Eskisehir Medical Journal
Eskişehir Tıp Dergisi

Eskisehir Med J. 2025; 6(3): 304-310
doi: 10.48176/esmj.2025.224

Original Article / Araştırma Makalesi

A COMPARISON OF A DEEP LEARNING NEURAL NETWORKS MODEL WITH HUMAN OBSERVATIONS FOR DETECTING NON-OBVIOUS RADIUS AND CARPAL BONE FRACTURES

AŞİKAR OLMAYAN RADİUS VE KARPAL KEMİK KIRIKLARININ TESPİTİ İÇİN DERİN ÖĞRENME SİNİR AĞLARI MODELİNİN İNSAN GÖZLEMLERİYLE KARŞILAŞTIRILMASI

AYÇA KOCA[1], KAAN ORHAN[2], AHMET BURAK OGUZ[1], YUSUF KAHYA[3], ATİLLA HALİL ELHAN[4], MÜGE GÜNALP[1], HUA YAN[5], GENGFEI LIU[5], EMRE CAN ÇELEBİOĞLI[6], AYTEN KAYI CANGIR[3],

[1] Ankara University Faculty of Medicine, Department of Emergency Medicine, Ankara, Turkey
[2] Ankara University Faculty of Dentistry, Department of Dental and Maxillofacial Radiology, Ankara , Turkey
[3] Ankara University Faculty of Medicine, Department of Thoracic Surgery, Ankara, Turkey
[4] Ankara University Faculty of Medicine, Department of Biostatistics, Ankara Turkey
[5] Huiying Medical Technology Company Limited, Beijing, People's Republic of China
[6] Ankara University Faculty of Medicine, Department of Radiology, Ankara , Turkey

ABSTRACT

Introduction: Patients with hand and wrist trauma are frequently diagnosed in the emergency department. Deep learning algorithms could potentially become powerful tools to diagnose fractures from X-ray wrist images. This study aims to assess the diagnostic performance of a deep learning algorithm in detecting wrist fractures that are difficult to detect through radiographs.

Methods: This retrospective study included adult patients with hand/wrist trauma who undergo CT imaging. CT imaging of injured areas, interpreted by an expert radiologist were considered as "ground truth" (GT). There were 313 cases, a total of 121 fractures (82 radius, 39 carpal bones) were identified as GT from CT images. Using the algorithm, fracture detection procedure was performed on dataset of hand and wrist X-ray images. The same datasets were evaluated by four emergency medicine doctors. Diagnostic performances such as accuracy, area under curve, sensitivity, precision and F1 score were calculated. Agreement (Kappa coefficient (κ)) between GT, observers and deep learning algorithm was determined.

Results: The algorithm showed 69.6% accuracy, 57% sensitivity and 61.6% precision. Emergency medicine doctors showed better diagnostic performance with higher accuracy, sensitivity and precision and AUC values. The interobserver agreement among four EM doctors was moderate whereas the agreement with the algorithm was only fair.

Conclusions: The Deep learning algorithm demonstrated an accurate detection of fractures in wrist X-rays and it had capabilities that were comparable to those of emergency medicine physicians, but the algorithm mentioned needs to be further improved to produce better outcome.

Keywords: Deep learning, neural networks, fractures, carpal bones, radius fractures

ÖZET

Giriş: Acil serviste sıklıkla el ve bilek travması olan hastalara tanı konur. Derin öğrenme algoritmaları, X-ışını bilek görüntülerinden kırıkları teşhis etmek için güçlü araçlar haline gelebilir. Bu çalışma, radyografilerle tespit edilmesi zor olan bilek kırıklarını tespit etmede derin öğrenme algoritmasının tanı performansını değerlendirmeyi amaçlamaktadır.

Yöntemler: Bu retrospektif çalışma, BT görüntülemesi yapılan el/bilek travması olan yetişkin hastaları içermektedir. Uzman bir radyolog tarafından yorumlanan yaralı bölgelerin BT görüntüleri "temel gerçek" (TG) olarak kabul edildi. 313 vaka çalışmaya dahil edildi, toplam 121 kırık (82 radius 39 karpal kemik) BT görüntülerinden TG olarak tanımlandı. Algoritma kullanılarak, el ve bilek X-ışını görüntülerinden oluşan veri setinde kırık tespit prosedürü gerçekleştirildi. Aynı veri setleri dört acil tıp doktoru tarafından değerlendirildi. Doğruluk, eğri altında kalan alan, duyarlılık, kesinlik ve F1 skoru gibi tanı performansları hesaplandı. TG, gözlemciler ve derin öğrenme algoritması arasındaki uyum (Kappa katsayısı (κ)) belirlendi.

Bulgular: Algoritma %69,6 doğruluk, %57 duyarlılık ve %61,6 kesinlik gösterdi. Acil tıp doktorları daha yüksek doğruluk, duyarlılık ve kesinlik ve AUC değerleriyle daha iyi tanı performansı gösterdi. Dört acil tıp doktoru arasındaki gözlemciler arası uyum orta düzeydeyken algoritmayla uyum yalnızca orta düzeydeydi.

