Use of deep learning model for paediatric elbow radiograph binomial classification: initial experience, performance and lessons learnt

Mark Bangwei Tan[1], MRCS, FRCR, Yuezhi Russ Chua[2], BEng, Qiao Fan[3], PhD, Marielle Valerie Fortier[4,5,6], MD, FRCPC, Peiqi Pearlly Chang[7], MBBS, MRCPCH

[1]Department of Diagnostic Radiology, Singapore General Hospital, [2]Agency for Science, Technology and Research, [3]Centre for Quantitative Medicine, Duke-NUS Medical School, [4]Department of Diagnostic and Interventional Imaging, KK Women's and Children's Hospital, [5]Duke-NUS Medical School, [6]Institute for Clinical Sciences, Agency for Science, Technology and Research, [7]Department of Paediatrics, KK Women's and Children's Hospital, Singapore

Abstract

Introduction: In this study, we aimed to compare the performance of a convolutional neural network (CNN)-based deep learning model that was trained on a dataset of normal and abnormal paediatric elbow radiographs with that of paediatric emergency department (ED) physicians on a binomial classification task.

Methods: A total of 1,314 paediatric elbow lateral radiographs (patient mean age 8.2 years) were retrospectively retrieved and classified based on annotation as normal or abnormal (with pathology). They were then randomly partitioned to a development set (993 images); first and second tuning (validation) sets (109 and 100 images, respectively); and a test set (112 images). An artificial intelligence (AI) model was trained on the development set using the EfficientNet B1 network architecture. Its performance on the test set was compared to that of five physicians (inter-rater agreement: fair). Performance of the AI model and the physician group was tested using McNemar test.

Results: The accuracy of the AI model on the test set was 80.4% (95% confidence interval [CI] 71.8%–87.3%), and the area under the receiver operating characteristic curve (AUROC) was 0.872 (95% CI 0.831–0.947). The performance of the AI model vs. the physician group on the test set was: sensitivity 79.0% (95% CI: 68.4%–89.5%) vs. 64.9% (95% CI: 52.5%–77.3%; $P = 0.088$); and specificity 81.8% (95% CI: 71.6%–92.0%) vs. 87.3% (95% CI: 78.5%–96.1%; $P = 0.439$).

Conclusion: The AI model showed good AUROC values and higher sensitivity, with the *P*-value at nominal significance when compared to the clinician group.

Keywords: Artificial intelligence, emergency radiology, machine learning, musculoskeletal radiology, paediatric radiology

INTRODUCTION

Timely interpretation of radiographs in the emergency department (ED) is important for the best patient outcomes. In Singapore, paediatric patients may present to primary care or adult hospitals instead of national tertiary paediatric hospitals in acute care settings. Primary care doctors, ED physicians and radiologists who do not routinely review paediatric imaging may be less familiar with interpreting such radiographs radiographs.[1] Among paediatric fractures, it is particularly challenging to diagnose acute elbow fractures, as they sometimes present with only effusion, with the abnormality on imaging seen as abnormal elevation of periarticular fat secondary to the abnormal accumulation of synovial fluid in the olecranon, coronoid or radial fossa (i.e., anterior or posterior fat pad signs).[2-4] The reported frequency of occult or initially missed acute paediatric elbow fractures is 17%–77%.[5-7] This is commonly due to either misdiagnosis of subtle elbow fractures

Access this article online

Quick Response Code:

Website: https://journals.lww.com/SMJ

DOI: 10.4103/singaporemedj.SMJ-2022-078

Correspondence: Dr Mark Bangwei Tan, Consultant, Department of Diagnostic Radiology, Singapore General Hospital, Outram Road, 169608, Singapore. E-mail: mark.tan.b.w@singhealth.com.sg

Received: 02 May 2022 **Accepted:** 15 Nov 2022 **Published:** 29 Nov 2023

How to cite this article: Tan MB, Chua YR, Fan Q, Fortier MV, Chang PP. Use of deep learning model for paediatric elbow radiograph binomial classification: initial experience, performance and lessons learnt. Singapore Med J 2025;66:208-14.

