Developing Deep Learning Models for the Classification of Pediatric Elbow Radiographic Abnormalities of Comparable Performance to Physicians: Strategies for Model Optimization With Small Sized Development Sets

Mark B. TAN (marktanbangwei@gmail.com)
Singapore General Hospital
Russ Y. CHUA
Agency for Science, Technology and Research
Qiao FAN
Duke NUS Graduate Medical School
Marielle V. FORTIER
KK Women's and Children's Hospital
Pearlly P. CHANG
KK Women's and Children's Hospital

Research Article

Keywords: Artificial Intelligence, Pediatric Radiology, Emergency Radiology, Musculoskeletal Radiology, Machine Learning

Posted Date: January 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1199983/v1

Abstract

Background

To compare the performance of an AI model based on strategies designed to overcome small sized development sets to pediatric ER physicians at a classification triage task of pediatric elbow radiographs.

Methods

1,314 pediatric elbow lateral radiographs (mean age: 8.2 years) were retrospectively retrieved, binomially classified based on their annotation as normal or abnormal (with pathology), and randomly partitioned into a development set (993 images), tuning set (109 images), second tuning set (100 images) and test set (112 images). The AI model was trained on the development set and utilized the EfficientNet B1 compound scaling network architecture and online augmentations. Its performance on the test set was compared to a group of five physicians (inter-rater agreement: fair). Statistical analysis: AUC of AI model - DeLong method. Performance of AI model and physician groups - McNemar test.

Results

Accuracy of the model on the test set - 0.804 (95% CI, 0.718 - 0.873), AUROC - 0.872 (95% CI, 0.831 - 0.947). AI model performance compared to the physician group on the test set - sensitivity 0.790 (95% CI 0.684 to 0.895) vs 0.649 (95% CI 0.525 to 0.773), p value 0.088; specificity 0.818 (95% CI 0.716 to 0.920) vs 0.873 (95% CI 0.785 to 0.961), p value 0.439.

Conclusions

The AI model for elbow radiograph triage designed with strategies to optimize performance for a small sized development set showed comparable performance to physicians.

Key Points

An AI model designed using strategies to overcome small sized development sets shows performance comparable to senior pediatric emergency medicine physicians in pediatric elbow radiograph triage.

The AI model network architecture, including that of compound scaling architectures, is an important factor in AI model performance along with other factors i.e. development set size.

Background

Timely radiological interpretation in the emergency room (ER) is important for best patient outcomes. In our practice setting, despite the presence of national tertiary pediatric hospitals, pediatric patients on occasion present to adult hospitals for acute care. Radiologists and emergency medicine physicians who do not review pediatric imaging routinely may be less confident in interpreting such studies [1]. This may

be due to various causes, for example, misdiagnosis of ossification centers as fractures, or, unfamiliarity with fracture patterns in the pediatric population [2]. Our group thus aimed to develop an AI model for the triage of pediatric elbow radiographs, given the risk of morbidity i.e. deficits in range of motion should diagnosis be delayed [3, 4]. This model is planned for future deployment within the pediatric and adult hospital setting.

Developing AI models for image classification as a rule requires a large dataset of annotated radiological images for the AI algorithm to be trained on. Our group however encountered restrictions on the dataset size made available for study as there was no means in the institution at the time of exporting bulk annotated data from the hospital RIS-PACS (Radiological Information System - Picture Archive and Communication System) without incurring prohibitive cost. The size of the dataset made available was about 1000 cases, 1/10th the size of previously published studies [5].

Our group in response employed development strategies to optimize performance of AI models in a 'data contested environment' where the size of the AI development set is small. First, an approach of practicable efficacy of the AI model was adopted. As the AI model was to be used for triage, it was developed as a binomial classifier of radiographs labelled as normal and abnormal, and it was benchmarked to meet or exceed the performance of a senior pediatric ER physician at radiograph classification. Second, the EfficientNet [6] network architecture was employed in the model as this architecture through compound scaling methods had previously shown superior performance and decreased overfitting for sets with fewer training parameters as compared to other larger network architectures [6, 7, 8]. Online augmentations were also performed. We hypothesized that an AI algorithm developed based on these strategies would perform comparably to senior pediatric ER physicians at a classification triage task of pediatric elbow radiographs.

Methods

Data Acquisition, Ground Truth, Data Partitions

In this study the data source was the RIS-PACS system of a tertiary pediatric hospital. No previous studies on this dataset were performed. Project approval was obtained from the IRB and informed consent was waived. There was no conflict of interest.