Sonuç: Derin öğrenme algoritması, bilek röntgenlerinde kırıkları doğru bir şekilde tespit etti ve acil tıp doktorlarınınkine benzer yeteneklere sahipti, ancak daha iyi sonuçlar elde etmek için bahsedilen algoritmanın daha da iyileştirilmesi gerekiyor.

Anahtar Kelimeler: Derin öğrenme, sinir ağları, kırıklar, karpal kemikler, Radius kırıkları

INTRODUCTION

The wrist, main functional joint involved in daily life activities is frequently exposed to traumatic injuries (1). Wrist trauma consist of distal radius, distal ulna and carpal bones fractures and represent 14 to 30 % of all traumas encountered in the emergency department (2-4). The anatomical complexity of the wrist may result in misdiagnosis or errors in interpretation. Radiographic evaluation plays an

Corresponding Author: Ayça Koca, Ankara University Faculty of Medicine, Department of Emergency Medicine, Ankara, Turkey
E-mail: aycakoca@hotmail.com
ORCID: 0000-0002-1546-3150

Submission Date: 14.05.2025 **Acception Date:** 07.08.2025
Cite as: Koca A, Orhan K, Oguz AB, et al. A comparison of a deep learning neural networks model with human observations for detecting non-obvious radius and carpal bone fractures. Eskisehir Med J. 2025; 6(3): x. doi: 10.48176/ esmj.2025.224.

Figure 1. Flowchart of the study population

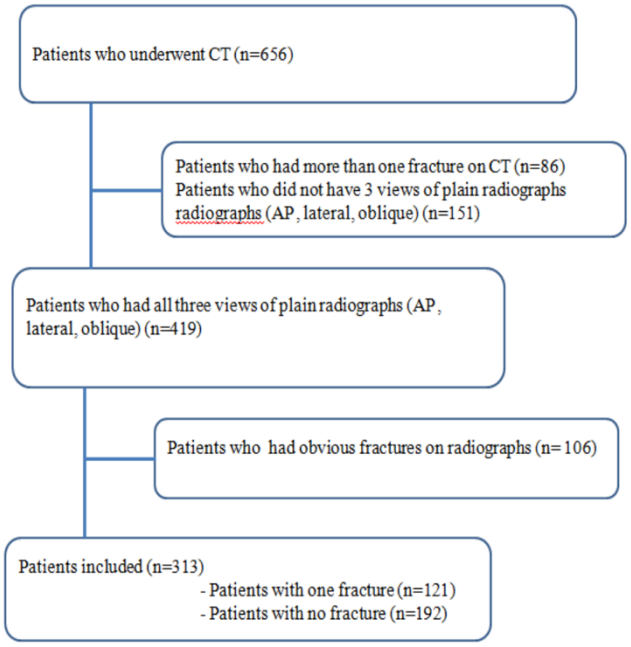


Abb. CT: computerized tomography, AP: antero-posterior

essential role in subsequent management of an injured wrist, serving as the standard imaging technique for fracture detection after trauma. However, wrist fractures can be overlooked using this modality. Furthermore, clinical signs may be subtle and both physical examination and standard X-ray inconclusive. When a wrist injury is initially evaluated, first, plain radiographs are typically ordered (5). Unrecognized fractures, often missed on initial radiographs may lead to complications such as malunion, nonunion, osteoarthritis and osteonecrosis, resulting in persistent pain and functional impairment (2). Moreover, in the busy setting of the emergency department (ED), failure to identify fractures is the most common diagnostic error (6). Conventional X-rays may be non-diagnostic and computed tomography (CT) may be required for prompt diagnosis, particularly when evaluating carpal bones. If initial finding are inconclusive and suspicion remains, a CT may be helpful for detecting occult or subtle fractures. However, CT is associated with additional costs, ionization radiation exposure and longer ED length of stay (7). In busy clinical settings where clinicians experience excessive workload, an accurate and efficient fracture detection method could assist and guide clinicians in avoiding misdiagnosis. In this context, deep learning algorithms (DLA) aim to facilitate clinician tasks and support clinical decision-making. Clinical studies have demonstrated the successful interpretation capabilities of DLA in various medical fields , including oncology and gastroenterology (8-11). In recent years, DLAs have achieved remarkable results in automatically detecting fractures in different body parts (4, 12). Nevertheless, DLA still encounter difficulties in identifying certain fractures such as scaphoid fractures, which are obvious to human observers(12).

This study aims to investigate the diagnostic performance of a DLA in detecting wrist bone fractures that are challenging to identify on antero-posterior and lateral plain radiographs. We then compare the diagnostic performance of the algorithm with that of emergency medicine (EM) doctors.

METHODS

Study design and population

This retrospective study was reviewed and approved by the Ankara University Research Ethics Review Board (approval number: 2021000367). The requirement for written informed consent was waived due to the retrospective nature of this study. Adult patients presenting with wrist injuries who underwent wrist or/and hand CT in the ED between 2019 and 2022 were reviewed. Patients with more than one fracture on CT, those who did not have three views of plain radiographs (antero-posterior, lateral, oblique) and those who had obvious fractures (displacement, fragmentation) on radiographs were excluded (Figure 1). Patient images were anonymized and stripped of any identifying clinical information.

