ACTA ORTHOPAEDICA et TRAUMATOLOGICA TURCICA

www.aott.org.tr

Research Article

Comparison of artificial intelligence algorithm for the diagnosis of hip fracture on plain radiography with decision-making physicians: a validation study

Salih Beyaz[1], Sahika Betul Yayli[2], Ersin Kılıç[2], Kutay Kılıç[2]

[1]*Department of Orthopedics and Traumatology, Başkent University, Adana Turgut Noyan Research and Training Centre, Adana, Turkey*
[2]*Turkcell Technology, Artificial Intelligence & Digital Analytic Solutions, İstanbul, Turkey*

ARTICLE INFO

Article history:
Submitted March 23, 2023
Received in revised form August 25, 2023
Last revision received form November 15, 2023
Accepted December 11, 2023
Publication Date February 5, 2024

Keywords:
Artificial Intelligence
Validation Study
Hip Fracture
Plain Radiography
Anteroposterior Pelvis X-ray

ORCID iDs of the authors:
S.B. 0000-0002-5788-5116;
S.BY. 0000-0001-5057-8634;
E.K. 0000-0002-0924-9246;
K.K. 0000-0001-8884-6833.

ABSTRACT

Objective: This study aimed to compare an algorithm developed for diagnosing hip fractures on plain radiographs with the physicians involved in diagnosing hip fractures.

Methods: Radiographs labeled as fractured (n=182) and non-fractured (n=542) by an expert on proximal femur fractures were included in the study. General practitioners in the emergency department (n=3), emergency medicine (n=3), radiologists (n=3), orthopedic residents (n=3), and orthopedic surgeons (n=3) were included in the study as the labelers, who labeled the presence of fractures on the right and left sides of the proximal femoral region on each anteroposterior (AP) plain pelvis radiograph as fractured or non-fractured. In addition, all the radiographs were evaluated using an artificial intelligence (AI) algorithm consisting of 3 AI models and a majority voting technique. Each AI model evaluated each graph separately, and majority voting determined the final decision as the majority of the outputs of the 3 AI models. The results of the AI algorithm and labelling physicians included in the study were compared with the reference evaluation.

Results: Based on F-1 scores, here are the average scores of the group: majority voting (0.942) > orthopedic surgeon (0.938) > AI models (0.917) > orthopedic resident (0.858) > emergency medicine (0.758) > general practitioner (0.689) > radiologist (0.677).

Conclusion: The AI algorithm developed in our previous study may help recognize fractures in AP pelvis in plain radiography in the emergency department for non-orthopedist physicians.

Level of Evidence: Level IV, Diagnostic Study.

Introduction

Artificial intelligence (AI) applications are popular in orthopedics for diagnosis and classification from radiography images.[1,2] Based on the results of these studies, the algorithms developed achieve near-perfect results.[3] It is important to conduct human comparison studies in order to translate this theoretical success into real-life medical practice.[4,5]

Although the incidence of hip fractures varies due to geographical differences, it increases in parallel with the aging population, and medical costs are rapidly increasing.[6,7] In suspected hip fracture cases, anteroposterior (AP) pelvis and lateral hip radiographs are taken and evaluated by emergency department specialists, general practitioners, orthopedic residents, radiologists, and/or orthopedists. The rate of hip fractures that cannot be distinguished on plain radiographs is 2.7%. Such suspected fractures require imaging methods with higher diagnostic sensitivity and specificity, such as computed tomography (CT) and magnetic resonance imaging (MRI).[8] Delayed diagnosis of hip fractures increases mortality and morbidity and causes legal problems.[8]

This study's aim was to compare an algorithm developed for diagnosing hip fractures on plain radiographs with the physicians' involvement in diagnosing hip fractures.

Materials and methods

For this study, permission was obtained from the Başkent University Non-Interventional Clinical Research Ethics Committee dated October 12, 2021 and numbered KA 21/420. Verbal or written constant was not obtained from the patiens to use the pelvis X-ray. All of patients data on the X-ray image were deleted. X-ray images were anomymized. Research and ethics committee approval was received under this condition.

