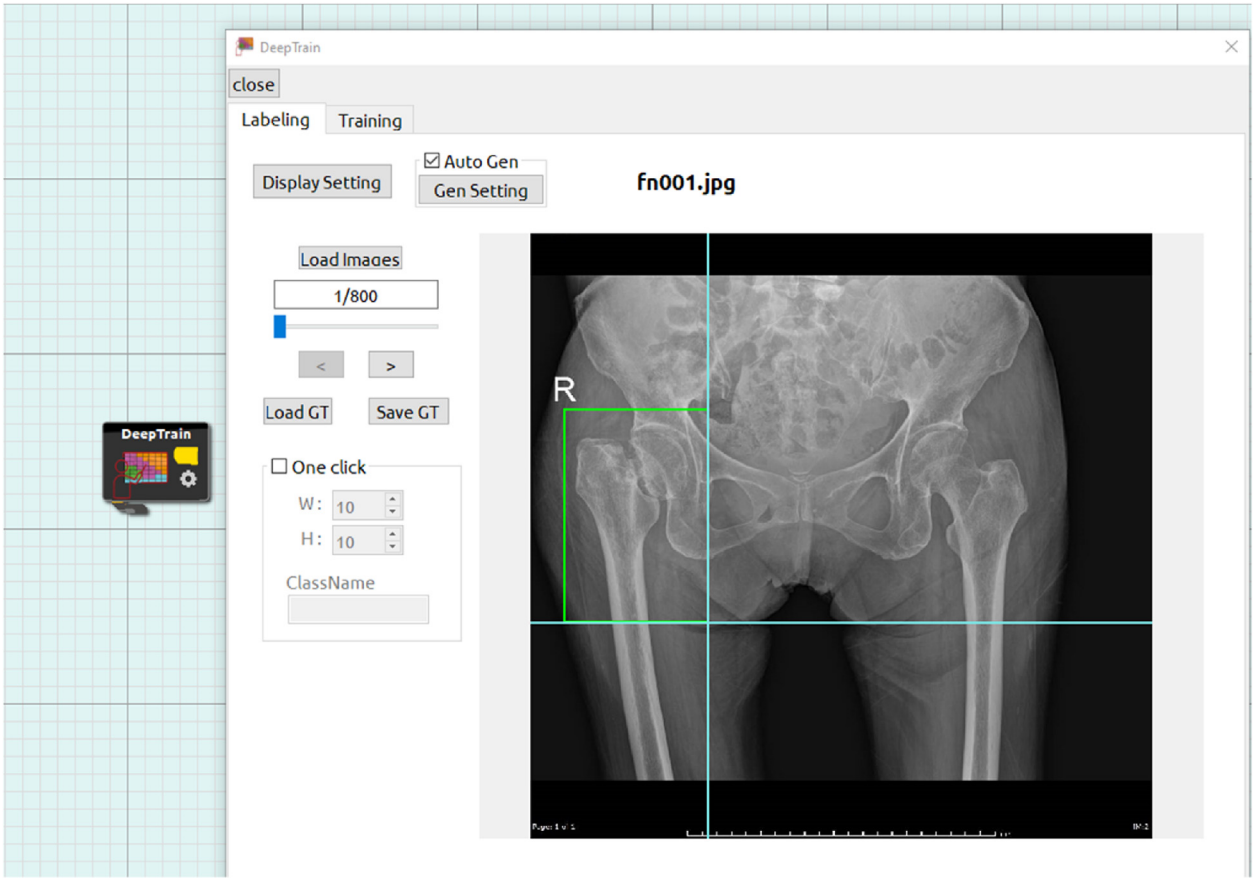

Figure 1. Annotation tool for image labeling (femoral neck fracture is shown).

fracture [3], intertrochanteric fracture, or [4] subtrochanteric fracture. Next, additional image augmentation was performed by adjusting rotation (−180°–180°) and contrast (−0.4 to 1.1). A computer program taking into account the augmentation parameters was then used to auto-generate 25,500 images, bringing the total number of ground truth images to 26,400.