Bone fracture Computer Aided Diagnosis (CAD) Model Pipeline (Initial dataset)

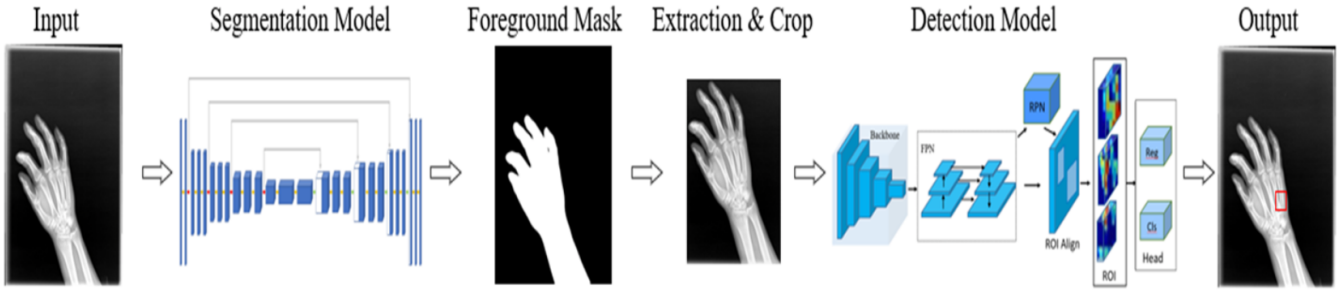
The proposed pipeline in this study consists of two models. The first model is a U-Net model, designed for foreground segmentation. In radius region images, background elements such as text annotations indicating the radius location may interfere fracture detection. Additionally, bilateral radius images may be appeared together, making it necessary to first isolate the foreground region (Figure 2).

U-Net is a convolutional neural network with the encoder-decoder structure. The encoder compresses the image into a low-resolution abstract representation, while the decoder reconstructs this encoded information to the original image resolution (13, 14). Skip connections between encoder and decode layers of the same resolution enable the integration of semantic information with precise spatial details. Finally, The model outputs the probability of each pixel belonging to the foreground or background.

The second model in the pipeline is Faster Region-based Convolutional Neural Network (Faster R-CNN)(15). Using the segmentation results, the foreground region is extracted from the original image and cropped to fill the field of view, and then resampled for model input. Feature extraction is performed by the Resnet-50 backbone, and multi-scale feature integration is achieved via a Feature Pyramid Network (FPN), which fuses spatial location information from lower layers with semantic information from higher layers (16). The Region Proposal Network (RPN) predicts candidate fracture bounding boxes based on these feature maps. Finally, proposals and corresponding features are passed through fully connected layers to classify fractures and refine bounding box coordinates for precise localization (Figure 3).

The U-Net model was trained on a dataset of 14,509 full-body digital radiography (DR) images for training and 1,196 images for validation. To enhance model robustness and generalization, a variety of data augmentation techniques were employed, including Random Gamma, Horizontal Flip, Random Brightness and Contrast, Elastic Transform, and Random Sized Crop. Key training parameters included an AdamW optimizer, a base learning rate of 3e-4, a weight decay of 0.0005, and a batch size of 16. The model was trained for 40,000 iterations with input images resized to 512x512 pixels. For the fracture detection task, the Faster R-CNN model was trained on a dedicated dataset of limb

Figure 2. The overview of Model architectıre



fractures. The training set consisted of 29,254 images (63.49% positive cases), and the validation set contained 2,037 images (50% positive cases). The data augmentation pipeline included Random Brightness, Random Contrast, Random Crop, Random Extent, Random Flip, Random Gamma, Random Saturation, and Random Lighting. The model was trained for 40,000 iterations using a Stochastic Gradient Descent (SGD) optimizer with a base learning rate of 0.001, momentum of 0.9, and a weight decay of 0.0001. A batch size of 16 was used, and input images were resized to have a minimum side of 800 pixels and a maximum side of 1333 pixels. During testing, a Non-Maximum Suppression (NMS) threshold of 0.1 was used.

Figure 3. Bone fracture Computer Aided Diagnosis- Model Pipeline a) U-Net Model for bone segmentation. B)Faster Region-based Convolutional Neural Network (Faster R-CNN), Faster R-CNN Model for fracture detection

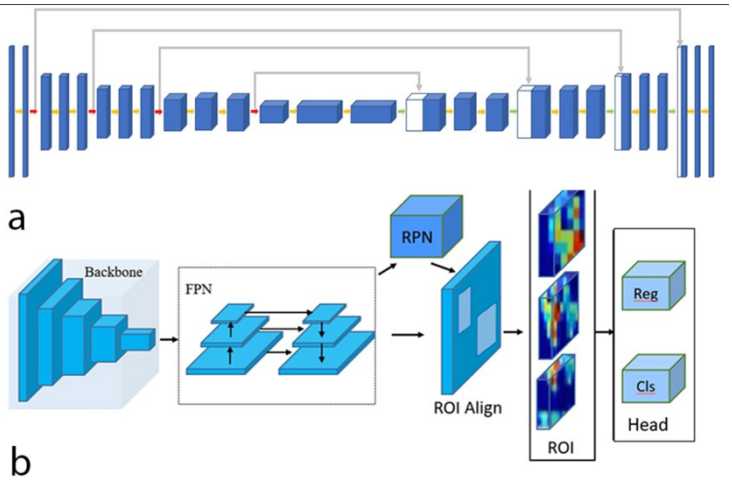


Abb. FPN: feature pyramid network, RPN: region proposal network, ROU: region of interest, Reg: regression, Cls: classification

