Avalanche decision schemes to improve pediatric rib fracture detection

Jonathan Burkow[a], Gregory Holste[a], Jeffrey Otjen[b], Francisco Perez[b], Joseph Junewick[c], and Adam Alessio[a]

[a]Michigan State University, East Lansing, MI, United States
[b]Seattle Children's Hospital, Seattle, WA, United States
[c]Spectrum Health, Grand Rapids, MI, United States

ABSTRACT

Rib fractures are a sentinel injury for physical abuse in young children. When rib fractures are detected in young children, 80-100% of the time it is the result of child abuse. Rib fractures can be challenging to detect on pediatric radiographs given that they can be non-displaced, incomplete, superimposed over other structures, or oriented obliquely with respect to the detector. This work presents our efforts to develop an object detection method for rib fracture detection on pediatric chest radiographs. We propose a method entitled "avalanche decision" motivated by the reality that pediatric patients with rib fractures commonly present with multiple fractures; in our dataset, 76% of patients with fractures had more than one fracture. This approach is applied at inference and uses a decision threshold that decreases as a function of the number of proposals that clear the current threshold. These contributions were added to two leading single stage detectors: RetinaNet and YOLOv5. These methods were trained and tested with our curated dataset of 704 pediatric chest radiographs, for which pediatric radiologists labeled fracture locations and achieved an expert reader-to-reader F2 score of 0.76. Comparing base RetinaNet to RetinaNet+Avalanche yielded F2 scores of 0.55 and 0.65, respectively. F2 scores of base YOLOv5 and YOLOv5+Avalanche were 0.58 and 0.65, respectively. The proposed avalanche inferencing approaches provide increased recall and F2 scores over the standalone models.

Keywords: pediatric imaging, radiology, deep learning, rib fractures, object detection

1. INTRODUCTION

The abuse and neglect of children is a widespread, persistent issue. The Department of Health & Human Services Children's Bureau estimates that approximately $650,000$ children were victims of abuse and neglect yearly between 2015-2019 in the United States.[1] The presence of fractures in children has proven to be highly predictive of child abuse. The most common location of bone fractures stemming from abuse is the ribs, representing over 70% of fractures compared with skull, humeral, or femoral fractures.[2] When rib fractures are detected in young children, studies have shown that 80-100% of the time it is the result of child abuse.[3]

It is challenging to detect rib fractures on pediatric x-rays. In a review of chest radiographs of 550 abused infants and children, it was found that over two-thirds of rib fractures were missed during the first reads.[4] In addition to the inherent challenge of reading these exams, there is a decreasing availability of radiologists to interpret these studies. While the raw count of trained radiologists increased 39.2% from 1995 to 2011, the ratio of radiologists to general physicians has decreased from 4.0% to 3.7% in the same time.[5] This lower ratio of specialized radiologists poses a significant risk to the ability of detecting rib fractures in the taken radiographs and the possibility of detecting cases of child abuse through them. However, recent technological advancements in deep learning for computer vision tasks are now making it possible to augment the capabilities of trained radiologists to both ease their workloads and improve the detection rates of these fractures.

In this paper, we present among the first application of deep learning for rib fracture detection on pediatric chest radiographs. Several efforts have developed machine-learned methods to detect rib fractures in adult patients using different modalities. Zhang *et al.*[6] found that utilizing deep learning as either a concurrent or

burkowjo@msu.edu; aalessio@msu.edu; http://www.egr.msu.edu/~aalessio/

secondary reader to a trained radiologist on CT images yielded significantly higher sensitivity to unassisted readings. Another study utilizing Faster R-CNN and YOLOv3 for rib fracture detection on CT images saw radiologist performance with CNN assistance improved precision by 10% and sensitivity by 24%.[7] Other works include additional segmentation via a U-Net to further refine locality predictions of rib fractures.[8,9] These works primarily utilize volumetric 3D CT-scans or slices extracted from them. Our work deviates from this by utilizing single-view, anterior-posterior 2D chest radiographs for detection.

The data used for this study is a custom dataset we have curated with expertly annotated bounding box labels from board-certified pediatric radiologists for CNN model training and evaluation. The main technical contribution of this work is the use of a novel avalanche decision scheme applied at inference time, inspired by the clinical knowledge that patients presenting with rib fractures are highly likely to have more than one fracture. The new contribution is compared against the baseline model performance of RetinaNet[10] and YOLOv5.[11] We show that the proposed avalanche decision schemes offer detection performance improvements, particularly in terms of recall and F2 score.

2. METHODS

We propose an object detection method tailored for detection of rib fractures in pediatric anterior-posterior chest radiographs. The general processing pipeline for inference is: original DICOM radiograph file → thoracic segmentation (based on a U-Net) → thoracic cropping → convolutional neural network (CNN) based detection. All methods were compared in terms of conventional detection metrics (precision, recall) and F2 score, which is a variant of F1 score with a greater emphasis placed on recall over precision.