2.3. Model training

This study used YOLO-v4-tiny, a state-of-the-art AI model that could detect objects with high accuracy and required reduced training time. The deep learning software was "CiRA CORE", a platform based on DARKNET framework constructed by government University (Please see Acknowledgement).

All 26,400 ground truth training images were loaded into the model, classifying each fractured or normal hip detected within the images. We then trained the model with 60 epochs until the loss function was 0.1.

2.4. Model testing

The pelvic X-ray picture test set (100 films, 200 hips) was loaded into the AI model. The test set contained 100 images: 50 images with a fracture and 50 images without a fracture (normal). Of the 50 fracture images, 23 were of a femoral neck fracture, 23 were of an intertrochanteric fracture, and 4 were of a subtrochanteric fracture.

The model then detected and classified the images continuously, immediately displaying each result, consisting of the automatically-labeled bounding boxes at the hip regions and the percent of confidence, as shown in Figures 2, 3, 4, 5, 6, and 7 below.

2.5. Model evaluation and statistical analysis

The training process was run on Microsoft Windows 10 Pro on a quad-core Intel(R) Core i7-7700 @ 3.6 GHZ processor having eight logical processors, and an NVIDIA GeForce GTX 1050 Ti GPU.

2.6. Statistical analysis

Stata 14 (StataCorp LLC) was used to compute model accuracy, sensitivity, specificity, precision, and F1 score. The performances of the model and human doctors were compared with a McNemar's test with a significance level of $p < 0.05$.

2.7. Diagnostic performance evaluation of the physicians (human doctors)

To appropriately compare the hip fracture diagnosis and classification performance of the model to human doctors, a test set was created for the doctors identical to the one used for the model.

The human doctors included [1]: one attending orthopedist [2], one attending radiologist [3], one chief orthopedic resident [4], one chief radiologist resident [5], one first-year orthopedic resident [6], one first-year radiologist resident, and [7] one general practitioner.

The test set was randomly arranged using a randomizer website (https://www.randomizer.org), and a web-based questionnaire with multiple-choice answers was created with Google Forms.

3. Results

3.1. Patient demographics

The 1000 patients corresponding to the radiographs selected for this study consisted of 367 males and 633 females, as shown in Table 1. The mean age of patients with a hip fracture was 68.54 years, significantly higher than the total mean age of 60.73 years. In normal and hip fracture cases, females outnumbered males.

3.2. Bounding box detection

With our AI model, bounding boxes accurately indicated the hip regions in every test image with two bounding boxes per image. Simultaneously, each bounding box correctly predicted the classification as

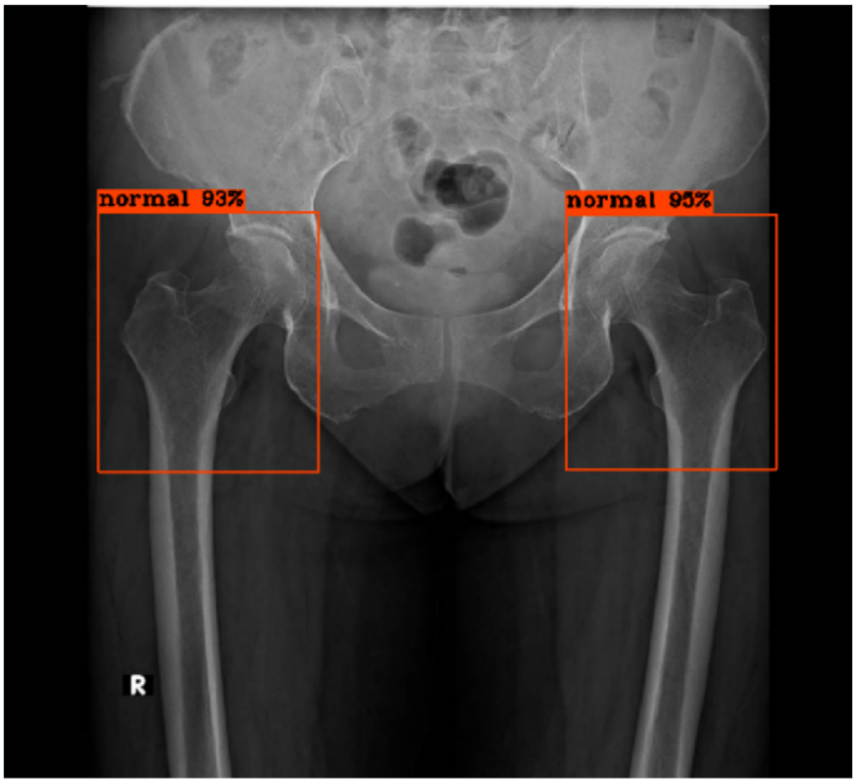

Figure 2. No fracture of bilateral hips, true negative results from the model.

