Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diagnostic Performance of a Deep Learning Model versus Emergency Medicine Specialists in the Detection of Paediatric Elbow Fractures: A Retrospective Study

Oğuzhan Oğuz

o.oguz1996@gmail.com

Şırnak Şehit Aydoğan Aydın State Hospital

Nurbanu Aksoy

Middle East Technical University

Havva Şahin Kavaklı

Ankara Yıldırım Beyazıt University

Selçuk Coşkun

Ankara Bilkent City Hospital

Research Article

Keywords: Emergency Medicine, Deep Learning, Paediatrics, Trauma, Radiography

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11018023/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

The aim of this study was to compare the diagnostic performance of a deep learning model developed for fracture detection on paediatric elbow radiographs with that of emergency medicine specialists and to evaluate its potential utility as a clinical decision-support tool in the emergency department.

Methods

Patients younger than 18 years who presented to the Emergency Department of Ankara Bilkent City Hospital between 14 March 2019 and 31 December 2024 and underwent both elbow radiography and computed tomography were retrospectively reviewed. A total of 492 patients were included in the study. To address class imbalance, uninjured contralateral elbow radiographs were added, yielding a dataset of 580 cases, each comprising an anteroposterior and a lateral elbow radiograph. The dataset was split into training, validation, and test sets at a ratio of 70:15:15, respectively. A deep learning model based on the EfficientNetV2-Small architecture was developed for fracture detection. Diagnostic performance was assessed using sensitivity, specificity, positive and negative predictive values, likelihood ratios, accuracy, and balanced accuracy, while pairwise comparisons between evaluators were assessed using McNemar's test.

Results

The deep learning model demonstrated a sensitivity of 32.56%, specificity of 61.36%, accuracy of 47.13%, a balanced accuracy of 0.47, and a macro-averaged F1 score of 0.46. The sensitivities of the five emergency medicine specialists ranged from 53.49% to 86.05%, specificities from 50.00% to 77.27%, and overall accuracy from 60.47% to 68.97%.

Conclusion

Although the developed deep learning model was insufficient for standalone clinical use, it has the potential to serve as a decision-support tool to complement clinical assessment in the emergency department. Future prospective, multicentre, and multimodal studies may help improve its diagnostic accuracy and further establish its clinical utility.

1. Introduction

Traumatic injuries in the paediatric age group are among the most common reasons for emergency department visits [1]. Fractures account for 10–25% of these injuries, while elbow fractures constitute approximately 5–10% of all fractures in childhood [2]. The most common type is supracondylar humeral

fracture [3]. Although two-view plain radiography is the first-line imaging modality for diagnosis, the complex anatomy of the elbow joint, difficulties in obtaining adequate cooperation from children, and the presence of epiphyseal growth plates may complicate radiographic assessment [4].

Although computed tomography (CT) provides high diagnostic accuracy, it is not the first-line imaging modality in paediatric patients because of ionising radiation exposure, cost, and limited availability [5]. In recent years, deep learning-based artificial intelligence models have attracted considerable attention because of their potential to reduce human error in radiological image interpretation and accelerate the diagnostic process [6].

In this study, the performance of a deep learning model developed for fracture detection on paediatric elbow radiographs was compared with the assessments of emergency medicine specialists to investigate the potential utility of artificial intelligence as a clinical decision-support tool in the emergency department.

2. Materials and Methods

2.1. Study Design and Dataset Construction

This was a single-centre, retrospective study (Fig. 1). Patients younger than 18 years who presented to the Emergency Department of Ankara Bilkent City Hospital between 14 March 2019 and 31 December 2024 and underwent both elbow radiography and CT were retrospectively reviewed. Only patients who underwent both elbow radiography and CT during the same diagnostic evaluation were included in the study. A total of 538 patients were assessed for eligibility, of whom 46 were excluded due to incomplete imaging, leaving 492 patients for inclusion. CT reports generated during routine clinical care were used as the reference standard for determining the presence or absence of fracture. The CT examinations were interpreted by radiologists as part of routine clinical practice, and the reports were generated prospectively during the patients' diagnostic evaluation. The CT images were not independently re-reviewed for this study.