Evaluation dataset and ground truth

In this study, the evaluation dataset comprised plain radiographs of patients over than 18 years of age presenting with hand injuries. The plain radiographs (antero-posterior, lateral and oblique) were retrieved from the Radiology Information System/Picture Archiving and Communication System (RIS/PACS; Centricity 5.0 RIS-i, GE Healthcare, Milwaukee, WI, USA) of Ankara University School of Medicine for fracture identification in order to identify the fractures. Only complete and adequate image series were included; —specifically, those consisting of all three standard views, free from plaster cast artifacts, and accompanied by corresponding CT images. All radiographs were provided in Digital Imaging and Communications in Medicine (DICOM) format and anonymized prior to evaluation. The dataset included only cases with a single fracture; patients with multiple or simultaneous fractures were excluded. CT images were obtained using a 64-slice scanner (Toshiba Aquilion 64, Otawara, Japan) with the following parameters: 0.5 mm detector collimation, 120 kVp tube voltage, 0.5 s gantry rotation time, 1 mm reconstructed section thickness, and 1 mm reconstruction intervals. The “ground truth” (GT) for each CT image was determined by a board-certified radiologist with eight years of experience. The radiologist was blinded to patients’ clinical conditions and prior CT reports. Four emergency medicine physicians (Expert 1, Expert 2, Expert 3, Expert 4), each with at least two years of clinical experience, independently analyzed the plain radiograph dataset. While they were aware that each patient had exactly one fracture to identify, they were blinded to the total number of fracture and non-fracture cases. Each reader annotated the fractures and their locations on the radiographs at individual workstations, without time constraints. The readers were blinded to the CT images, CT reports, and each other's assessments.

Evaluation method and model performance

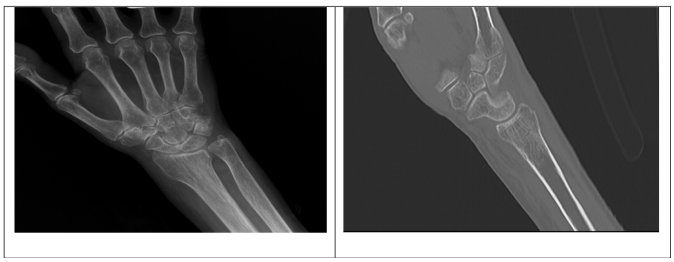
Each image was analyzed using the DLA algorithm. The findings from the GT and algorithm analysis for all cases were compared. The model performance was evaluated using the Confusion Matrix. The metrics used to assess the performance of the fracture detection model were as follows: True Positive (TP): the number of accurate detected fractures; False Positive (FP): the number of detected fractures even though there were no fractures; False Negative (FN): the number of fractures not detected. The performance metrics of the model were determined according to the formulas using TP, FP, and FN, as follows: Accuracy= (TP + TN) / (TP + TN + FP + FN); Sensitivity (Recall, True positive rate (TPR)= TP / (TP + FN); Precision (Positive predictive value (PPV))= TP / (TP + FP); F1 Score= 2TP / (2TP + FP + FN). From these formulas, accuracy, sensitivity, precision and F1 scores for all radius and carpal fractures were determined. F1 score is defined as the harmonic mean of sensitivity and precision.

Statistical Analysis

Continuous variables were reported with mean and standard deviation (SD) and categorical variables with frequencies and percentages (%). We assessed the diagnostic performance of DLA and the four experts using a receiver operating curve (ROC). ROC curves were used to

compare the diagnostics performance of DLA and the four experts. An area under curve (AUC) equal to 1.00 represents perfect classification, and an AUC of 0.5 indicates a prediction equal to chance. The difference between GT and DLA, and the four EM doctors, in terms of categorical variables, was tested by McNemar test. Cohen's kappa coefficient (κ) analysis was performed to determine the agreement between GT, DLA and the four EM doctors. Fleiss' generalized kappa coefficient was used to measure the degree of interobserver agreement among the four EM doctors. The degree of agreement was defined according to the value of the κ as follows: between 0.00 and 0.20 slight, between 0.21 and 0.40 fair, between 0.41 and 0.60 moderate, 0.61 and 0.80 substantial, and 0.81 to 1.00 almost perfect. A P-value less than 0.05 was considered significant.

Figure 4. Fracture of radius bone on plain radiographs and computerized tomography images



RESULTS

A total of 313 patients with wrist trauma who underwent CT were included. There were 155 female (49.5%) and 158 male (50.5%) patients. The mean age of the study population was 43.1±19.5 years old. After radiologic evaluation (GT), a total of 121 fractures was identified (38.7%), 82 patients were diagnosed with a radius fracture and 39 patients with a carpal bone fracture (Figure 4). The remaining 192 patients (61.3%) with no fractures were assigned to the control group According to DLA, 112 fractures were detected (35.8%) whereas no fracture was detected in 201 of the patients (64.2%). The final demographic and clinical characteristics of the patients and diagnosis based on GT and DLA are provided in Table 1.