The IRB authorized retrospective extraction from the RIS-PACS radiographs of both the elbows in lateral view, of male and female aged from 3-16 years (mean 8.2 years), in a consecutive series from January to November 2015. As the aim of the model was for triage of patients who presented with their first onset of elbow symptoms in the ER, radiographs which showed casts or orthopedic hardware were not extracted for use.

Of the 1,696 radiographs presented to the study team, 1,314 unique radiographs were extracted. Manual extraction of the entire image excluding identifiers without cropping was performed by a study team member (T.B.M.) who possessed radiology postgraduate qualifications (FRCR). The included studies

were then annotated as ground truth as either normal or abnormal (with pathology i.e. fracture, effusion, dislocation). These annotators all possessed postgraduate radiology qualifications (FRCR), and had been deemed qualified by the department to read out pediatric 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 (M.B.T.) would review and independently annotate the radiograph; if there was no discrepancy between the annotation of the first and the second annotator, the final annotation would follow their consensus opinion. In the event of discrepancy, the radiograph would be referred to the third annotator (M.V.F.), a specialist pediatric radiologist, for adjudication and designation of final annotation.

These images were separated into different partitions of 993 images in a development set (450 normal and 543 abnormal images), 109 images in a tuning set (to prevent model over-fitting, 49 normal and 60 abnormal, representing 10% of the number of images in the development set), 100 images in a second tuning set (50 normal and 50 abnormal images), and 112 images in a test set (which was not exposed to the model in its development and tuning, 57 abnormal and 55 normal images). This testing sample size was determined by what was reasonably achievable by a physician during a test sequence, a practice consistent with previous studies [9].

Model Development

Each image was downsampled to 240 x 240 pixels. The tensorflow-gpu 2.3.1 built-in function: tf.keras.preprocessing.image.ImageDataGenerator was used. The EfficientNet B1 model was used [6]. These versions of the following software were employed: EfficientNet v1.1.1, Keras-Applications v1.0.8, Keras-Preprocessing v1.1.2, Scikit-image v0.17.2, Scikit-learn v0.24.1, Scipy v1.5.3, Tensorboard v2.3.0, Tensorboard-plugin-wit v1.7.0, Tensorflow-estimator v2.3.0, Tensorflow-gpu v2.3.1. A NVIDIA RTX 2060 GPU was used. The parameters of the model were initialized with pre-trained weights from the ImageNet dataset [10, 11] as a further strategy to combat overfitting in this relatively small dataset [12]. None of the model parameters were frozen due to the intrinsic differences of the images of the ImageNet dataset from the development dataset. The online augmentations employed on the development dataset were random rotations (10°), flips, positional shifts (0.05) and affine transformations (zoom, 0.05), applied to each epoch. 300 Epochs with a minimum patience of 50 epochs were run. An Adam optimizer was used with a 1e-4 learning rate without a scheduler. The Batch size was 8. The model optimized a categorical cross entropy loss on 2 output classes enabling the heatmaps of the two classes to be visualized distinctly. The best performing model out of the total number of training epochs was selected based on the model which had the best accuracy on the second tuning set, which in our case also had the highest AUROC. To evaluate if our model was making predictions on the right pathological features and not artifacts within the elbow radiograph image, we produced Class Activation Maps [13] (CAM). Specifically, we introduced an image into our network and multiplied 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 most salient features, and then overlaid this map on our images to produce a heat map visualization (Figure 3).

Statistical Analysis

The AI model performance on the test set was compared to a clinical group of five (5) senior pediatric ER physicians who as a group had performed their job role and obtained their postgraduate pediatric medicine qualifications for an average of 9 years and 12 years respectively. This clinical group was administered the same test set as the AI model, which was presented to the physicians in the form of a slideshow presentation (Microsoft PowerPoint) which the physician could navigate through (Fig. 1) on a monitor of similar resolution to that of their usual practice. No time limit was set, each slide presented a single case, and the subject was to indicate if the radiograph was normal or abnormal, and if the image was abnormal, the subject was to describe the abnormality as well as place a marker over its site. Note that only the designation of the radiograph as normal or abnormal was taken in the analysis, with the marker placement done in an attempt to have the physician aim for accuracy rather than sensitivity of detection. The physicians took an average of 52 minutes 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, 0.40 to 0.75 as fair to good, and over 0.75 as excellent. Among the 5 members of the clinical group, a composite score of summation of ratings at a cut-off value of ≥ 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 the statistical software STATA (version 16.1).

<Figure 1>

Figure 1: Format of test set presentation for physicians.

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 5 physicians using McNemar test in R v4.1.0 [14]. A P value < 0.05 was considered statistically significant. The ground truth for both AI model and physicians' performance was the graded image (abnormal vs. normal). The 95% confidence intervals (CI) for AUC were calculated using the DeLong method. The study statistician was F.Q.