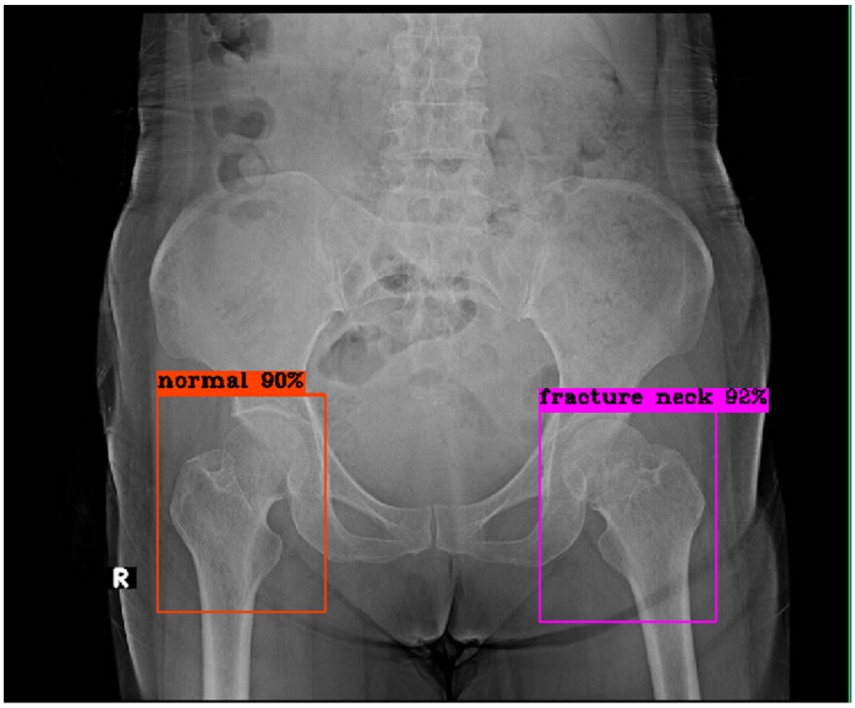

Figure 3. Right hip: No fracture, true negative result; Left hip: Fracture femoral neck, true positive result.

normal, femoral neck fracture, intertrochanteric fracture, or subtrochanteric fracture.

3.3. Model performance

The deep learning CNN model had a sensitivity of 96.2% (86.8–99.5%) and a specificity of 94.6% (89.6–97.6%) in distinguishing between fractured and normal hips, as shown in Table 2. The model performed with an F1 score of 0.909, and accuracy of 0.950, and a precision of 0.862, as shown in Table 3.

3.4. Multi-class detection

The methods of the deep learning CNN showed that the diagnostic performance was 94% accurate for identifying femoral neck fractures, 99% for intertrochanteric fractures, and 100% for subtrochanteric