To increase the number of examinations without fractures, 271 uninjured contralateral elbow radiographs were added to the dataset. As most cases belonged to the no-fracture class, the majority class was randomly undersampled to match the fracture class (n = 290), and 183 cases were excluded accordingly. This procedure yielded a balanced dataset of 580 cases in total, each comprising an anteroposterior (AP) and a lateral elbow radiograph. The dataset was subsequently split into training, validation, and test subsets at a ratio of 70:15:15, respectively, using stratified sampling so that the class ratios were preserved across subsets [7].

The admission number, age, sex, and fracture status recorded for each patient were processed and cleaned in Python using the PyTorch library; sex and fracture status were encoded as binary variables, and the corresponding image file paths were generated automatically.

2.2. Image Preprocessing and Data Augmentation

During preprocessing, the black borders surrounding the images were removed using a content-based cropping method based on Otsu thresholding and contour detection. The cropped images were then resized to 224 × 224 pixels, converted to RGB format, and transformed into tensors. For the training set, data augmentation was applied, comprising rotation, flipping, and brightness and contrast adjustments, together with ImageNet normalisation; for the validation and test sets, only resizing and normalisation were performed [8].

2.3. Model Architecture and Training

For the binary classification task, a multi-view convolutional neural network (CNN) was implemented, in which the EfficientNetV2-Small architecture served as the backbone and the AP and lateral radiographs were processed through separate branches (Fig. 1). The resulting feature vectors were concatenated and passed to a classification head consisting of a dropout layer (0.2), a fully connected layer with 256 units, a rectified linear unit (ReLU) activation, a second dropout layer (0.5), and a final fully connected layer with a single output unit. The model was trained with the binary cross-entropy loss (BCEWithLogitsLoss), using a batch size of 16, with the data shuffled at the beginning of each epoch.

A gradual fine-tuning strategy was adopted: during the first four epochs, only the classification head was trained; from the fifth epoch onwards, the final blocks of the backbone were unfrozen; and from the tenth epoch onwards, all layers were trained. The Adam optimiser was used with an initial learning rate of 1e-4, and the learning rate was adjusted using the ReduceLROnPlateau scheduler (patience = 2, factor = 0.5).

To prevent overfitting, early stopping was applied (patience = 7), and the training process was monitored through the validation performance using the learning curves. Because both the model, evaluated at a fixed classification threshold of 0.50, and the emergency medicine specialists each provided a single operating point, discriminative performance was summarised using balanced accuracy (the mean of sensitivity and specificity) rather than a threshold-swept area under the receiver operating characteristic curve.

2.4. Evaluation by Emergency Medicine Specialists

Five emergency medicine specialists independently evaluated the AP and lateral elbow radiographs from the test dataset using a standardized online questionnaire accessible exclusively to them. Only radiographic images were provided, without clinical information, age, or sex. Cases were classified as fracture or no fracture, without a time limit and with specialists blinded to CT results and each other’s assessments.

2.5. Model Interpretability

To visualise the image regions that contributed most to the model's classification decisions, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied. For each test image, which had already undergone cropping and normalisation, heat maps were generated from the pooled gradients and then

normalised. The resulting colour-coded maps were overlaid on the original radiograph, and the anteroposterior and lateral outputs for each patient were saved together as a single .png file (Fig. 2).

3. Results

The training set comprised 406 cases, the validation set 87 cases, and the test set 87 cases. In the training set, fractures were present in 203 cases and absent in 203 cases; 305 cases were male, 101 were female, and the mean age was 10.41 ± 4.33 years. In the validation set, fractures were present in 44 cases and absent in 43 cases; 65 cases were male, 22 were female, and the mean age was 10.08 ± 4.26 years. In the test set, fractures were present in 43 cases and absent in 44 cases; 59 cases were male, 28 were female, and the mean age was 10.06 ± 4.11 years (Table 1).

Table 1
Demographic and clinical characteristics by dataset

	Total Dataset	CT Report		Sex		Age ± SD (years)
		Fracture	No Fracture	Male	Female	
Training	406 (70%)	203	203	305	101	10.41 (± 4.33)
Validation	87 (15%)	44	43	65	22	10.08 (± 4.26)
Test	87 (15%)	43	44	59	28	10.06 (± 4.11)

The sensitivity of the deep learning model was 32.56% (95% CI, 20.49–47.48), specificity was 61.36% (95% CI, 46.62–74.28), positive predictive value was 45.16%, negative predictive value was 48.21%, and accuracy was 47.13% (95% CI, 36.98–57.51). The sensitivities of the five emergency medicine specialists ranged from 53.49% to 86.05%, specificities from 50.00% to 77.27%, and accuracies from 60.47% to 68.97% (Table 2).