Results

In this study of pediatric elbow radiograph classification, the accuracy of the model on the external held out validation set was 0.820 (95% confidence interval [CI], 0.731 - 0.890) and the AUROC was 0.896 (95% CI, 0.848 -0.966). The accuracy of the model on the test set was 0.804 (95% CI, 0.718 - 0.873) and the AUROC was 0.872 (95% CI, 0.831 - 0.947).

The performance of the model on the test set compared to the physician group was: sensitivity 0.790 (95% confidence interval 0.684 to 0.895) vs 0.649 (95% confidence interval 0.525 to 0.773), p value 0.088; specificity 0.818 (95% confidence interval 0.716 to 0.920) vs 0.873 (95% confidence interval 0.785 to 0.961), p value 0.439; A composite score cut-off of 3 was used (Fig. 2). The inter-rater agreement for physicians was fair (Fleiss' Kappa coefficient, 0.399, 95% CI, 0.340 to 0.457). The AI model compared to

the physician group achieved superior sensitivity with the p-value at nominal significance (0.08) although specificity was inferior.

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 on the area of pathology on abnormal images (Fig. 3).

Conclusions

These results generally prove the hypothesis that the performance of an AI model developed based on the aforementioned strategies of overcoming data contested environments is comparable to senior pediatric emergency medicine physicians in pediatric elbow radiograph triage. This study complements the earlier study on binomial classification of pediatric elbow fractures by Rayan et al [5], with our study utilizing strategies for model development in a more data contested and resource limited environment, as well as human annotators as opposed to natural language processing for image curation.

Several points to this study were identified. 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 ER radiograph triage. Second, in this paper we used an EfficientNet B1 model with 240 x 240 resolution. Recognizing the obvious differences in the set characteristics, this model using a lower resolution model and smaller dataset size nonetheless managed respectable results as compared to other studies which used higher resolution images and dataset size [5]. The reasons for this would require further study but it may be conceptually important that the performance of the model depends on factors other than image resolution [15] or set size alone, with the network architecture possibly also an important factor contributing to model performance. The EfficientNet [6] family of models has shown among other Convolutional Neural Networks efficacy in terms of performance and speed using commercially available GPU processing capabilities in the classification of skin lesions [16], CT lung scans [17] and diabetic retinopathy [18] but this is the probably one of the first papers employing this model in paediatric elbow radiographs. In this study, a lower powered B1 version of the model was employed as compared to higher (i.e. B4 to B7) versions due to limitations in processing power, directions for a future study may include comparing the relative performance of higher powered versions of the EfficientNet model. Third, 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 physicians indicate the abnormal diagnosis if present and place a marker over the site where the abnormality was seen in an attempt to have the physician 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 which compare the performance of human raters to AI models.

There were a number of limitations of this study. First, the cases were retrieved from a single institution potentially limiting its generalizability. Second, a multiview approach to classification by the AI (i.e.

analysis of AP and lateral projection radiographs) as performed in earlier studies [5] was not possible in this study due to resource limitations. Third, the resolution of the model was also below the standard resolution of a radiograph which may affect the model's sensitivity to subtle abnormality [9, 15], this may be potentially overcome through using more advanced versions of EfficientNet i.e. B6 and B7 models [19]. Fourth, despite the stated aims, the intrinsic limitations of an AI model developed on a small sized development dataset should also not be lightly viewed. The AI model specificity may have potentially improved with using a larger sized development set; this should however be taken in the context of the real world challenges in obtaining large volumes of high quality annotated data.

We note that our study had limitations in development set size not seen in previous papers. However, the prevalence of such data-contested environments may not exactly be rare in the setting of supervised learning, currently the standard for image classification tasks, as the burden of accurate data annotation is a perennial limiting factor. In summary, this study shows that an AI algorithm developed based on strategies of overcoming data contested environments has value in creating clinically relevant models.

Declarations

Ethics approval and consent to participate: Project approval was obtained from the Singhealth Centralised Institutional Review Board (Singapore) and informed consent was waived. All methods were performed in accordance with the relevant guidelines and regulations of the Institutional Review Board. There was no conflict of interest.

Consent for Publication: Not applicable

Availability of data and materials: The data that support the findings of this study are available from the corresponding author, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the corresponding author.

Competing interests: The authors declare that they have no competing interests

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

Authors' contributions: MBT directed the conduct of the study, performed data collection and analysis and prepared the final manuscript. RYC developed the AI model. FQ performed statistical calculations and analysis on the data. MVF performed data analysis. PPC supervised the clinician arm of the study. All authors read and approved the final manuscript.