as normal or normal ossification growth centres as fractures, or unfamiliarity with and non-identification of fracture patterns seen mainly in the paediatric population (e.g., greenstick or torus fractures).[8] Misdiagnosis of paediatric elbow fracture is not without consequence, as the risk of morbidity increases when the diagnosis is delayed.[9,10]

As paediatric elbow radiographs are often reviewed by non-paediatric radiologists initially, accurate automated binomial classification and triage of elbow radiographs into normal and abnormal groups is of practical value for clinical management at the point of care. Recent developments in machine learning have made it possible to develop such algorithms for elbow radiograph triage into normal and abnormal groups using convolutional neural network (CNN)-based deep learning techniques, with reported accuracy of 88%–91%, sensitivity of 91%–93% and specificity of 84%–92%.[11-14] The current study aimed to develop an artificial intelligence (AI) model based on available local data for paediatric elbow radiograph binomial classification and compare its performance with those of clinicians on a test set assessing such a classification task. We hypothesed that the AI model would meet or exceed the performance of senior paediatric ED physicians at this task. To our knowledge, this is the first local study with these specific aims. We hope that the experience and lessons learnt from this project will be incorporated in AI models developed for other radiography types and imaging modalities in future.

METHODS

Data collection

The data source for our study was the radiology information system-picture archiving and communication system (RIS-PACS) of KK Women's and Children's Hospital, Singapore. No studies had been performed on this dataset previously. Approval for the study was obtained from SingHealth Institutional Review Board (IRB) (IRB no. 2019/2523), and informed consent was waived.

We retrospectively extracted the radiographs of the left and right elbows (in lateral view) of male and female patients aged 3–16 (mean 8.2) years, in a consecutive series from January 2015 to November 2015. Radiographs that showed casts or orthopaedic hardware were not extracted, as the aim of the model was triage of patients who presented with first onset of elbow symptoms in the ED. Of the 1,696 selected radiographs, 1,314 unique radiographs were extracted for analysis by a study team member with radiology postgraduate qualifications. Manual extraction of the entire image (excluding identifiers) without cropping was performed.

Data labelling

To increase the accuracy of the labels of the radiographs, the included radiographs were then annotated as ground truth — normal [Figure 1] or abnormal (with pathology, i.e., fracture, effusion, dislocation) [Figure 2] using the joint input of the original annotators and the annotators from the study team. All the annotators possess postgraduate radiology qualifications and have been deemed qualified by the paediatric radiology department to read out paediatric radiographs. The first annotator was the reader (one of a group of approximately 10–15 annotators) who prepared the initial radiographic report and, as part of the report, annotated the radiograph as either normal or abnormal. The second annotator from the study team reviewed and independently annotated the radiograph. If there was no discrepancy between the annotation of the first and second annotators, the final annotation would follow their consensus opinion. In the event of a discrepancy, the radiograph was referred to a third annotator, a specialist paediatric radiologist from the study team, for adjudication and designation of the final annotation.

Data partitions

The images were randomly separated into different partitions: 993 images in a development set (450 normal and 543 abnormal

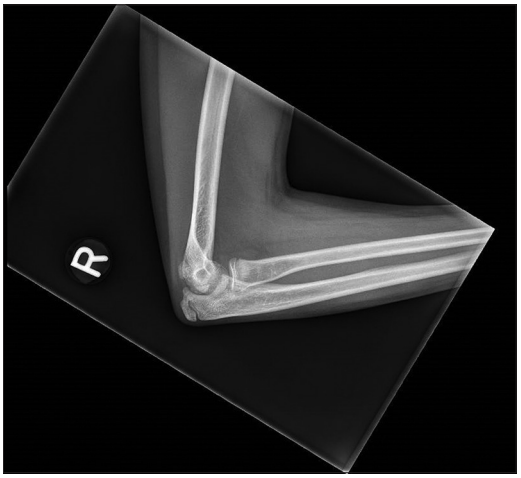

Figure 1: Example of a normal lateral elbow radiograph in this study.