Table 2
Diagnostic performance of the deep learning model and specialists in the test dataset

Variables	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)	PPV (95% CI)	NPV (95% CI)	Accuracy (95% CI)
DL model	32.56 (20.49–47.48)	61.36 (46.62–74.28)	0.84 (0.48–1.49)	1.10 (0.80–1.50)	45.16 (29.16–62.23)	48.21 (35.67–60.99)	47.13 (36.98–57.51)
Specialist-1	74.42 (59.76–85.07)	63.64 (48.87–76.22)	2.05 (1.33–3.14)	0.40 (0.23–0.70)	66.67 (52.54–78.32)	71.79 (56.22–83.46)	68.97 (58.62–77.71)
Specialist-2	62.79 (47.86–75.62)	61.36 (46.62–74.28)	1.63 (1.05–2.52)	0.61 (0.39–0.95)	61.36 (46.62–74.28)	62.79 (47.86–75.62)	62.07 (51.57–71.55)
Specialist-3	53.49 (38.92–67.49)	77.27 (63.01–87.16)	2.35 (1.28–4.34)	0.60 (0.42–0.86)	69.70 (52.66–82.62)	62.96 (49.63–74.58)	65.52 (55.06–74.66)
Specialist-4	61.90 (46.81-75.00)	59.09 (44.41–72.31)	1.51 (0.99–2.32)	0.64 (0.41–1.02)	59.09 (44.41–72.31)	61.90 (46.81-75.00)	60.47 (49.90-70.14)
Specialist-5	86.05 (72.74–93.44)	50.00 (35.83–64.17)	1.72 (1.25–2.37)	0.28 (0.13–0.62)	62.71 (49.95–73.92)	78.57 (60.46–89.79)	67.82 (57.43–76.70)

The model had a balanced accuracy (defined as the mean of sensitivity and specificity) of 0.47 and a macro-averaged F1 score of 0.46. The balanced accuracy of the emergency medicine specialists ranged from 0.61 to 0.69, while their F1 scores ranged from 0.61 to 0.67.

The performance of the deep learning model and the five emergency medicine specialists was compared using McNemar's test. All emergency medicine specialists achieved a higher number of correct classifications than the deep learning model, and this difference was statistically significant.

No statistically significant differences were observed in the pairwise comparisons among the emergency medicine specialists (Table 3).

Table 3
Pairwise comparison of diagnostic performance using McNemar’s test

Evaluator Pair		A Correct, B Incorrect	B Correct, A Incorrect	McNemar's Chi-Square Test	p-value
A	B				
DL model	Specialist-1	10	29	9.26	0.002
DL model	Specialist-2	12	25	4.57	0.033
DL model	Specialist-3	8	24	8.00	0.005
DL model	Specialist-4	10	22	4.50	0.034
DL model	Specialist-5	14	32	7.04	0.008
Specialist-1	Specialist-2	15	9	1.50	0.221
Specialist-1	Specialist-3	15	12	0.33	0.564
Specialist-1	Specialist-4	16	9	1.96	0.162
Specialist-1	Specialist-5	10	9	0.05	0.819
Specialist-2	Specialist-3	9	12	0.43	0.513
Specialist-2	Specialist-4	13	12	0.04	0.841
Specialist-2	Specialist-5	8	13	1.19	0.275
Specialist-3	Specialist-4	15	11	0.62	0.433
Specialist-3	Specialist-5	12	14	0.15	0.695
Specialist-4	Specialist-5	8	14	1.64	0.201

4. Discussion

This study developed a multi-view deep learning model for fracture detection on paediatric elbow radiographs and compared its diagnostic performance with that of five emergency medicine specialists

evaluating the same test set under identical, blinded conditions. Although the model's specificity fell within the range achieved by the specialists, its sensitivity was substantially lower. McNemar's test demonstrated statistically significant differences between the model and each of the five specialists, whereas no statistically significant differences were observed among the specialists themselves. These findings indicate that, in its current form, the model does not achieve diagnostic performance comparable to that of emergency medicine specialists, particularly for fracture detection.