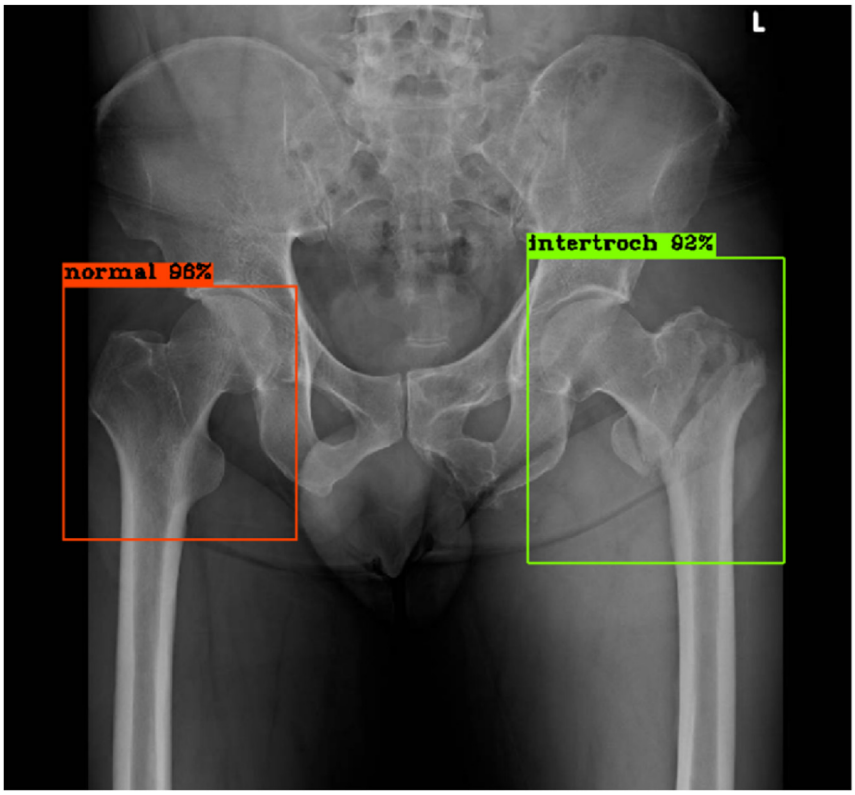

Figure 4. Right hip: No fracture, true negative result; Left hip: Intertrochanteric fracture, true positive result.

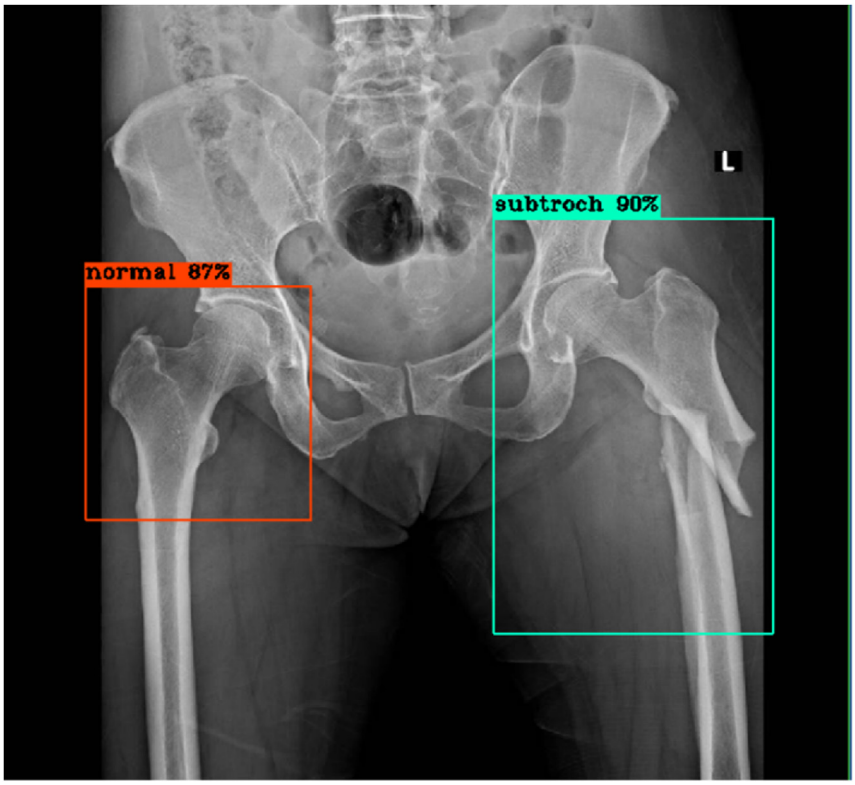

Figure 5. Right hip: No fracture, true negative result; Left hip: Subtrochanteric fracture, true positive result.

fractures, as shown in Table 4. In calculating the type-specific performance, other positive findings not consistent with the subgroup were treated as negative.