Development of an artificial intelligence algorithm

In a previous study, we developed a decision support system that automatically detects hip fractures on x-ray images of the pelvis AP.[9] A large multicenter dataset of 19,583 radiographs collected from 5 centres was created to achieve success at the clinical expert level. The radiographs of patients < 16 years, those with implants, and those with lateral hip view were excluded. In the previous study, 10,849 graphs were used for training, validation, and testing. The Xception, EfficientNet-B7, and NFNet-F3 model architectures were trained and tested among the most successful image classification methods in the literature. In the

Corresponding author:
Salih Beyaz
sbeyaz@baskent.edu.tr

Cite this article as: *Beyaz S, Yayli SB, Kılıç E, Kılıç K. 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; 10.5152/j.aott.2024.23065 [Epub Ahead of Print].*

DOI: 10.5152/j.aott.2024.23065

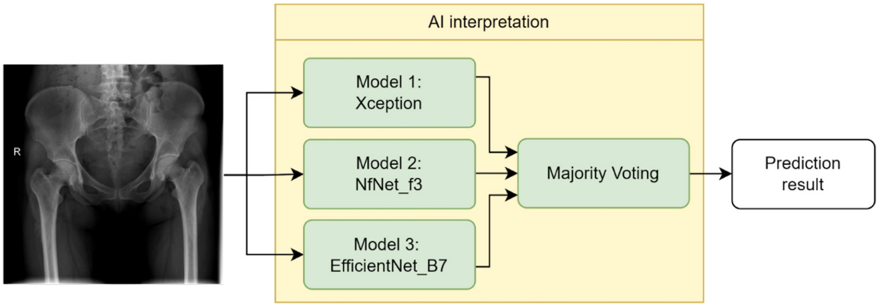

Figure 1 . Majority voting technique for final prediction.

test results, the Xception, EfficientNet-B7, and NfNet-F3 models' F1 scores were 0.89, 0.92, and 0.90, respectively. In the proposed system, the final decision-making system uses these classifiers' results to determine the final decision using a majority voting technique. The final system's F1 score was 0.93 (published data) (Figure 1).

Selection of test radiographs

Anteroposterior pelvis radiographs of 3,655 patients taken in the emergency departments of 5 different centers between 2015 and 2022 were included. A total of 678 radiographs of patients < 18 years and 48 radiographs of patients with an implant in at least 1 hip were excluded. The International Classification of Diseases 11th revision (ICD-11) codes of the remaining 2,929 patients were scanned: 333 patients diagnosed with proximal femur fracture (ICD-11 codes s72.0, s72.1, and s72.3) were labeled fractured, and the remaining 2,596 patients were labeled non-fractured (Figure 2). Of the 333 patients' radiographs labeled fractured according to CT results, 36 were non-displaced, and the radiographs of 297 patients were displaced.

Patient data (name, surname, application date, age, etc.) on the radiographs were deleted and anonymized. Radiographs converted from Digital Imaging and Communications in Medicine (DICOM) to JPEG format were recorded with a descriptive code number. Two board-certified orthopedics and traumatology specialists with ≥ 10 years of trauma experience labeled the radiographs as fractured or non-fractured in a reference evaluation process. All the results agreed upon by the 2 experts were included in the study. Radiographs with different diagnoses were excluded. The Kappa statistic of inter-observer correlation for fracture presence or absence was 0.95.

The fracture prevalence in the dataset used while developing the algorithm was 21%, and the minimum number required to be evaluated with a sensitivity of 90% and a precision of 5% was 659.[10] For this study, we included 724 AP pelvic radiographs, 10% more than the minimum number. Since the prevalence of fractures in the dataset used in the algorithm development was 21%, a minimum of 182 fractured and 542 non-fractured patients were randomly selected and included in the study. The demographic structure of the selected radiographs is given in Table 1.

HIGHLIGHTS

- Artificial intelligence applications are becoming more popular in orthopedics. This study aimed to compare the success of the physicianswith an artificial intelligence algorithm in recognizing hip fracture from direct x-ray.
- F1 scores in the success of hip fracture recognition from radiographs were determined as AI > orthopedic surgeon >orthopedic resident > emergency medicine > general practitioner > radiologist.
- Artificial intelligence-based hip fracture recognition algorithm can evaluate radiography as well as an orthopedic surgeon.

Identifying the labelers

Physicians involved in the final decision-making process were selected as labelers to evaluate the radiological images of patients who presented to the emergency department with hip pain. General practitioners (n=3) with at least 3 years of experience in emergency deperment, emergency medicine (n=3), radiologists (n=3), orthopedic residents (n=3) with at least 3rd year in residency, and orthopedists (n=3) participated. The other participants' identities were not disclosed to the participating clinicians, who performed the labeling on a single computer in the hospital and were not permitted to have mobile communication devices with them.