Acknowledgements: The authors would like to acknowledge Dr. Choi Yoon Seong for her review and inputs on the manuscript.

References

1. Taves J, Skitch S, Valani R Determining the clinical significance of errors in pediatric radiograph interpretation between emergency physicians and radiologists. CJEM. 2018; May;20(3):420-424.

2. Iyer RS, Thapa MM, Khanna PC, Chew FS. Pediatric bone imaging: imaging elbow trauma in children–a review of acute and chronic injuries. AJR Am J Roentgenol. 2012; May;198(5):1053-68.

3. Nakamura K, Hirachi K, Uchiyama S et al. Long-term clinical and radiographic outcomes after open reduction for missed Monteggia fracture-dislocations in children. J Bone Joint Surg Am. 2009; Jun;91(6):1394-404.

4. 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; Dec;5(6):449-57. doi: 10.1007/s11832-011-0372-0. Epub 2011 Oct 16.

5. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A. Binomial Classification of Pediatric Elbow Fractures Using a Deep Learning Multiview Approach Emulating Radiologist Decision Making. 2019; Radiol Artif Intell. Jan 30;1(1):e180015.

6. Tan, M., & Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2019. *ArXiv, abs/1905.11946*. Accessed 8 Nov 2021.

7. Image Classification on ImageNet. 2021. https://paperswithcode.com/sota/image-classification-on-imagenet. Accessed 8 Nov 2021.

8. Tan MX, Quoc VL. EfficientNet: Improving accuracy and efficiency through AutoML and Model Scaling. 2019. https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html accessed 8 Nov 2021.

9. Krogue JD, Cheng KV, Hwang KM et al. Automatic Hip Fracture Identification and Functional Subclassification with Deep Learning. 2020; Radiol Artif Intell. Mar 25;2(2):e190023.

10. J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255

11. ImageNet. (n.d.). http://image-net.org/index. Accessed 30 June 2021.

12. Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. 2014; *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 512-519.

13. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. 2016. IIn Proceedings of the IEEE conference on computer vision and pattern recognition (pp.

2921-2929)

14. Zhou X, Obuchowski N, McClish D. Statistical Methods in Diagnostic Medicine. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2011.

15. Sabottke CF, Spieler BM. The Effect of Image Resolution on Deep Learning in Radiography. Radiol Artif Intell. 2020; Jan 22;2(1):e190015.

16. Gessert N, Nielsen M, Shaikh M, Werner R, Schlaefer A. Skin lesion classification using ensembles of multi-resolution EfficientNets with metadata. 2020; MethodsX. Mar 19;7:100864.

17. Lawton S, Viriri S. Detection of COVID-19 from CT Lung Scans Using Transfer Learning. Comput Intell Neurosci. 2021; Apr 8;2021:5527923. doi: 10.1155/2021/5527923.

18. Pak, A., Ziyaden, A., Tukeshev, K., Jaxylykova, A., & Abdullina, D.. Comparative analysis of deep learning methods of detection of diabetic retinopathy. Cogent Eng. 2020; *7*(1), 1805144. https://doi.org/10.1080/23311916.2020.1805144

19. Y Fu. Image classification via fine-tuning with EfficientNet. 2020. https://keras.io/examples/vision/image_classification_efficientnet_fine_tuning/ Accessed 30 June 2021.

Figures

Figure 1

Format of test set presentation for physicians.