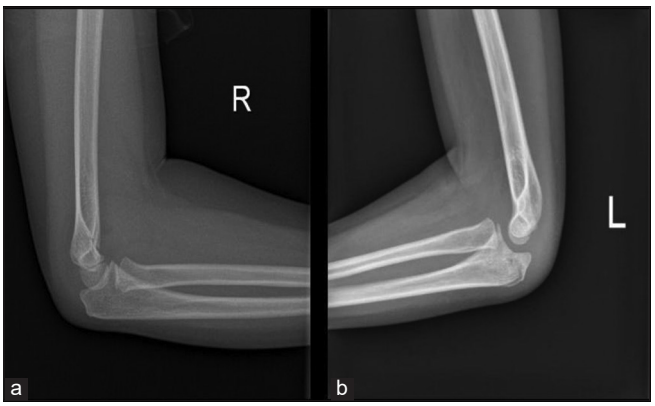

Figure 2: Examples of abnormal lateral elbow radiographs in this study. (a) Radiographs show an elbow effusion with elevation of the anterior and visible posterior fat pad and (b) a mildly displaced fracture of the proximal ulna.

images); 109 images in a tuning (validation) set (49 normal and 60 abnormal images, representing 10% of the number of images in the development set); 100 images in a second tuning (validation) set (50 normal and 50 abnormal images, representing 10% of the number of images in the development set), and 112 images in a test set (which was not exposed to the model in its development and tuning [validation], 57 abnormal and 55 normal images). This testing sample size was determined by what was reasonably achievable by a clinician during a test sequence, a practice consistent with previous studies.[15]

Image preparation and model development

In model development, the following software packages were used: tensorflow-gpu 2.3.1 built-in function, tf.keras.preprocessing.image.ImageDataGenerator, and the EfficientNet B1 network architecture.[16] These versions of the software were employed: EfficientNet version 1.1.1, Keras-Applications version 1.0.8, Keras-Preprocessing version 1.1.2, Scikit-image version 0.17.2, Scikit-learn version 0.24.1, Scipy version 1.5.3, Tensorboard version 2.3.0, Tensorboard-plugin-wit version 1.7.0, Tensorflow-estimator version 2.3.0 and Tensorflow-gpu version 2.3.1. The model was developed on a NVIDIA RTX 2060 graphics processing unit. The images were extracted in JPG or PNG format, and resized to 240 × 240 resolution by the AI model upon presentation of the original scale set images to it (this was the maximum image resolution possible, with the limitations of the AI model and hardware made available to the study team being unable to process images at higher resolutions).

The training of CNN-based deep learning algorithms requires the data scientist to predefine several variables that affect the training process. These are known as hyperparameters, and include manually set variables such as the deep learning network architecture, its learning rate, as well as the type and number of data augmentations (i.e. image manipulation, e.g., zooming, panning, rotation, contrast to increase dataset size), should they be employed during the training process. In this study, given the relatively modest dataset size, the model was also initialised with pretrained parameter weights from the ImageNet dataset[17,18] (this is a large dataset used to adapt models for image recognition, and this process is also known as transfer learning; 'parameters' refers to variables of the model derived by data fitting internal to its system). None of these model parameters was frozen due to the intrinsic differences in the images of the ImageNet dataset from the development dataset. The data augmentations employed on the training dataset were random rotations (10°), flips, positional shifts (0.05) and affine transformations (zoom, 0.05), applied to each epoch (training iteration where the model analyses all images in the training set). Loss is defined as the amount of error generated by each model; models start training with high loss values with an accuracy of 50%, and optimisers such as Adam optimisers may reduce this loss by changing the parameter weights of the network during training towards a minimum. In this case, an Adam optimiser was used with a 1e-4 learning rate without a scheduler. The batch size was 8.