Labeling of radiographs

Open-source web-based Computer Vision Annotation Tool server version 2.0, core version 4.2.1, (CVAT, Intel, Santa Clara, CA, USA) was preferred for labeling the radiographs.[11] To ensure data security, the software was installed on a special server where the radiographs were uploaded so that users could easily access the labeling environment and the data were safe, as they were not uploaded to the cloud. A special username and password were provided for this labeling program, which each labeler could enter through the web browser. The labelers were asked to label the radiograph as fractured or non-fractured on the right and left sides in the proximal femoral region of the AP plain pelvis radiographs. The patients' demographic characteristics, including age, sex, comorbidity, and origin of the radiographs, were not shared with the labeler. After each labeling, the labeler was asked to save the labels. No more than 50 graphs were allowed to be labeled in a day.

At the AI algorithm stage, the radiographs were individually labeled by 3 different AI models. For each radiograph, at least a two-thirds agreement was sought to result in fractured/non-fractured as the final decision in a majority voting process (Figure 1).

Comparison of performances

The AI and labeling physicians' results were compared with the reference evaluation. For each labeler, the area under the curve (AUC), CI, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score were calculated (Table 2).[12,13]

Three different labelers from each specialty participated in this study. The results were compared with the reference evaluation, and the values (true positive/negative and false positive/negative) of the results of the labelers in the same specialization were determined for easier interpretation. The AUC, CI, sensitivity, specificity, PPV, NPV,

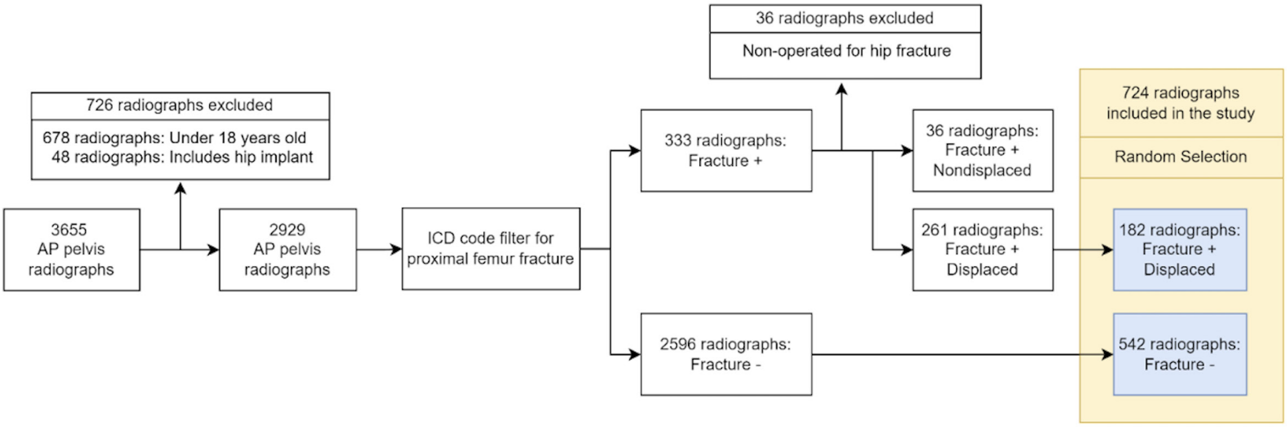
36 radiographs excluded
Non-operated for hip fracture
726 radiographs excluded
678 radiographs: Under 18 years old
48 radiographs: Includes hip implant
3655 AP pelvis radiographs
2929 AP pelvis radiographs
ICD code filter for proximal femur fracture
333 radiographs: Fracture +
36 radiographs: Fracture + Nondisplaced
261 radiographs: Fracture + Displaced
2596 radiographs: Fracture -
724 radiographs included in the study
Random Selection
182 radiographs: Fracture + Displaced
542 radiographs: Fracture -

Figure 2. Selection of test radiographs.

and F1 scores were calculated by averaging the results for each specialization group (Table 2). The F1 score shows the harmonic average of the sensitivity and specificity values and is the main evaluation metric of this study.[14]

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 score} = \frac{2 \times \text{Prescision} \times \text{Recall}}{\text{Precision} \times \text{Recall}}$$

TP = True Positive

FN = False Negative

Results

In Table 2, the results for each labeler are given in terms of their sensitivities, specificities, AUCs, and F1 scores. There is a comparison of the mean values of each expert group in Table 3.