The model demonstrated relatively preserved performance in identifying fracture-free examinations. One possible explanation is that exposure to a sufficient number of normal elbow radiographs during training may have facilitated the learning of normal anatomical patterns. In addition, fracture-free examinations may have more consistently demonstrated optimal positioning and image quality and may have been less frequently complicated by splints, dislocations, or other factors that can obscure anatomical structures. Reliable recognition of the fracture-free class may have clinical value, as accurate exclusion of fracture could potentially contribute to reducing unnecessary advanced imaging and specialist consultation. However, the relatively preserved specificity observed in this study should not be interpreted as sufficient evidence for clinical use without further prospective validation.

In contrast, the lower sensitivity of the model relative to the emergency medicine specialists may reflect several challenges intrinsic to paediatric elbow trauma. Pain and limited cooperation may make optimal positioning difficult, particularly on lateral radiographs, while splints or other immobilisation devices and concomitant dislocations may further obscure relevant anatomical structures. Variable image quality and subtle radiographic findings may also make fracture detection challenging for an image-based model. In particular, fractures associated with subtle cortical abnormalities or indirect findings, such as fat pad signs, may be difficult to identify when the radiographic appearance is suboptimal. The relatively limited training dataset may have further constrained the model's ability to generalise across the broad spectrum of paediatric elbow fracture patterns and more challenging presentations. These factors may collectively explain the lower sensitivity observed in the present study.

The growing integration of artificial intelligence and machine learning into medical imaging makes paediatric elbow trauma a particularly relevant application, given the diagnostic challenges associated with these injuries and the need for timely assessment in the emergency department. Nevertheless, our findings suggest that the model should currently be considered a supportive rather than an independent diagnostic tool. Its potential incremental value when used alongside clinical assessment and physician interpretation should be evaluated in larger, prospective and multicentre studies. In particular, future models that integrate radiographic findings with clinical variables such as age, mechanism of injury, physical examination findings, and clinical history may provide a more clinically relevant representation of the diagnostic process.

The diagnostic performance of previously reported artificial intelligence models for paediatric elbow fracture detection has generally been higher than that observed in the present study. Li et al. reported a sensitivity of 93.2% and a specificity of 94.8% for a DenseNet-201 model trained using radiographs from

1370 paediatric patients [9]. Ye et al. reported a sensitivity of 89% and a specificity of 92% on an independent test set, with performance approaching that of emergency medicine specialists and paediatric radiologists [10]. Similarly, a systematic review and meta-analysis by Le Nguyen Binh et al. reported pooled sensitivity and specificity of 93% (95% CI, 91–96) and 89% (95% CI, 85–92), respectively, for deep learning models detecting paediatric elbow fractures [11]. The lower performance observed in our study may be attributable to methodological differences among studies, including differences in dataset size and diversity, fracture-type distribution, image quality, reference standards, patient selection, and model architecture. In addition, differences in dataset construction and class distribution may influence diagnostic metrics and limit direct comparison between studies.

Systematic analysis of misclassified images using visualisation techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) may provide further insight into the regions contributing to model predictions and whether these regions correspond to clinically relevant anatomical structures. Such analyses may help distinguish clinically meaningful model errors from predictions driven by image quality, positioning, immobilisation devices, or other non-anatomical features and could inform future model refinement. Ziegner et al. reported that artificial intelligence-assisted fracture detection systems may improve the diagnostic performance of emergency department physicians and that high-performing systems may serve as adjunctive tools in clinical decision-making [12]. Taken together with our findings, this supports positioning artificial intelligence not as a stand-alone diagnostic tool intended to replace physicians, but as a decision-support system that complements clinical assessment and may assist physicians in the interpretation of paediatric elbow radiographs.

5. Limitations

This study has several limitations. First, its retrospective, single-centre design and relatively limited dataset may restrict the generalisability of the findings to other patient populations, institutions, and imaging protocols. The retrospective nature of image acquisition also resulted in variability in radiographic positioning and image quality. In particular, the inability to achieve an ideal degree of flexion on lateral radiographs, as well as the potentially subtle appearance of indirect radiographic findings such as fat pad signs, may have increased the difficulty of fracture classification.

Second, the dataset was constructed using random undersampling to achieve class balance. Although this approach prevented the majority class from dominating model training, it reduced the number of available fracture-free examinations and may have limited the diversity of normal presentations represented in the training data. In addition, uninjured contralateral elbow radiographs were used to augment the fracture-free class. These images may differ from radiographs obtained from children presenting with suspected elbow trauma and therefore may not fully represent the spectrum of fracture-free examinations encountered in routine emergency practice.