There were 300 epochs (iterations where the model analyses the data in the training set) run with a minimum patience of 50 epochs (number of epochs performed before stopping the model to training when an apparent minimum to the loss value is encountered) set. The best performing model out of the total number of training epochs was selected based on the model that had the best accuracy on the second tuning (validation) set, which, in our case, also had the highest area under the receiver operating characteristic curve (AUROC). In our case, two tuning (validation) sets were used instead of one to prevent overfitting, a known entity in machine learning where the developed AI model is specific to the data in the training set and unable to subsequently generalise to similar but non-identical data presented to it. To evaluate if our model was making predictions on the right pathological features and not artefacts within the elbow radiograph image, we produced class activation maps (CAM).[19] Specifically, this involved multiplication of the outputs of the model's last activation layer with the weights leading up to the model's top prediction class to yield a map with the model's most salient features. The model was set to optimise a categorical cross entropy loss on two output classes, enabling the heatmaps of the two classes to be visualised distinctly. The resultant derived map could then be overlaid on a given image to produce a heat map visualisation. An overall flow diagram of the study method is depicted in Figure 3.

Clinician group characteristics

The AI model performance on the test set was compared to the performance of a clinician group consisting of five[5] senior paediatric ED clinicians, whose job scope includes the initial interpretation of paediatric elbow radiographs performed at the point of care at the ED (pending the formal paediatric radiology report), and the subsequent management of these patients. These five clinicians had an average of 9 years' experience working in the ED and had obtained their postgraduate paediatric medicine qualifications for an average of 12 years. This clinical group was administered the same test set as the AI model (before resizing). This was presented to the clinicians in a slideshow presentation (Microsoft PowerPoint) [Figure 4], which was viewed on a computer monitor of similar resolution to that used in their regular practice. No time limit was set. Each slide presented a single case. The clinician had to indicate if the radiograph was normal or abnormal; if the image was abnormal, the clinician had to describe the abnormality and place a marker over its site. Note that only the designation of the radiograph as normal or abnormal was taken in the analysis; the placement of the marker was an attempt to encourage the clinician aim for accuracy rather than sensitivity of detection.