Based on F-1 scores, here are the average scores of the group: majority voting (0.942) > orthopedic surgeon (0.938) > AI models (0.917) > orthopedic resident (0.858) > emergency medicine (0.758) > general practitioner (0.689) > radiologist (0.677).

Based on sensitivity scores, here are the average scores of groups: majority voting (0.970) > orthopedic surgeon (0.946) > AI models (0.916) > orthopedic resident (0.788) > emergency medicine (0.645) > general practitioner (0.549) > radiologist (0.528).

Based on specificity scores, here are the average scores of groups: orthopedic resident (0.943) > radiologist (0.942) > orthopedic surgeon (0.929) > general practitioner (0.925) > emergency medicine (0.919) > AI models (0.917) > majority voting (0.915).

Table 1 Demographic structure of patients

		Male	Female	Male + Female
All radiographs	Minimum age	17	19	17
	Maximum age	99	102	102
	Average age	64.64	68.33	66.88
	Count	286	438	724
Non-fractured radiographs	Minimum age	17	19	17
	Maximum age	99	94	99
	Average age	60.71	65.12	63.40
	Count	212	330	542
Fractured radiographs	Minimum age	31	46	31
	Maximum age	98	102	102
	Average age	76.63	79.38	78.28
	Count	89	93	182

Discussion

This is the first study to compare an AI-based fracture recognition algorithm with different clinicians involved in decision-making. We have taken an important step for active clinical use by comparing our success with previous data with clinicians. Two types of studies in the literature evaluate AI models' success. In the first type, a test dataset separate from the data is collected to develop a model. The second type of study evaluates changes in sensitivity and specificity when doctors use AI as decision support. Our study differs from the literature in that it compares the AI algorithm with the physicians involved in the decision-making process to diagnose hip fractures.

We developed an AI system for hip fracture diagnosis that provides high accuracy, sensitivity, and specificity.[9] According to the F1 scores, the majority voting results were the most successful compared to the specialist groups' averages and each AI model's average. For the first time, the capacity of emergency medicine physicians and radiologists to diagnose hip fractures correctly was revealed. The orthopedic surgeons and residents made the highest number of correct diagnoses after the AI algorithm, and the radiologists made the lowest (Table 2).

Many studies using AI algorithms in hip fractures have reported their high diagnostic value.[15-20] Adams et al. reported an AI accuracy of 90.6% by an algorithm to detect femoral neck fractures on radiographs.[15] Urakawa et al[16] used 3,346 hip radiographs (1,773 fractured and 1,573 non-fractured) to compare a CNN and 5 orthopedic surgeons, reporting a sensitivity of 93.9%, a specificity of 97.4%, and an AUC of 0.97.As a result, he suggested that algorithms can be used in cases where it is not possible to reach the orthopedic surgeon who will evaluate the radiographs. In a study of a deep CNN for the detection and localization of hip fractures on plain frontal pelvic radiographs, the algorithm detected hip fractures with 91% accuracy, 98% precision, a 2% false-negative rate, and an AUC of 0.98.[17] Yamada et al[18] evaluated diagnostic performance using AP and lateral hip radiographs to distinguish between CNNs and femoral neck fractures,

Table 2. Diagnostic performances of the emergency department specialist, general practitioner, radiologist, orthopedic surgeon, orthopedic resident, AI models and majority voting model

		AUC	95% CI	Sensitivity (%)	Specificity (%)	PPV	NPV	F1 score
Emergency medicine	1	0.712	0.678-0.745	0.513	0.911	0.778	0.756	0.656
	2	0.819	0.790-0.847	0.724	0.915	0.744	0.906	0.808
	3	0.844	0.815-0.869	0.757	0.931	0.794	0.915	0.835
General practitioner	1	0.659	0.623-0.694	0.396	0.922	0.856	0.568	0.554
	2	0.850	0.822-0.875	0.758	0.943	0.833	0.912	0.840
	3	0.772	0.740-0.802	0.638	0.907	0.733	0.862	0.749
Radiologist	1	0.756	0.723-0.787	0.580	0.932	0.822	0.803	0.715
	2	0.774	0.742-0.804	0.600	0.948	0.867	0.809	0.735
	3	0.694	0.659-0.727	0.757	0.931	0.794	0.915	0.835
Orthopedic surgeon	1	0.938	0.918-0.955	0.946	0.931	0.778	0.985	0.938
	2	0.953	0.935-0.968	0.979	0.928	0.767	0.994	0.953
	3	0.922	0.901-0.941	0.915	0.930	0.778	0.976	0.922
Orthopedic resident	1	0.941	0.921-0.957	0.937	0.945	0.828	0.982	0.941
	2	0.805	0.775-0.834	0.664	0.947	0.856	0.857	0.781
	3	0.877	0.850-0.900	0.816	0.938	0.811	0.939	0.872
AI models	1	0.937	0.917-0.954	0.957	0.918	0.733	0.989	0.937
	2	0.881	0.855-0.904	0.830	0.931	0.789	0.947	0.878
	3	0.947	0.928-0.962	0.992	0.902	0.672	0.998	0.945
Majority voting		0.943	0.923-0.958	0.970	0.915	0.722	0.993	0.942