Table 1. Characteristics of the study group

Gender, n (%)	Female		155 (49.5)
	Male		158 (50.5)
Age, (years old) Mean±SD			43.1±19.5
Fracture according to GT, n (%)	No		192 (61.3)
	Yes		121 (38.7)
		Radius fracture	82 (26.2)
		Carpal bone fracture	39 (12.5)
Fracture according to DLA, n (%)	No		201 (64.2)
	Yes		112 (35.8)

Abb. DLA: deep learning algorithm; GT: Ground truth; SD: Standard deviation

DLA results were compared with GT; 52 false negatives, 43 false positives and 149 true negatives and 69 true positives were determined. According to these results, the sensitivity of DLA was calculated as 57% and the precision as 61.6% in detecting fractures in hand and wrist traumas. The DLA had an AUC of 0.673±0.032 while the diagnostic performances of the four EM doctors were higher with AUC values of 0.822±0.027, 0.726±0.029, 0.807±0.027, 0.808±0.028, respectively (Figure 5). The DLA correctly detected 69 out of the total fractures while 43 of 192 patients confirmed as having no fracture were incorrectly diagnosed as having a fracture. The accuracy of the DLA was calculated as 69.6% (95% CI 64.2% to 74.5%). Accuracy, sensitivity, precision and F1 score favored the four EM doctors over DLA (Table 2).

Figure 5. The diagnostic performances of DLA and emergency medicine doctors in detecting hand and wrist traumas were compared with the area under the receiver operating curves (AUC)

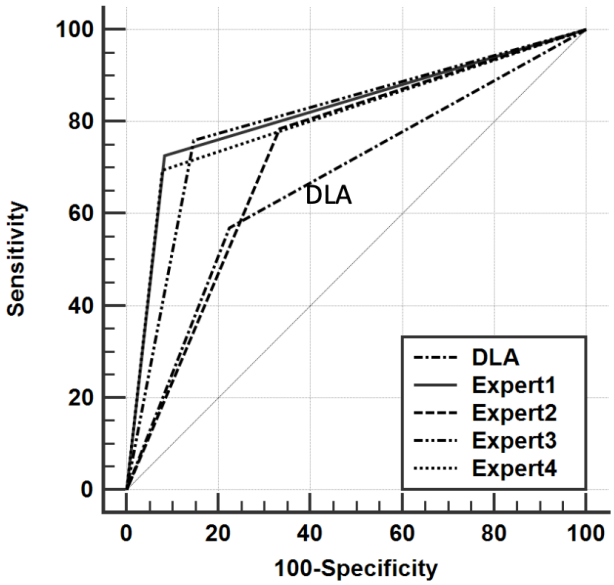


Abb. *DLA: deep learning algorithm*

The diagnostic performance of DLA was lower than that of expert 1, 3 and 4 (p<0.001 for all comparisons). There was no statistically significant difference between diagnostic performance of DLA and Expert 2 (p= 0.110) (upper diagonal of the Table 3). EM experts' agreements with each other were higher than with DLA (lower diagonal of the Table 3).

The generalized kappa coefficient among four experts was indicated substantial agreement with a κ= 0.624±0.023, (95% CI: 0.579-0.670, p<0.001). There was a significantly fair agreement between GT and DLA. All four experts showed a better agreement with GT than DLA. Expert 1, 3 and 4 demonstrated substantial agreement with DLA. A moderate agreement between GT and Expert 2 was detected (Table 4).

When considering accuracy according to anatomic sites of fractures, DLA and all four experts were better at diagnosing radius fractures than carpal ones (p<0.001) (Table 5).

For detecting radius fractures, there were statistically significant differences between DLA and the four experts (p= 0.004, p= 0.004, p= 0.027 and p= 0.008 respectively).

Table 2. Diagnostic performances of deep learning algorithm and four emergency medicine experts

		DLA		Expert 1		Expert 2		Expert 3		Expert 4	
		Normal	Fracture	Normal	Fracture	Normal	Fracture	Normal	Fracture	Normal	Fracture
GT	Normal	149	43	176	16	128	64	164	28	177	15
	Fracture	52	69	33	88	26	95	29	92	37	84
Total		201	112	209	104	154	159	193	120	214	99
p value		<0.001		<0.001		<0.001		<0.001		<0.001	
Accuracy (95% CI)		0.696 (0.642-0.745)		0.843 (0.798-0.880)		0.712 (0.659-0.760)		0.818 (0.771-0.857)		0.834 (0.788-0.872)	
AUC (95% CI)		0.673 (0.618-0.725) (p<0.001)		0.822 (0.775-0.863) (p<0.001)		0.726 (0.673-0.775) (p<0.001)		0.807 (0.759-0.849) (p<0.001)		0.808 (0.760-0.850) (p<0.001)	
Sensitivity (95% CI)		0.570 (0.481-0.655)		0.727 (0.642-0.799)		0.785 (0.704-0.849)		0.760 (0.677- 0.828)		0.694 (0.607-0.769)	
Precision (95% CI)		0.616 (0.524-0.701)		0.846 (0.765-0.903)		0.597 (0.520-0.671)		0.767 (0.683-0.833)		0.848 (0.765-0.906)	
F1 Score (95% CI)		0.592 (0.536-0.646)		0.782(0.732-0.825)		0.679(0.625-0.729)		0.763(0.712-0.708)		0.764(0.713-0.808)	