3.5. Comparison between CNN and human doctors

Seven human doctors completed the web-based questionnaire. Table 5 displays the diagnostic performance of the human doctors, and Table 6 shows the sensitivity comparison between the model and human doctors using a McNemar's test. The model performance was comparable to those of the attending and chief residents of radiology and orthopedics, with no statistically significant difference. On the other hand, with statistical significance, the model outperformed first-year orthopedic and radiology residents and general practitioners.

4. Discussion

The results showed that without coding competency and with only the YOLO-v4-tiny algorithm, the model detected hip fractures with high

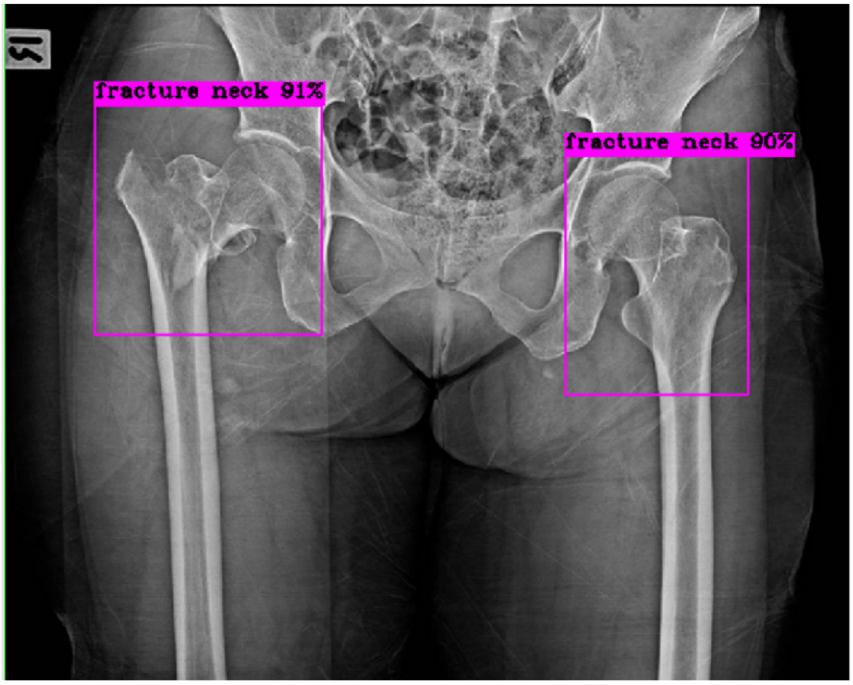

Figure 6. Right hip: Intertrochanteric fracture, true positive result, but fracture type misclassification; Left hip: No fracture, false positive result.