AUC, area under the curve; AI, artificial intelligence; PPV, positive predictive value; NPV, negative predictive value.

Table 3. Average of diagnostic performances of emergency medicine, general practitioners, radiologists, orthopedic surgeons, orthopedic residents, and artificial intelligence

	Sensitivity (%)	Specificity (%)	PPV	NPV	F1 Score
Emergency medicine	0.645	0.919	0.426	0.859	0.758
General practitioner	0.549	0.925	0.807	0.781	0.689
Radiologist	0.528	0.942	0.861	0.746	0.677
Orthopedic surgeon	0.946	0.929	0.774	0.985	0.938
Orthopedic resident	0.788	0.943	0.831	0.926	0.858
AI	0.916	0.917	0.731	0.978	0.917
Majority voting	0.970	0.915	0.722	0.993	0.942

AI, artificial intelligence; NPV, negative predictive value; PPV, positive predictive value.

trochanteric fractures, and non-fractures using 1,703 straight hip AP radiographs and 1,220 straight hip lateral radiographs,[18] and the CNNs had a mean F1 score of 0.98. CNN accuracy has been reported as comparable or statistically significantly better than that of orthopedic surgeons, regardless of the radiographic image used.[18] In a study using 3,026 radiographs to detect hip fractures with deep learning, the CNN's sensitivity was 93.2%, the specificity was 94.2%, and the AUC was 0.98.[19] In a study using 3,605 radiographs, AI had a sensitivity of 98% and a specificity of 84%.[21] It has been suggested that this system can be used in emergency departments and that it improves physician's performance in diagnosing hip fractures.[21] In a study using 10,484 radiographs with the correct diagnosis basis established by the Gradient-weighted Class Activation Mapping technique, the accuracy, sensitivity, specificity, F-value, and AUC were 96.1%, 95.2%, 96.9%, 0.961, and 0.99, respectively.[20] In a controlled trial, clinicians' diagnostic accuracy improved significantly when they used a Computer-Aided Diagnosis system.[20] These studies' common feature is an AUC of between 0.97 and 0.99. Our study's results support these findings. AI models' primary purpose is to assist clinicians in decision-making rather than being direct decision-makers. Studies in which AI models that can diagnose bone fractures are compared with doctors are given in Table 4.

In our study, when the AI models were evaluated individually, the F1 scores of Model 1 and Model 2 were lower than the majority voting. The clinical equivalent of this situation is consultation. In emergency services, considering the patient density, workload, and long working hours, it is not always possible for physicians who examine radiographs to make a diagnosis as a joint decision by seeking a colleague's opinion. It is thought that using AI radiography as a decision support system will increase diagnostic efficiency while evaluating radiography in emergency services.[16]

Table 4. Comparative studies with health professionals of fracture identification algorithms from X-ray image developed based on artificial intelligence

Author (year)	Localization	Artificial intelligence training data	Test data	Artificial intelligence vs.
Gan K. et al (2019)	Femur neck	2,040	300 (150 fracture)	Orthopedist and radiologist
Blüthgen C. et al (2020)	Distal radius	524	100 (42 fracture	Radiologist
Hendrix N. et al (2022)	Scaphoid	3,000	219 (65 fracture)	Radiologist
Oppenheimer et al (2023)	Multiple	60,170	1,163 (367 fracture)	Radiologist
Urakawa T. et al (2019)	Intertrochanteric femur	3,346	334 (185 fracture)	Orthopedist
Tvinprai N. et al (2022)	Hip fracture	900	100 (50 fracture)	Orthopedist, radiologist, and residents
Guermazi A. et al (2022)	Multiple	60,170*	480 (240 fracture)	Orthopedist, radiologist, emergency medicine, family medicine