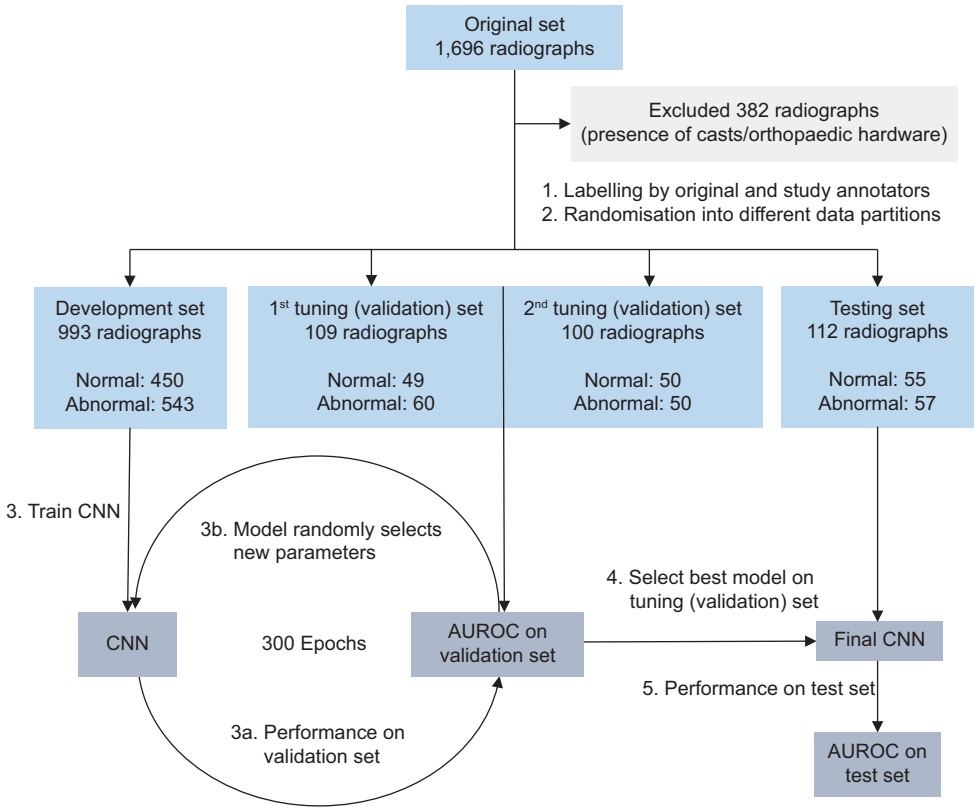

Figure 3: Flow diagram of the study methodology. AUROC: area under the receiver operating characteristic curve, CNN: convolutional neural network.

The clinicians took an average of 52 min to complete the test set. The inter-rater reliability of the clinical group was assessed using Fleiss' Kappa coefficients; Kappa values ≤0 indicate no agreement, values 0.40–0.75 indicate fair to good agreement and values >0.75 indicate excellent agreement. Among the five members of the clinical group, a composite score of summation of ratings at a cut-off value ≥3 was used to determine the classification status of a particular test set image by the group as either normal or abnormal. Statistical analysis on Kappa coefficients was performed using Stata version 16.1 (StataCorp LP, College Station, TX, USA).

Statistical analysis

The performance of the AI model on the test set in terms of sensitivity and specificity was compared to the binary classification aggregated from the five clinicians using McNemar test in R v4.1.0. A *P* value <0.05 was considered statistically significant. The ground truth for both AI model and clinician performance was the graded image (abnormal vs. normal) by the study team. The 95% confidence intervals (CIs) for AUROC were calculated using the DeLong method.

RESULTS

In this study of paediatric elbow radiograph classification, the accuracy of the model on the external held out validation set was 82.0% (95% CI: 73.1%–89.0%) and the AUROC was 0.896 (95% CI: 0.848–0.966). The accuracy of the model on the test set was 80.4% (95% CI: 71.8%–87.3%) and the AUROC was 0.872 (95% CI: 0.831–0.947). The performance of the model on the test set compared to the clinician group was as follows: sensitivity 79.0% (95% CI: 68.4%–89.5%) versus 64.9% (95% CI: 52.5%–77.3%), *P* value 0.088; specificity 81.8% (95% CI: 71.6%–92.0%) versus 87.3% (95% CI: 78.5%–96.1%), *P* value 0.439. Figure 5 shows the AUROC, sensitivity and specificity of the AI model compared to the sensitivity and specificity of the clinician group on the test

Figure 4: Diagram shows the format of test set presentation for the clinician group. The marker (red circle) is placed over the abnormality.

set. A composite score cut-off of 3 was used. The inter-rater agreement for clinician was fair (Fleiss' Kappa coefficient: 0.399, 95% CI: 0.340–0.457).

Class activation mapping performed on the test set images generally showed that the AI model focused, as expected, on the elbow joint on normal images and the area of pathology on abnormal images [Figure 6]. The misclassification of paediatric elbow radiographs by the AI model included cases of fracture, effusion and abnormal elbow position [Table 1].

DISCUSSION

In this study, we developed a CNN-based deep learning AI model for the binomial classification of paediatric elbow radiographs. The AUROC of the model was 0.872. This model on a test set demonstrated higher sensitivity, with the *P* value at nominal significance (0.08), when compared to the clinician group. The AI model showed inferior specificity, with the *P* value not significant (0.439), when compared to the clinician group.