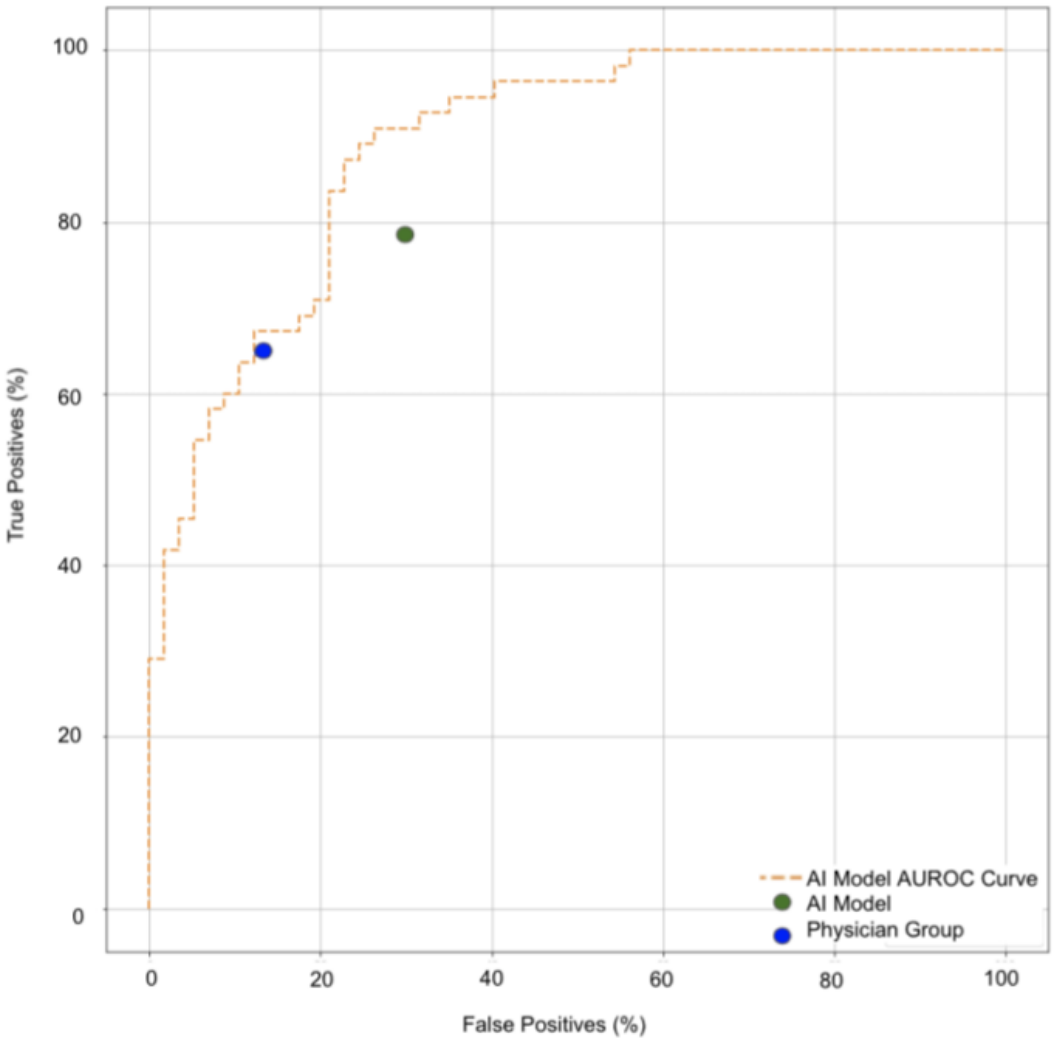

Figure 2

AI Model AUROC curve vs Physician Group and AI model sensitivity and specificity on the test set

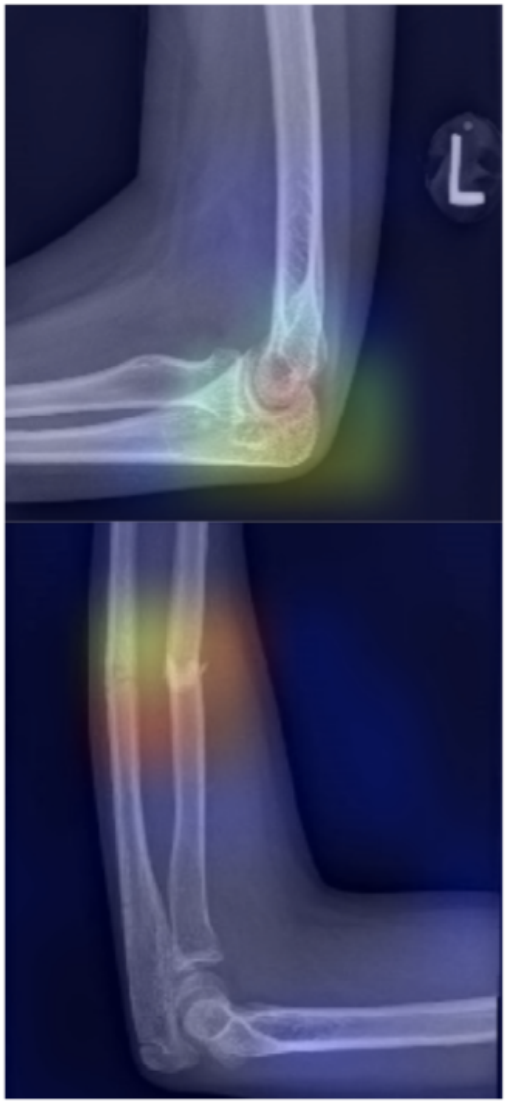

Figure 3

Class activation mapping heat map analysis of test images. The AI model generally focuses as expected on the elbow joint on normal images (top image), and on the area of pathology on abnormal images (bottom image).