*Same datase.Oppenhaimer et all and Guermazi et all used

While developing AI-based algorithms, the dataset is split into training, validation, and test datasets. Success metrics (i.e., the sensitivity, specificity, accuracy, and F1 score obtained with the test dataset) indicate the algorithm's success. However, metrics can be obtained with the test dataset for reasons such as bias or lack of data diversity (i.e., a limited number of sources for the dataset is obtained, limited distribution of patient demographics in the dataset, etc.) and overfitting. This causes applications not to achieve the same success in daily practical use. Our study tested data from 5 different centers and devices with a method more suitable for daily practice.

The sensitivity of the AI algorithm we developed to diagnose hip fracture was 0.97, the specificity was 0.915, and the F1 score was 0.917. Compared with the average physician groups in the study, the AI algorithm was more successful than all the physician groups involved in diagnosing hip fractures. The algorithm we developed can be used when it is not possible to reach orthopedics or to support physicians working in the emergency department and interpreting radiographs. Radiological imaging is vital for hip fracture diagnosis.[8] It is not possible in the near future to leave the decision to AI to definitively diagnose patients by interpreting radiological images and revealing the treatment plan due to legal reasons and problems in explaining the working mechanism of these algorithms. Therefore, the developed applications should be used to assist physicians in decision-making or screening. When the workload is heavy and/or it is difficult or impossible to reach experienced physicians to interpret radiographs, it will help prevent an increase in morbidity and mortality due to delayed or missed diagnoses. In approximately 15% of hip fracture cases, the hip fracture is not displaced (non-displaced), and radiographic changes may be minimal.[22] In another 1% of cases, the fracture is not visible on plain radiographs, further diagnostic tests are needed, and MRI is the preferred method.[23] Therefore, using AI to diagnose hip fractures will be advantageous when it is difficult to diagnose (orthopedic surgeons are difficult to reach, take time, or in centers where MRI is unavailable). It is predicted that using AI as a decision support system will increase due to the frequency of hip fractures and increased morbidity and mortality, and serious legal problems when they are overlooked.[24]

More efficient use of AI in surgical decision-making could eliminate risk factors and human error.[24] AI technology's successful results in analyzing and interpreting data are increasing day by day compared to humans.[25] AI for future orthopedics and trauma surgery is much more than just fiction; optimized and personalised patient care is still a long way away.[26] The development and diffusion of AI is inevitable, as it reduces healthcare and administrative costs, improves medical efficiency, and predicts and prevents major disease complications.[27]

Another important finding from this study is that the performance of emergency specialists, general practitioners, and radiologists, who are the addressees of the subject, was evaluated for the first time in the diagnosis of hip fracture. The radiologists and general practitioners working in the emergency department had a lower F1 score. This is thought to be because these physicians work in a multidisciplinary manner with a high workload. This shows that AI can be used as an aid in emergency services.

Many studies aimed at diagnosing fractures have achieved a theoretical success rate of over 90%. The reflection of these dazzling success rates in daily practice is currently impossible currently due to data security and legal problems. The homogeneity of radiographs used during the development of algorithms and the fact that they are obtained from different centers and even from different countries will directly affect success in practical use. For this reason, we have made the data set and the 724 radiographs used in this study available so that researchers comply with the ethical rules. Decision support systems based on AI-based image processing will pave the way for each individual to reach an equal health system regardless of location, time, religion, language, or race.

This study's biggest limitation was that lateral hip radiographs were not included. The radiographs used while developing the algorithm were obtained from 5 devices in 5 centers. However, the study would be more reliable if the radiographs used in the development of the algorithm were obtained from different centers and even from different countries.

This study has shown that our AI algorithm is as successful as non-orthopedist physicians evaluating patients in the emergency department in recognizing fractures on AP pelvis plain radiography. Algorithms will help physicians make more decisions, although we still have a long way to go in final decision-making by AI alone.

Ethics Committee Approval: This study was approved by Research & Ethics committee of Başkent University (Approval No: KA21/420, Date: October 12,2021).