This form of binomial classification models may have a role as a decision support tool for clinicians in the acute care setting and as a decision support algorithm for radiologists, for radiology worklist management and for identification of potentially abnormal radiographic images for expedited interpretation by a radiologist.[12] The performance of the model that we developed was, however, inferior to that of previously developed models, which showed an accuracy of 88%–91%, a sensitivity of 91%–93% and a specificity of 84%–92%.[11-14] The reasons for this are not immediately apparent; possible explanations for the variance may include different cases, set sizes and set hyperparameters. For example, in the study by England *et al*.,[12] the model was initialised from scratch using a DenseNet-BC architecture instead of initialisation with pretrained parameter weights from the ImageNet dataset, as done in our study, and the AI model resized the images presented to it at a 512 × 512 resolution compared to a 240 × 240 resized resolution in our study. In the study by Rayan *et al*.,[13] an Xception network architecture and a significantly larger dataset were used (58,817 images), and the images presented to it were resized to 500 × 500 resolution. In the study by Choi *et al*.,[14] a ResNet-50 network architecture was used and the AI model centred the olecranon process at the centre of the image at a 600 × 600 resized resolution. Although increased resolution may not always correlate with model performance owing to decreases in maximum batch size[20], other studies assessing the effect of image resolution on model performance have described improved model performance for detection of small and subtle features with increased image resolution.[21]

Several points are highlighted in this study. First, we assessed the performance of the AI model and clinical group in using a metric of sensitivity instead of accuracy, as this was deemed important in its planned deployment for ED radiograph triage. Second, the method of administration and testing of the reference clinical group should be carefully considered. In our study, we took care not to prime the clinical group on the breakdown of normal and abnormal cases in the test set. We also had the clinician indicate the abnormal diagnosis if present and place a marker over the site where the abnormality was seen, in an attempt to have the clinician aim for accuracy rather than sensitivity of detection as an outcome, as per their usual clinical practice. These factors should be taken into account in the design of trials that compare the performance of human raters to AI models. Finally, the AI model is potentially applicable in daily practice even if the source images are in high-resolution DICOM format, as these would be automatically converted to image files of size 240 × 240 pixels by the developed AI model as part of the model development pipeline.

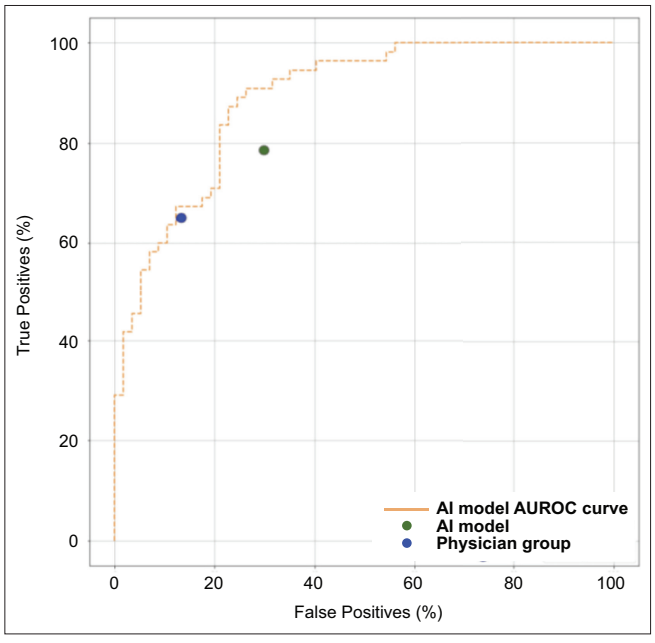

Figure 5: Graph shows artificial intelligence (AI) model area under the receiver operating characteristic curve (AUROC) versus physician (clinician) group, and AI model sensitivity and specificity on the test set.

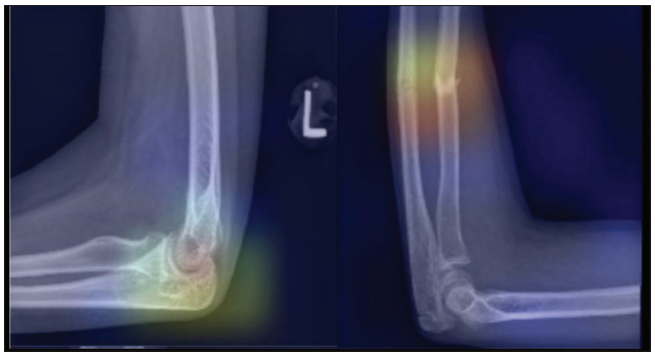

Figure 6: Class activation mapping heat map analysis of test images. The artificial intelligence model generally focuses on the elbow joint on normal images (left) and the area of pathology on abnormal images (right).