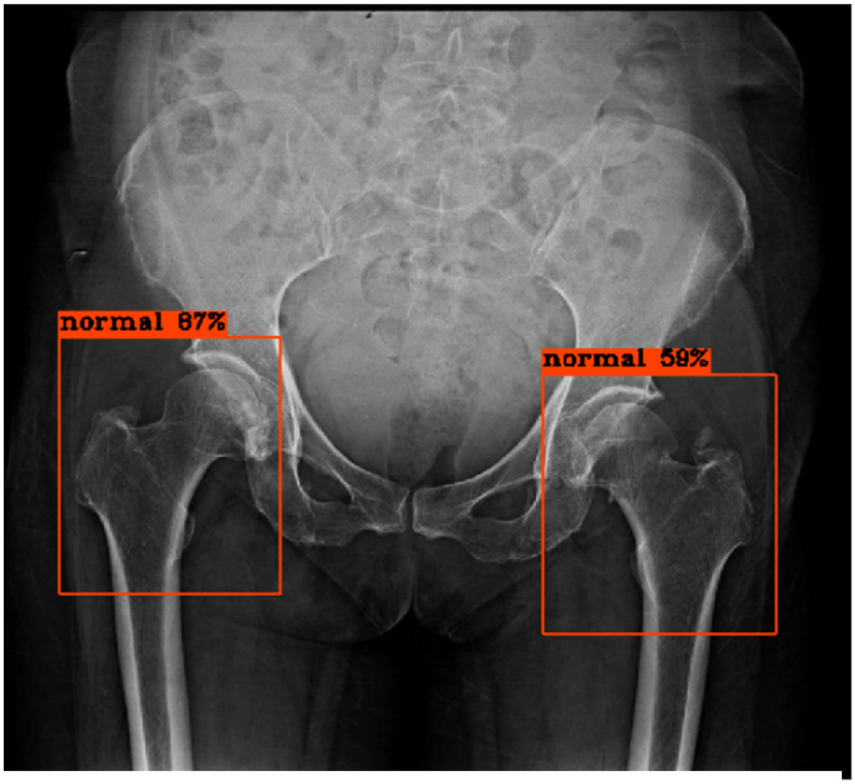

Figure 7. Right hip: No fracture, true negative result; Left hip: Femoral neck fracture, false negative result.

Table 1. Demographics of the pelvic and hip radiographic set.

Factor	Hip Fracture	Normal	Total
Number of patients	500	500	1000
Mean age, years (SD)	68.54 (19.15)	54.28 (22.71)	60.73 (21.22)
Number of males (%)	139 (27.8)	228 (45.6)	367 (36.7)
Number of females (%)	361 (72.2)	272 (54.4)	633 (63.3)

Table 2. Model sensitivity and specificity.

			Reality		Total
			Fracture	Normal	
Diagnostic Test	Fracture	Count	50	8	58
		% within Reality	**96.20%**	5.40%	29.00%
	Normal	Count	2	140	142
		% within Reality	3.80%	**94.60%**	71.00%
Total		Count	52	148	200
		% within Reality	100.00%	100.00%	100.00%

sensitivity (96.2%) and specificity (94.6%). In addition, the model classified fracture types highly accurately.

Compared to human doctors, the model outperformed the first-year residents in orthopedics and radiology and general practitioners with statistical significance. The model sensitivity is comparable to the attending physician and chief residents in radiography and orthopedics with no statistical difference. This finding confirmed that the model performed similarly to a well-trained radiologist or orthopedist.

Previous successful hip fracture detection studies include the models of Cheng [2] and Krogue JD [3], which used DenseNet-121 for a sensitivity of 98% and 93.2%, respectively, and an accuracy of 91% and 93.7%, respectively. Lee [6] successfully used the meta-learning deep neural network GoogLeNet (Inception v3) to classify femoral fractures in pelvic radiographs. Adams [7] compared AlexNet and GoogLeNet for femoral neck fracture detection with an accuracy of 88.1% and 89.4%, respectively. Gale [8] used DenseNet to predict hip fractures with 97% accuracy and 99% precision. We are not aware of any previous study that used the YOLO model for hip fracture detection.

Table 3. Model diagnostic performance.

Measure	Value
F1 Score	0.909
Accuracy	0.950
Precision	0.862

Table 4. Multi-class performance of the model for each classification subtype.

Category	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)
Femoral neck fracture	93.3 (77.9–99.2)	94.1 (89.4–97.1)
Intertrochanteric fracture	90.5 (69.6–98.8)	100 (98–100)
Subtrochanteric fracture	100 (29.2–100)	100 (98.1–100)

Table 5. Diagnostic performance of human doctors.

	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)
Doctor, GP	69.2 (54.9–81.3)	96.6 (92.3–98.9)
Resident 1st, Ortho	73.1 (59–84.4)	98 (94.2–99.6)
Resident 1st, Radio	76.9 (63.2–87.5)	98 (94.2–99.6)
Chief resident, Ortho	96.2 (86.8–99.5)	97.3 (93.2–99.3)
Chief resident, Radio	92.3 (81.5–97.9)	93.9 (88.8–97.2)
Attending, Ortho	96.2 (86.8–99.5)	97.3 (93.2–99.3)
Attending, Radio	96.2 (86.8–99.5)	95.3 (90.5–98.1)

Table 6. Statistical comparison of sensitivities between the model and human doctors using a McNemar's test.

	Model vs doctor, GP	Model vs resident 1st, ortho	Model vs resident 1st, radio	Model vs chief resident, ortho	Model vs chief resident, radio	Model vs attending, ortho	Model vs attending, radio
p-Value	<0.001	<0.001	0.006	1.000	0.625	1.000	1.000

Our model is based on YOLO-v4-tiny. As implied by its name, You Only Look Once [9], YOLO allows for object detection and classification while processing the image only once. Previous successful YOLO tasks in medical imaging fields include detecting suspicious regions in mammogram images [10], identifying cholelithiasis, and classifying gallstones on CT images [11].