Abb. DLA: deep learning algorithm; GT: Ground truth

Significance with DLA versus Expert 1, 2, 3, and 4 was seen for detecting carpal fractures (p= 0.125, p= 0.008, p= 0.002, and p= 0.648, respectively), total fractures (p= 0.001, p<0.001, p<0.001, and p= 0.020, respectively) and normal radiographs (p<0.001, p= 0.020, p= 0.063, and p<0.001 respectively).

Table 3. Pairwise comparison of ROC curves (upper diagonal) and inter-observer agreement for evaluations (lower diagonal)

	DLA	Expert 1	Expert 2	Expert 3	Expert 4
DLA		0.149±0.028 p<0.001	0.053±0.033 p=0.110	0.134±0.030 p<0.001	0.135±0.030 p<0.001
Expert 1	0.449±0.053 p<0.001		0.096±0.021 p<0.001	0.015±0.018 p=0.417	0.014±0.023 p=0.552
Expert 2	0.268±0.052 p<0.001	0.625±0.041 p<0.001		0.081±0.023 p=0.001	0.082±0.027 p=0.003
Expert 3	0.370±0.054 p<0.001	0.806±0.035 p<0.001	0.637±0.042 p<0.001		0.0008±0.024 p=0.973
Expert 4	0.379±0.055 p<0.001	0.657±0.046 p<0.001	0.428±0.047 p<0.001	0.630±0.046 p<0.001	

Abb. DLA: deep learning algorithm; kappa coefficients (κ±Standard Error) were given at the lower diagonal of the table; ΔArea Under the Curve±Standard Error were given at the upper diagonal of the table,

DISCUSSION

In emergency medicine settings, the high volume of patients and substantial workload, often necessitate rapid evaluation of patients and X-rays. The lack of dedicated time for reading radiographs should not lead doctors to misinterpretation of results, as misdiagnosis of musculoskeletal injuries can have undesirable outcomes such as disability, restriction of range of motion and similar complications (17, 18). Therefore, a deep learning algorithm designed to interpret plain radiographs for orthopedic injuries could potentially improve diagnostic accuracy for extremity injuries and enhance patient management. In this retrospective study of 313 patients, we assessed the diagnostic performance of DLA and four emergency medicine doctors in detecting wrist bone fractures. Our findings indicate that the diagnostic ability of DLA was not as accurate as that of the EM doctors with the DLA's sensitivity (57%) being lower than that of the EM physicians (72.7%, 78.5%, 76%, 69.4%) in detecting wrist bone fractures. Furthermore, our DLA exhibited lower accuracy, sensitivity and specificity compared to the four EM doctors, as indicated by its lowest AUC. These results suggest that the DLA may not be sufficient to replace human doctors. However, the potential for artificial intelligence-assisted diagnosis lies not in replacement, but in conjunction with physicians to enhance their capabilities and facilitate clinical integrations. An algorithm, despite its stand-alone limitations observed in this study, could be highly useful as a triage tool to prompt further imaging when suspicion persists, particularly in busy settings like emergency departments.

Table 4. Agreement between ground truth, deep learning algorithm and emergency medicine experts

	Kappa (versus GT) (κ±Standard Error)	95% Confidence Interval	p value
DLA	0.351±0.054	0.245-0.458	<0.001
Expert 1	0.661±0.044	0.575-0.747	<0.001
Expert 2	0.427±0.050	0.330-0.524	<0.001
Expert 3	0.615±0.046	0.526-0.705	<0.001
Expert 4	0.638±0.045	0.549-0.726	<0.001

Abb. DLA: deep learning algorithm

Table 5. Diagnostic accuracy of DLA and emergency medicine experts

	Radius bone n (%)		Carpal bone n (%)		Fracture n (%)		Normal n (%)		Total n (%)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
DLA	60 (73.2)	22 (26.8)	9 (23.1)	30 (76.9)	69 (57.0)	52 (43.0)	149 (77.6)	43 (22.4)	218 (69.6)	95 (30.4)
Expert 1	74 (90.2)	8 (9.8)	14 (35.9)	25 (64.1)	88 (72.7)	33 (27.3)	176 (91.7)	16 (8.3)	264 (84.3)	49 (15.7)
Expert 2	74 (90.2)	8 (9.8)	21 (53.8)	18 (46.2)	95 (78.5)	26 (21.5)	128 (66.7)	64 (33.3)	223 (71.2)	90 (28.8)
Expert 3	71 (86.6)	11 (13.4)	21 (53.8)	18 (46.2)	92 (76.0)	29 (24.0)	164 (85.4)	28 (14.6)	256 (81.8)	57 (18.2)
Expert 4	72 (87.8)	10 (12.2)	12 (30.8)	27 (69.2)	84 (69.4)	37 (30.6)	177 (92.2)	15 (7.8)	261 (83.4)	52 (16.6)

Abb. *DLA: deep learning algorithm*

Previous studies have shown that artificial intelligence (AI) led to an improvement in the diagnostic performance of both radiologists and EM physicians when evaluating trauma radiographs [19, 20]. Nehrer et al. also showed that the accurate diagnosis of knee osteoarthritis could be ameliorated with the aid of AI, especially when compared to stand-alone AI or physician-alone performances (21). The integration of AI as a decision support tool can help clinicians avoid misdiagnosis in settings with extensive workload acting as an accurate and efficient fracture detection method. Our study found that the DLA detected 112 fractures, while 121 fractures were identified according to GT. These findings indicate that DLA underestimated the number of fractures.