Third, only patients who underwent both elbow radiography and computed tomography during the same diagnostic evaluation were included. This inclusion criterion may have introduced selection bias because

patients undergoing CT may have had greater diagnostic uncertainty, more complex injuries, or clinical or radiographic findings that prompted advanced imaging. Consequently, the study population may not be representative of the broader population of children presenting to the emergency department with elbow trauma.

Fourth, both the deep learning model and the emergency medicine specialists were evaluated using radiographic images alone and without access to clinical information. The specialists therefore could not incorporate potentially relevant factors such as age, mechanism of injury, physical examination findings, or clinical history into their assessments. Although restricting both groups to the same information ensured a controlled and blinded comparison, this design does not fully reflect real-world emergency department practice, where clinical context is routinely integrated with radiographic interpretation.

Finally, the model was evaluated at a single operating point and was not externally validated in an independent cohort. Therefore, its performance in other institutions, populations, imaging protocols, and clinical settings remains uncertain. Future studies should include larger and more diverse datasets, external validation across multiple centres, prospective evaluation under real-world clinical conditions, and multimodal approaches integrating radiographic and clinical information. These approaches may improve the generalisability and clinical utility of artificial intelligence for paediatric elbow fracture detection.

6. Conclusion

The deep learning model developed in this study demonstrated lower diagnostic performance than the emergency medicine specialists in detecting fractures on paediatric elbow radiographs. The model's low sensitivity indicates that its standalone clinical use is not appropriate in its current form. Larger, multicentre datasets, standardisation of imaging acquisition, and multimodal approaches incorporating clinical data appear to be priorities for future studies.

Abbreviations

AP
Anteroposterior
CI
Confidence Interval
CNN
Convolutional Neural Network
CT
Computed Tomography
DL
Deep Learning

Grad-CAM
Gradient-weighted Class Activation Mapping
LR+
Positive Likelihood Ratio
LR-
Negative Likelihood Ratio
NPV
Negative Predictive Value
SD
Standard Deviation
PPV
Positive Predictive Value
ReLU
Rectified Linear Unit

Declarations

Ethics Approval and Consent to Participate:

The study was approved by the Scientific and Ethical Evaluation Board for Medical Research No. 2 of Ankara Bilkent City Hospital on 16 April 2025 (decision no: TABED 2-25-969). The requirement for informed consent was waived because of the retrospective nature of the study. The study was conducted in accordance with the Declaration of Helsinki.

Human Ethics and Consent to Participate

Not applicable

Consent for Publication:

Not applicable.

Competing Interests:

The authors declare that they have no conflicts of interest related to this study.

Funding:

No funding or financial support was received for this study.

Author Contribution

OO and HSK conceived and designed the study. OO collected the data, and NA contributed to the development of the deep learning model. All authors contributed to data analysis and manuscript preparation, critically reviewed the final manuscript, and approved the final version.

Acknowledgements:

Not applicable.

Data Availability

The datasets generated and analysed during the current study are not publicly available because they contain potentially identifiable clinical and radiographic data and are subject to institutional and ethical restrictions. Deidentified data may be available from the corresponding author on reasonable request and with permission from the relevant institutional authorities.

References

1. Danseco ER, Miller TR, Spicer RS. Incidence and costs of 1987–1994 childhood injuries: demographic breakdowns. Pediatrics. 2000;105(2):e27.
2. Spady DW, Saunders DL, Schopflocher DP, Svenson LW. Patterns of injury in children: a population-based approach. Pediatrics. 2004;113(3 Pt 1):522–9.
3. Emery KH, Zingula SN, Anton CG, Salisbury SR, Tamai J. Pediatric elbow fractures: a new angle on an old topic. Pediatr Radiol. 2016;46(1):61–6.
4. Kappelhof B, Roorda BL, Poppelaars MA, The B, Eygendaal D, Mulder PGH, et al. Occult fractures in children with a radiographic fat pad sign of the elbow: a meta-analysis of 10 published studies. JBJS Rev. 2022;10(10). e22.00114.
5. Kraus R, Dresing K. Rational usage of fracture imaging in children and adolescents. Diagnostics (Basel). 2023;13(3):538. 10.3390/diagnostics13030538.
6. Kutbi M. Artificial intelligence-based applications for bone fracture detection using medical images: a systematic review. Diagnostics (Basel). 2024;14(17):1879. 10.3390/diagnostics14171879.
7. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
8. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6:60. 10.1186/s40537-019-0197-0.
9. Li J, Hu W, Wu H, Chen Z, Chen J, Lai Q, et al. Detection of hidden pediatric elbow fractures in X-ray images based on deep learning. J Radiat Res Appl Sci. 2024;17(2):100893.