One of the most significant pain points regarding deep learning in the medical field is the black box problem. Machine learning can use various features to create algorithms for diagnosing, predicting, and forecasting outcomes without providing much information about the reasoning behind all these vectors [12, 13]. Some studies [4, 8, 14] attempted to overcome the black box problem by using cropped images to help the model see only essential features. Cheng [1] applied Grad-CAM to visualize the heatmap of the regions that the model saw and found images with incorrect activation sites: one site was at the wrong side, and the other was over the iliac bone.

The strength of our model was the output, correctly presented in all test images as bounding boxes at the hip areas. The certainty with which the model predicted the results at these regions helped reduce any uncertainty regarding the black box problem.

In a previous study, Yu's successful model [17] was 97% accurate but required initial cropping of the femoral neck images. In contrast, our model uses the entire routinely-acquired patient image as input, similar to Cheng's model [1]. In this way, our model is easy to apply and simple to use in a real-life scenario.

Another strength of this work was that the dataset was prepared with PACS images carefully acquired by a medical doctor specialist, who could directly review subsequent patient system images. Our dataset was more accurate than in studies where a doctor did not review the dataset.

The ground truth images were prepared as best as possible in our setting. All fracture cases were confirmed with postoperative films. All available hip CT and MRI results associated with the selected PACS radiographs were reviewed. It is believed that similar to human brain training, a good AI result can be obtained if a good dataset is provided. However, good dataset preparation must be traded for time. Hence, only 1000 X-ray images were retrieved for this study.

There are limitations to this work. Our dataset was collected exclusively at our institution, which limits its generalizability. Another drawback is the small dataset totaling 1000 images. In previous research, Adams [7] found that increasing the sample size improved model accuracy, with magnification playing only a minor improvement role.

For the false-positive cases, we carefully reexamined the images and found that the images had poor film positioning and lucent lesions, such as artifacts of skin folds in the buttock area and diaper folds, which might have resembled fracture lines. In addition, some of the images were taken with the hip externally rotated such that the short femoral neck resembled a femoral neck fracture. In the false-negative cases, they were nondisplaced fractures, and no fracture was detected. After reviewing the cases again with the medical records, the cases were highly suspicious for a fracture. Patients were followed up with either repeated film or further CT or MRI to confirm the presence of a fracture. For improving model accuracy and subsequent hip fracture diagnosis, the model should be further trained with these artifacts to ensure that the model can differentiate between an artifact and fracture.

With or without AI, when obtaining X-ray images, care should be taken to remove external artifacts, such as diapers, clothing, or other exterior items that may cause artifacts. In addition, positioning basics should be performed such that hip radiographs are taken with an internal hip rotation of 15–20° for adequately accessing the femoral neck [15].

Consider that even with careful inspection, the number of occult hip fractures can be as high as 10% [16]. If an occult hip fracture is suspected, further MRI or CT or close patient follow-up is recommended. In conjunction with clinical information, our model is a promising tool for reducing missed hip fracture diagnoses. Moreover, with the well-known radiologist shortage, most radiograph films have no associated report, and our model may also help diagnose hip fractures in those scenarios.

5. Conclusions

The YOLO model provides hip fracture detection sensitivity comparable to well-trained radiologists and orthopedists and high accuracy in hip fracture classification.

Application development with this model is a future goal. Using this model in conjunction with clinical data may assist the primary care physician to reduce hip fracture misdiagnosis, especially in rural or remote areas.

Declarations

Author contribution statement

Nattaphon Twinprai: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

Artit Boonrod, Arunnit Boonrod, Jarin Chindaprasirt, Wichien Sirithanaphol: Contributed reagents, materials, analysis tools or data.

Prinya Chindaprasirt: Analyzed and interpreted the data; Wrote the paper.

Prin Twinprai: Performed the experiments, Analyzed, and interpreted the data; Wrote the paper.