Similar to our results, Langerhuizen et al. demonstrated the failure of a DLA in identifying hand fractures on radiographs; showing lower accuracy, sensitivity and specificity compared to surgeons (12). In our study, the performance of DLA and human observers differed significantly depending on anatomical area. Both DLA and EM experts were more effective at detecting radius bone fractures than carpal ones. Similarly, in Cohen et al.'s study, among all missed fractures, AI missed radius fractures (38%) less frequently than carpal ones (58%)(22). It is widely acknowledged that diagnosing carpal bone fractures using plain radiographs is challenging for physicians due to their complex anatomical structures (23). Like human readers, the DLA encountered some difficulties analyzing carpal bone fractures suggesting that future algorithms should be specially trained to recognize these challenging injuries. Despite these challenges, the collaborative use of AI with physicians holds significant promise

For instance, Nguyen et al demonstrated that AI assistance significantly increased the detection of non-obvious and difficult pediatric fractures, such as Salter or greenstick fractures, by 14.32 points, even though their stand-alone AI performed better than human observers in that specific context. This highlights AI's potential to complement human expertise, particularly in identifying subtle findings that might otherwise be overlooked (24).

For successful clinical integration, several factors must be considered. Firstly, AI programs for fracture detection have the potential to reduce the burden on EM doctors, leading to benefits such as decreased emergency department length of stay, fewer unnecessary CT orders, and reduced costs. (25,26).

This highlights the efficiency gains AI can bring. Secondly, AI can serve as a valuable training tool in the era of digitalization. It is crucial for both junior and senior medical doctors to familiarize themselves with the clinical practice of these technologies. However, it is paramount that junior doctors with less experience do not base all diagnostic decisions solely on AI, emphasizing the need for physician oversight and critical evaluation (27). One of the strengths of our study is the establishment of GT through CT images whicih provides a more definitive reference compaed to many studies that rely on plain radiographs interpreted by expert radiologists or orthopedic surgeons (24, 28). This robust GT is essential for accurately evaluating AI performance.

Our study had some limitations. First, the retrospective nature of this study did not allow the readers to examine the patient and focus on a specific anatomical area. Examining the patient's injured area before viewing the radiograph improves the final diagnostic decision (20, 29). Similarly, incorporating demographics or injury signs may improve the algorithm performance. As the study population was restricted to one single institution with a small number of patients, having larger cohorts might have enhanced the results.

Transformative change in medical practice occurs over time; however, AI, as one of the most important technologies ever developed in recent decades, must ensure safety and benefit for the long-term future. It is generally believed that AI tools will primarily facilitate and enhance human work, rather than replace the work of physicians and other healthcare staff. While AI in emergency radiology is still in its early stages and necessitates investment in research and development, particularly for subtle findings like wrist fractures where training data can be challenging to acquire, the synergistic potential of AI and human expertise is undeniable.

AI's application in orthopedic imaging can integrate past clinical experiences and a vast amount of knowledge to guide physicians in making more accurate diagnostic and therapeutic decisions, thus making the diagnosis and treatment of orthopedic diseases more efficient, standardized, and automated (30).

Continued research, innovation, and interdisciplinary collaboration are important to unlock the full potential of AI in healthcare. With successful integration, AI is anticipated to revolutionize healthcare, leading to improved patient outcomes, enhanced efficiency, and better access to personalized treatment and quality care (31).

CONCLUSION

In conclusion, our study demonstrated that while the DLA accurately detected radius fractures in wrist X-rays with capabilities comparable to emergency medicine physicians, the algorithm requires further improvement, especially for carpal bone injuries. The findings underscore that AI-assisted diagnosis, when integrated thoughtfully, can serve as a valuable adjunct to physician expertise, particularly in busy emergency settings, by enhancing efficiency, providing decision support, and acting as a robust triage tool. Continued research and development are crucial to refine AI algorithms, ensuring their reliability, and seamless integration into clinical workflows to ultimately improve patient outcomes through a collaborative human-AI approach.

Ethics Committee Approval: This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics committee of Ankara University School of Medicine (Ethics approval number: 2021000367).

Informed Consent: The requirement for written informed consent was waived due to the retrospective nature of this study.

Authorship Contributions: All authors contributed to the study conception and design. Methodology: Ayça Koca, Ahmet Burak Oğuz Formal analysis and investigation: Ayça Koca, Atilla Halil Elhan, Hua Yan, Pengfei Liu, Ahmet Burak Oğuz, Yusuf Kahya, Emre Can Çelebioğlu. Writing-original draft preparation: Ayça Koca, Ahmet Burak Oğuz. Writing-review and editing: Ayça Koca, Ahmet Burak Oğuz Hua Yan, Pengfei Liu. Supervision: Ayten Kayı Cangır, Kaan Orhan, Müge Günalp. All authors read and approved the final version of the manuscript.