10.1016/j.jrras.2024.100893.

10. Ye Q, Wang Z, Lou Y, Yang Y, Hou J, Liu Z, et al. Deep learning approach based on a patch residual for pediatric supracondylar subtle fracture detection. Biomol Biomed. 2025;25(7):1631–46. 10.17305/bb.2024.11341.
11. Binh LN, Nhu NT, Nhi PTU, Son DLH, Bach N, Huy HQ, et al. Impact of deep learning on pediatric elbow fracture detection: a systematic review and meta-analysis. Eur J Trauma Emerg Surg. 2025;51(1):115. 10.1007/s00068-025-02779-w.
12. Ziegner M, Pape J, Lacher M, Brandau A, Kelety T, Mayer S, et al. Real-life benefit of artificial intelligence-based fracture detection in a pediatric emergency department. Eur Radiol. 2025;35(10):5881–90. 10.1007/s00330-025-11554-9.

Figures

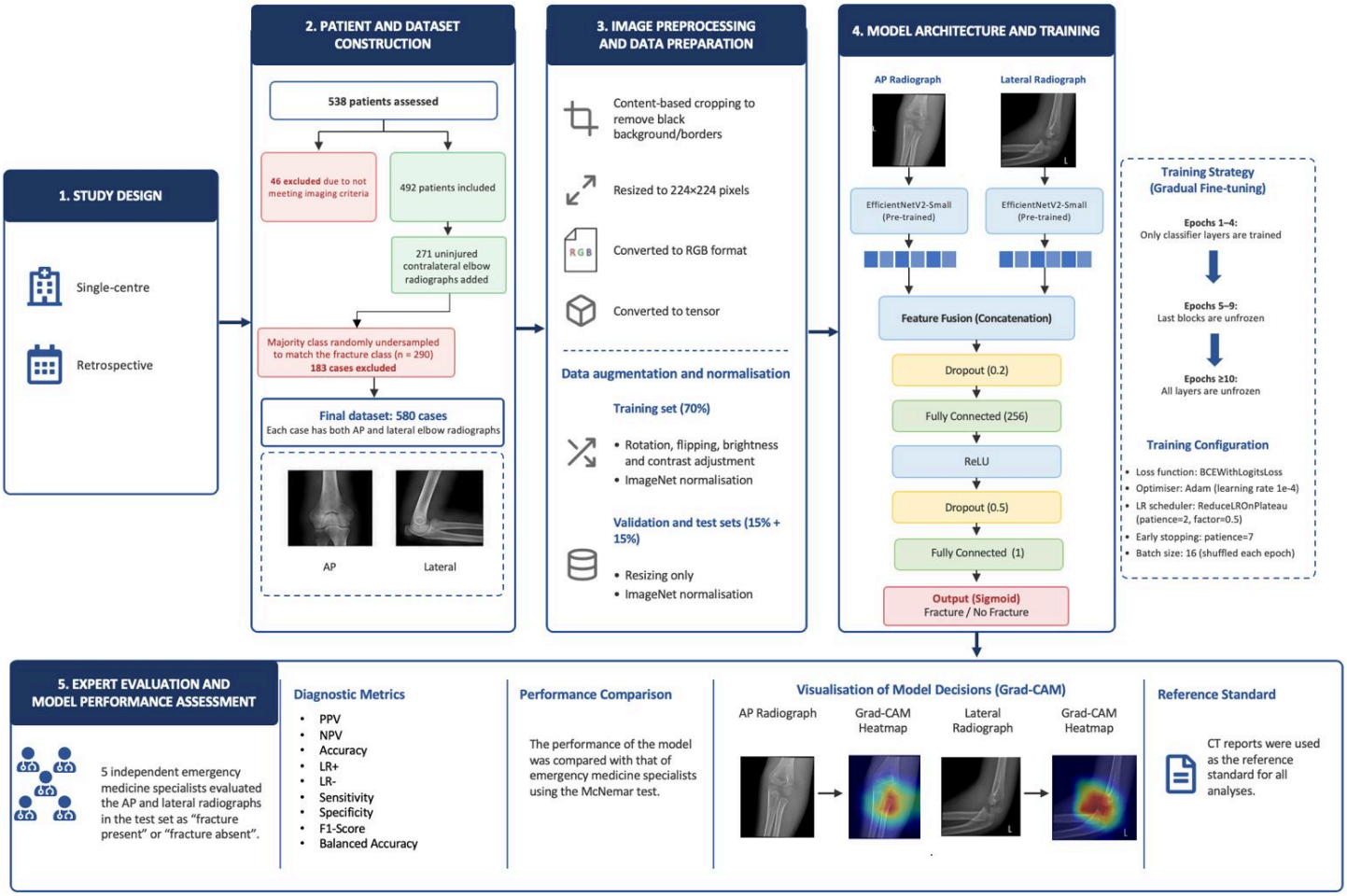


Figure 1

Overview of the study design and methodology

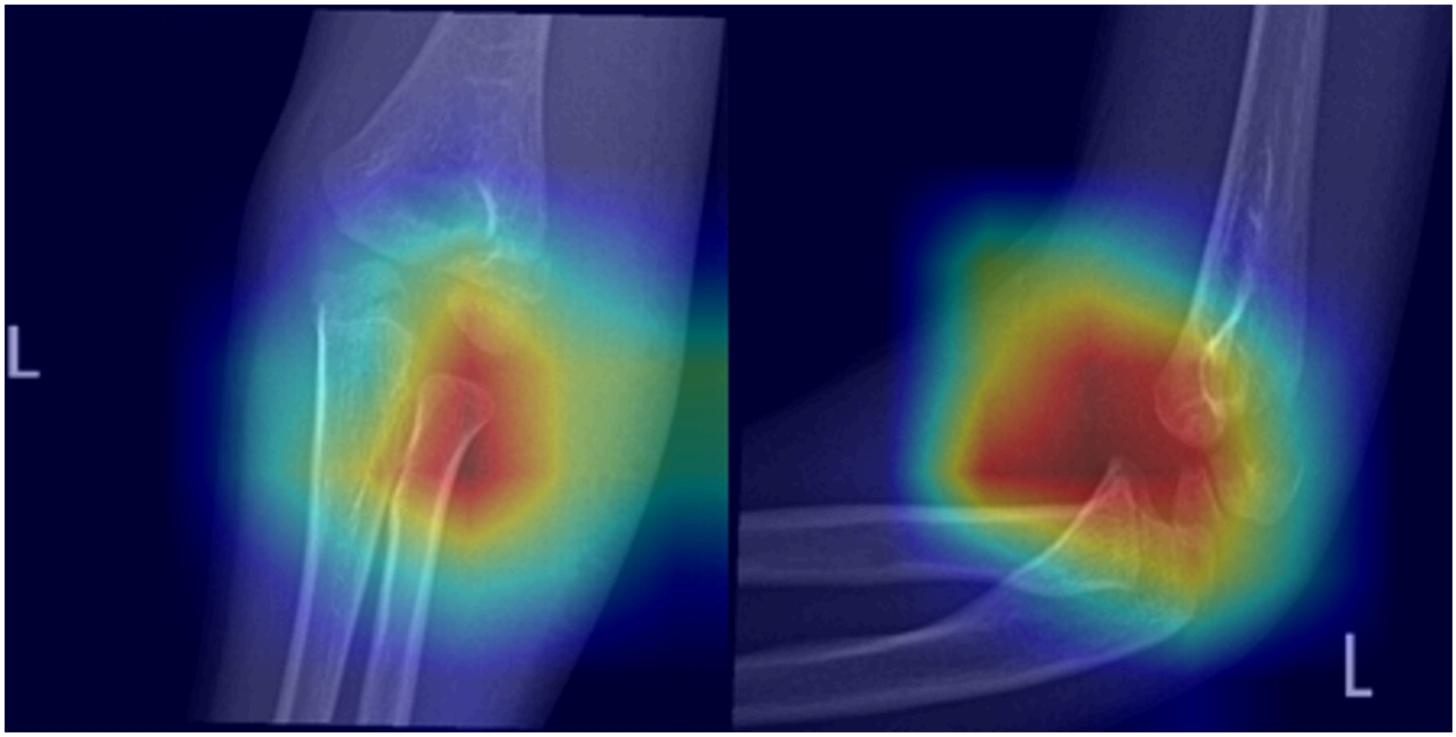


Figure 2

Grad-CAM output for the AP (left) and lateral (right) elbow radiographs of a representative patient.