Funding statement

Arunnit Boonrod was supported by Khon Kaen University's Research and Graduate Studies [RP64-8-003].

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

This study is supported by research and graduated studies, Khon Kaen University. We also thank the College of Advanced Manufacturing Innovation, King Mongkut's Institute of Technology Ladkrabang for providing the deep learning platform, CiRA CORE and software to support the research project. We want to acknowledge Mr. Kevin McCracken for editing the manuscript via the Publication Clinic, Khon Kaen University, Thailand.

References

[1] C.T. Cheng, T.Y. Ho, T.Y. Lee, C.C. Chang, C.C. Chou, C.C. Chen, et al., Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs, Eur. Radiol. 29 (10) (2019) 5469–5477.

[2] WHO, Ageing and health [Internet]. https://www.who.int, 2018 [cited 2020 Jun]. Available from: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

[3] J.D. Krogue, K.V. Cheng, K.M. Hwang, P. Toogood, E.G. Meinberg, E.J. Geiger, et al., Automatic hip fracture identification and functional subclassification with deep learning, Radiol. Artif. Intell. 2 (2) (2020), e190023.

[4] T. Urakawa, Y. Tanaka, S. Goto, H. Matsuzawa, K. Watanabe, N. Endo, Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network, Skeletal Radiol. 48 (2) (2019) 239–244.

[5] V. Kittichai, T. Pengsakul, K. Chumchuen, Y. Samung, P. Sriwichai, N. Phatthamolrat, et al., Deep learning approaches for challenging species and gender identification of mosquito vectors, Sci. Rep. 11 (1) (2021) 4838.

[6] C. Lee, J. Jang, S. Lee, Y.S. Kim, H.J. Jo, Y. Kim, Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network, Sci. Rep. 10 (1) (2020).

[7] M. Adams, W. Chen, D. Holcdorf, M.W. McCusker, P.D. Howe, F. Gaillard, Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures, J. Med. Imaging Radiat. Oncol. 63 (1) (2019) 27–32.

[8] W. Gale, L. Oakden-Rayner, G. Carneiro, Lyle Andrew, Detecting Hip Fractures with Radiologist-Level Performance Using Deep Neural Networks, 2017 arXiv pre-print server.

[9] C.Y. Wang, A. Bochkovskiy, H.Y.M. Liao, Scaled-YOLOv4: Scaling Cross Stage Partial Network. https://arxiv.org/abs/2011.08036, 2020.

[10] R. Platania, S. Shams, S. Yang, J. Zhang, K. Lee, S.-J. Park (Eds.), Automated Breast Cancer Diagnosis Using Deep Learning and Region of Interest Detection (BC-DROID), ACM, 2017.

[11] S. Pang, T. Ding, S. Qiao, F. Meng, S. Wang, P. Li, et al., A novel YOLOv3-arch model for identifying cholelithiasis and classifying gallstones on CT images, PLoS One 14 (6) (2019), e0217647.

[12] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, et al., Artificial intelligence in healthcare: past, present and future, Stroke Vasc. Neurol. 2 (4) (2017) 230–243.

[13] A.I.F. Poon, J.J.Y. Sung, Opening the black box of AI-Medicine, J. Gastroenterol. Hepatol. 36 (3) (2021) 581–584.

[14] S.W. Chung, S.S. Han, J.W. Lee, K.S. Oh, N.R. Kim, J.P. Yoon, et al., Automated detection and classification of the proximal humerus fracture by using deep learning algorithm, Acta Orthop. 89 (4) (2018) 468–473.

[15] S.J. Lim, Y.S. Park, Plain radiography of the hip: a review of radiographic techniques and image features, Hip Pelvis 27 (3) (2015) 125–134.

[16] A. Pinto, D. Berritto, A. Russo, F. Riccitiello, M. Caruso, M.P. Belfiore, et al., Traumatic fractures in adults: missed diagnosis on plain radiographs in the Emergency Department, Acta Biomed. 89 (1–s) (2018) 111–123.

[17] J.S. Yu, S.M. Yu, B.S. Erdal, M. Demirer, V. Gupta, M. Bigelow, et al., Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept, Clin. Radiol. 75 (3) (2020), 237.e1–.e9.