Conflict of Interest: The authors declare that they have no conflict of interest.

Financial Disclosure: The authors declare that no funds, grants, or other support were received during the preparation of the manuscript.

REFERENCES

1. Eschweiler J, Li J, Quack V et al. Anatomy, Biomechanics, and Loads of the Wrist Joint. Life 2022;12:188.
2. Shahabpour M, Abid W, Van Overstraeten L, De Maeseneer M. Wrist Trauma: More Than Bones. Journal of the Belgian Society of Radiology 2021;105:90.
3. Obert L, Loisel F, Jardin E, Gasse N, Lepage D. High-energy injuries of the wrist. Orthopaedics & traumatology, surgery & research 2016;102:81-93.
4. 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. European journal of trauma and emergency surgery 2022;48:585–592.
5. Bruno F, Arrigoni, Palumbo P, et al. The Acutely Injured Wrist. Radiologic clinics of North America 2019;57:943–955.
6. Pinto A, Reginelli A, Pinto F, et al. Errors in imaging patients in the emergency setting. The British journal of radiology 2016;89:20150914.
7. Elzinga JL, Dunne C L, Vorobeichik A, et al. A Systematic Review Protocol to Determine the Most Effective Strategies to Reduce Computed Tomography Usage in the Emergency Department. Cureus 2020;12: e9509.
8. Ma Y, Lin C, Liu S, et al. Radiomics features based on internal and marginal areas of the tumor for the preoperative prediction of microsatellite instability status in colorectal cancer. Frontiers in Oncology 2022;12:1020349.
9. Wu J, Fang Q, Yao J, et al. Integration of ultrasound radiomics features and clinical factors: A nomogram model for identifying the Ki-67 status in patients with breast carcinoma. Frontiers in Oncology 2022;12:979358.
10. Xie D, Xu F, Zhu W, et al. Delta radiomics model for the prediction of progression-free survival time in advanced non-small-cell lung cancer patients after immunotherapy. Frontiers in Oncology 2022;12:990608.
11. Cangir AK, Orhan K, Kahya Y, et al. A CT-Based Radiomic Signature for the Differentiation of Pulmonary Hamartomas from Carcinoid Tumors. Diagnostics 2022;12:416.
12. Langerhuizen DWG, Bulstra AEJ, Janssen SJ, et al. Is Deep Learning On Par with Human Observers for Detection of Radiographically Visible and Occult Fractures of the Scaphoid?. Clinical orthopaedics and related research 2020;478:2653-2659.
13. Long, J., E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation. In:2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.
14. Ronneberger O, Fischer P, Brox T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2015, Lecture Notes in Computer Science 2015; 9351.
15. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017;39:137-1149.
16. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017;2117-2125.
17. Zech JR, Santomartino SM, Yi PH. Artificial Intelligence (AI) for Fracture Diagnosis: An Overview of Current Products and Considerations for Clinical Adoption, From the AJR Special Series on AI Applications. American journal of roentgenology 2022;219:869-878.
18. Lamb L, Kashani P, Ryan J, et al. Impact of an in-house emergency radiologist on report turnaround time. Canadian journal of emergency medicine 2015;17:21-26.
19. Duron L, Ducarouge A, Gillibert A, et al. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. Radiology 2021;300:120-129.
20. Guermazi A, Tannoury C, Kompel AJ, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. Radiology 2022;302:627-36.
21. Nehrer S, Ljuhar R, Steindl P, et al. Automated Knee Osteoarthritis Assessment Increases Physicians' Agreement Rate and Accuracy: Data from the Osteoarthritis Initiative. Cartilage 2021;13:957S-965S.
22. Cohen M, Puntonet J, Sanchez J, et al. Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs. European radiology 2023;33:3974-3983.
23. Heo YM, Kim SB, Yi JW, et al. Evaluation of associated carpal bone fractures in distal radial fractures. Clinics in orthopedic surgery 2013;5:98-104.
24. Nguyen T, Maarek R, Hermann AL, et al. Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. Pediatric Radiology. 2022;52:2215-2226.
25. Herpe G, Nelken H, Vendeuvre T, et al. Effectiveness of an artificial intelligence software for limb radiographic fracture recognition in an emergency department. Journal of Clinical Medicine 2024;13:5575.
26. Canoni-Meynet L, Verdot P, Danner A, Calame P, Aubry S. Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. Diagnostic and interventional imaging 2022;103:594-600.
27. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLoS One 2022;17:e0271872.
28. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proceedings of the National Academy of Sciences 2018;115:11591-11596.
29. Snaith BA, Lancaster A. Clinical history and physical examination skills–a requirement for radiographers ?. Radiography 2008;14150-153.
30. Huang X, Han F, Chen YF, et al. Bibliometric analysis of the application of artificial intelligence in orthopedic imaging. Quantitative Imaging in Medicine and Surger 2025;15:3993
31. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education 2023;23:689.