Table 1. Performance of AI model compared to the performance of clinical group on the test set and AI model misclassifications.

Variable	%			AUROC
	Sensitivity	Specificity	Accuracy	
Performance				
AI model	79.0 (68.40–89.5)	81.8 (71.6–92.0)	80.4 (71.8–87.3)	0.872 (0.831–0.947)
Clinical group	64.9 (52.5–77.3)	87.3 (78.5–96.1)	–	–
AI model misclassifications				
False negative	Ulna fracture (2), supracondylar fracture (1), missed elbow effusion (7)			
False positive	Abnormal elbow position (2), elbow effusion (8)			

AI: artificial intelligence, AUROC: area under receiver operating characteristics

This study had several limitations. First, the cases were retrieved from a single institution, potentially limiting its generalisability. Second, the prevailing circumstances and regulations by the designated IRB and the hospital at the time of this study allowed only the investigators access to anonymised radiographs of the elbow performed within the institution and obtained within a specified time frame; no identifiers were provided, and only the date the radiograph was acquired and the patient's age were available on the image. It was thus not possible to determine the exact number of patients included in the study and the gender distribution of the radiographs. Hence, the data partitions (i.e. the development, tuning [validation], and the test sets) were also disjoint at the image level and not at the patient level or higher in this study. Therefore, it was not possible to completely exclude that there could be different radiographs from the same patient in the different partitions, although this was deemed unlikely by the study team, given the nature of the data. Nonetheless, the radiographs were independently assessed as being normal or abnormal regardless of the origin of the patient, and there was no possibility that there was data overlap of a given radiograph among different datasets. These could be improved upon in future studies with evolution of the regulatory environment in local medical AI research. Third, a multiview approach to classification by the AI (i.e. analysis of anteroposterior [AP] and lateral projection radiographs), as performed in earlier studies,[13,14] was not possible in this study due to resource limitations. Fourth, the lower resized resolution of the model compared to other studies may affect the model's sensitivity to subtle abnormality. This may potentially warrant further study and include the use of more advanced versions of EfficientNet, such as the B6 and B7 models, which have higher intrinsic resolutions run upon GPU systems with higher processing power.[22] A program that centres the image at the olecranon, similar to the study by Choi *et al*.,[14] may also be employed. Fifth, despite our stated aims, the intrinsic limitations of an AI model developed on a small-sized development dataset should not be viewed lightly. The AI model specificity may have potentially been improved if a large sized development set was used, this should, however, be taken in the context of the real-world challenges of obtaining large volumes of high-quality annotated data. Using natural language processing tools to aid in the classification of large datasets may be a potential strategy for exploration.[13] Sixth, this study did not compare the EfficientNet model to existing architectures such as Resnet and InceptionNet, and it would be useful to perform these analyses in future work to assess the merit of this compound scaling network architecture compared to other established models. Seventh, Microsoft PowerPoint was used to administer the test set to the clinicians. This is suboptimal compared to the dedicated imaging viewing software used in clinical setting (Citrix Vuemotion, which has viewing tools such as panning and zooming), but it was not employed in the study due to technical factors regarding image conversion and test set creation. Eighth, a future study with more cases that is appropriately powered to evaluate the degree to which age-related changes affect test performance may be considered. Ninth, a future study could also explore the use of the musculoskeletal radiograph dataset by the Stanford University Medical Center dataset,[22] which contains a large number of elbow radiographs with binary labels indicating whether the image is normal or abnormal for pretraining of parameter weights of the model. Lastly, to improve results, future studies could consider further analysis of the composition of randomised sets and false-negative/-positive studies to assess for their characteristics, correlating with CAM, with retraining of the model based on these cases using augmentation or zooming means, as well as the use of object detection methods through subdomain decomposition or bounding box methods.[13]