Informed Consent: Verbal or written constant was not obtained from the patiens to use the pelvis X-ray. All of patients data on the X-ray image were deleted. X-ray images were anomymized. Research and ethics committee approval was received under this condition.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – S.B.; Design – S.B., S.B.Y.; Supervision – S.B.; Resources – S.B., S.B.Y.; Materials – S.B., E.K.; Data Collection and – S.B., E.K.; Analysis and/or Interpretation – S.B., K.K.; Literature Search – S.B., S.B.Y.; Writing – S.B., S.B.Y., E.K., K.K.; Critical Review – S.B., S.B.Y., E.K., K.K.

Declaration of Interests: Associate Prof. Dr Salih Beyaz is a consultant to Turkcell Technology company on "Artificial Intelligence Applications in the Health Department." The other authors have no conflict of interest to declare.

Funding: This study was supported by Başkent University Reseach Fund and Turkcell Technology Artificial Intelligence & Digital Solutions.

References

1. Park CW, Oh SJ, Kim KS, et al. Artificial intelligence-based classification of bone tumors in the proximal femur on plain radiographs: system development and validation. *PLoS One*. 2022;17(2):e0264140. [CrossRef]
2. Ozkaya E, Topal FE, Bulut T, Gursoy M, Ozuysal M, Karakaya Z. Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. *Eur J Trauma Emerg Surg*. 2022;48(1):585-592. [CrossRef]
3. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop*. 2017;88(6):581-586. [CrossRef]
4. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol*. 2018;73(5):439-445. [CrossRef]
5. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018;89(4):468-473. [CrossRef]
6. Veronese N, Maggi S. Epidemiology and social costs of hip fracture. *Injury*. 2018;49(8):1458-1460. [CrossRef]
7. Zhang C, Feng J, Wang S, et al. Incidence of and trends in hip fracture among adults in urban China: a nationwide retrospective cohort study. *PLoS Med*. 2020;17(8):e1003180. [CrossRef]
8. Parker M, Johansen A. Hip fracture. *BMJ*. 2006;333(7557):27-30. [CrossRef]
9. Beyaz S, Yaylı SB, Kılıç E, Doktur U. The ensemble artificial intelligence (AI) method: detection of hip fractures in AP pelvis plain radiographs by majority voting using a multicenter dataset. *Digit Health*. 2023;9:20552076231216549. [CrossRef]
10. Wang X, Ji X. Sample size estimation in clinical research: from randomized controlled trials to observational studies. *Chest*. 2020;158(1S):S12-S20. [CrossRef]
11. Sekachev B, Manovich N, Zhiltsov M, et al. *opencv/cvat: v1. 1.0*. 2020. [CrossRef]
12. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*. 2013;4(2):627-635.

13. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799. [CrossRef]

14. Sasaki Y. *The truth of the f-measure*; 2007. https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

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

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

17. 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(10):5469-5477. [CrossRef]

18. Yamada Y, Maki S, Kishida S, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. *Acta Orthop*. 2020;91(6):699-704. [CrossRef]

19. Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. *Radiol Artif Intell*. 2020;2(2):e190023. [CrossRef]

20. Sato Y, Takegami Y, Asamoto T, et al. Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: a multicenter study. *BMC Musculoskelet Disord*. 2021;22(1):407. [CrossRef]

21. 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(11):e19416. [CrossRef]

22. Cameron ID, Handoll HH, Finnegan TP, Madhok R, Langhorne P. Co-ordinated multidisciplinary approaches for inpatient rehabilitation of older patients with proximal femoral fractures. *Cochrane Database Syst Rev*. 2001:CD000106.

23. Gillespie LD, Gillespie WJ, Robertson MC, Lamb SE, Cumming RG, Rowe BH. Interventions for preventing falls in elderly people. *Cochrane Database Syst Rev*. 2003:CD000340.

24. Beyaz S. A brief history of artificial intelligence and robotic surgery in orthopedics & traumatology and future expectations. *Jt Dis Relat Surg*. 2020;31(3):653-655. [CrossRef]

25. Beyaz S, Açıcı K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. *Jt Dis Relat Surg*. 2020;31(2):175-183. [CrossRef]

26. Tjardes T, Heller RA, Pförringer D, Lohmann R, Back DA, AG Digitalisierung der DGOU. Artificial intelligence in orthopedics and trauma surgery. *Chirurg*. 2020;91(3):201-205. [CrossRef]

27. Lorkowski J, Grzegorowska O, Pokorski M. Artificial intelligence in the healthcare system: an overview. *Adv Exp Med Biol*. 2021;1335:1-10. [CrossRef]