In summary, a CNN-based deep learning AI model for the binomial classification of paediatric elbow radiographs was developed. The AI model showed good AUROC values and higher sensitivity with the *P* value at nominal significance when compared to a clinician group. The initial model developed in this study is planned for further refinement with a view for future deployment within the paediatric and adult hospital setting for image triage into normal and abnormal groups at the point of care by the attending clinician. The near- and mid-term plan for this project is to further improve the AI model with better model development techniques, to integrate the model into a frontline clinical platform, as well as to conduct a proof-of-value trial using the frontline clinical platform in a real-world scenario where the testing dataset

comprises scans that have be performed immediately before analysis by the AI model.

Acknowledgement

We would like to acknowledge Dr Choi Yoon Seong for her review and input on the manuscript.

Financial support and sponsorship

Funding was provided by the Singhealth Duke-NUS Radiological Sciences Academic Clinical Programme Clinical & Systems Innovation Support Grant.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Taves J, Skitch S, Valani R. Determining the clinical significance of errors in paediatric radiograph interpretation between emergency physicians and radiologists. CJEM 2018;20:420-4.
2. Goswami GK. The fat pad sign. Radiology 2022;222:419-20.
3. Norell HG. Roentgenologic visualisation of the extracapsular fat; Its importance in the diagnosis of traumatic injuries to the elbow. Acta radiol 1954;42:205-10.
4. Bledsoe RC, Izenstark JL. Displacement of fat pads in diseases and injury of the elbow: A new radiographic sign. Radiology 1959;73:717-24.
5. Donnelly LF, Klostermeier TT, Klosterman LA. Traumatic elbow effusions in paediatric patients: Are occult fractures the rule? AJR 1998;171:243-5.
6. Major NM, Crawford ST. Elbow effusions in trauma in adults and children: Is there an occult fracture? AJR 2002;178:413-8.
7. Morewood DJ. Incidence of unsuspected fractures in traumatic effusions of the elbow joint. Br Med J (Clin Res Ed) 1987;295:109-10.
8. Iyer RS, Thapa MM, Khanna PC, Chew FS. Paediatric bone imaging: Imaging elbow trauma in children – A review of acute and chronic injuries. AJR Am J Roentgenol 2012;198:1053-68.
9. Nakamura K, Hirachi K, Uchiyama S, Takahara M, Minami A, Imaeda T, *et al*. Long-term clinical and radiographic outcomes after open reduction for missed Monteggia fracture-dislocations in children. J Bone Joint Surg Am 2009;91:1394-404.
10. Rahbek O, Deutch SR, Kold S, Søjbjerg JO, Møller-Madsen B. Long-term outcome after ulnar osteotomy for missed Monteggia fracture dislocation in children. J Child Orthop 2011;5:449-57.
11. Offiah AC. Current and emerging artificial intelligence applications for paediatric musculoskeletal radiology. Pediatr Radiol 2022;52:2149-58.
12. England JR, Gross JS, White EA, Patel DB, England JT, Cheng PM. Detection of traumatic paediatric elbow joint effusion using a deep convolutional neural network. AJR Am J Roentgenol 2018;211:1361-8.
13. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A. Binomial classification of paediatric elbow fractures using a deep learning multi view approach emulating radiologist decision making. Radiol Artif Intell 2019;1:e180015
14. Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, *et al*. Using a dual-input convolutional neural network for automated detection of paediatric supracondylar fracture on conventional radiography. Investig Radiol 2020;55:101-10.
15. Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, *et al*. Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell 2020;2:e190023.
16. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. 2019. ArXiv, abs/1905.11946. [Last accessed on 2021 Nov 08].
17. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L, *et al*. ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition; 2009. p. 248-55.
18. ImageNet. 2009. Available from: http://image-net.org/index. [Last accessed on 2021 Jun 30].
19. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. p. 2921-9.
20. Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 1994;5:537-50.
21. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. Radiol Artif Intell 2020;2:e190015.
22. Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, *et al*. Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv: 1712.06957. 2017 Dec 11.