2.1 Rib Fracture Dataset

Our dataset was collected through an IRB-approved study at Seattle Children's Hospital. The dataset contains 704 unique patients, of which 515 are fracture present and 189 are fracture absent. There are 241(34.2%) female and 463(65.8%) male patients. The average age of patients is 268.76 ± 784.93 days (range $0 - 6935$, median 84, IQR 196). After removing outliers (missing, $age = 0$, or $age \geq Q3 + 1.5IQR$), the average age of patients is 128.11 ± 111.43 days (range $1 - 476$, median 84, IQR 140). The images are chest radiographs in an anterior-posterior perspective, provided in DICOM file format.

Ground truth annotations of rib fracture locations were provided by eight board-certified pediatric radiologists. When grading the fracture present section of the dataset the radiologists were given prior knowledge that at least one fracture was present in each image, thus there is a slight bias towards the labeling performance of the radiologists. All studies in the fracture absent section were cleared by the radiologists as being fracture free with no prior knowledge of either presence or absence of rib fractures.

2.2 Deep Learning Models

In this work we compare two single-stage object detector models, RetinaNet[10] and YOLOv5,[11] and the effect of our proposed avalanche decision scheme on their performance. RetinaNet is a leading object detector that includes a feature pyramid network backbone that feeds into two subnets responsible for classification and bounding box regression.[10] RetinaNet introduced the use of focal loss to overcome class imbalance common in object detection tasks. We adapted an open-source RetinaNet implementation[12] using PyTorch[13] to train all of our RetinaNet models. We maintained the standard ResNet-50 backbone with pre-trained weights from ImageNet. Each RetinaNet model used an early-stopping algorithm to stop training if validation set performance did not improve within five epochs.

The YOLO method has evolved since its first iteration introduced in 2016.[14] A major improvement came with YOLOv3 that used a feature pyramid network to improve performance across multiple scales of objects.[15] In addition to the improvements in version 1 through 4, YOLOv5 includes mosaic data augmentation to help with small training volumes. We adapted Ultralytics' open-source YOLOv5 implementation[11] and used the large (L6) model size pre-trained on COCO to train all YOLOv5 models on our dataset.

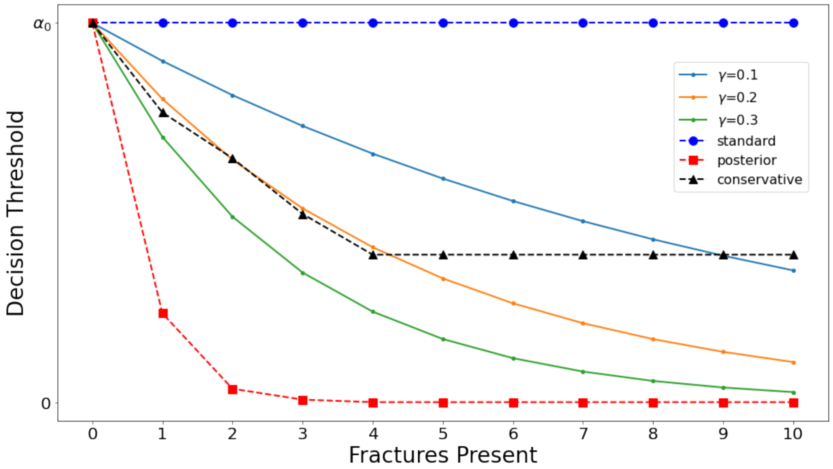

Figure 1. Visual representation of the dynamic changes in decision threshold for bounding box acceptance as a function of the number of proposed fractures.

2.3 Avalanche Decision

We investigated making the decision threshold for fracture-present model predictions a function of the number of already detected bounding box proposals. In other words, the decision threshold is not fixed, but rather changes depending on the number of high probability proposed regions. The motivation for this technique is that, in this medical application, if a patient has one fracture they are very likely to have more than one fracture, and a patient with two fractures is very likely to have three, and so on. This reality is highlighted in Tab. 1 which summarizes the posterior likelihood for fractures in our training images.

Table 1. Posterior likelihood summarizing the probability of more fractures given at least X fractures in the training dataset. This provides motivation for proposed avalanche decision scheme.

X $\geq$ x Fractures	N	$\mathbb{P}(X > x \mid X \geq x)$
X $\geq$ 1	382	76.4%
X $\geq$ 2	292	84.2%
X $\geq$ 3	246	77.2%
X $\geq$ 4	190	78.4%

For this method, we start with a base threshold α_0 and then reduce that threshold as a function of the number of proposals that clear that threshold. Starting from a higher threshold, this leads to cascading the threshold down which we denote as the avalanche decision scheme.

In order to determine how to adjust the decision threshold as a function of number of cleared proposals, we used the training dataset to calculate the probabilities of there being more proposals given that at least X fractures are currently present in the images, i.e., $P(X > 1|X \geq 1), \ldots, P(X > 4|X \geq 4)$ as presented in the third column of Tab. 1. Then, for a given starting model threshold α_0, if the model predicted 1 bounding box with a probability greater than α_0, we scale down the threshold to $\alpha_1 = \alpha_0 \cdot (1 - P(X > 1|X \geq 1))$ and the number of bounding box predictions that clear this threshold will be re-evaluated. If now three proposals have probabilities greater than α_1, we scale the threshold down to $\alpha_3 = \alpha_1 \cdot (1 - P(X > 2|X \geq 2) \cdot (1 - P(X > 3|X \geq 3)$.

We explored an alternative calculation where if one proposal was found in an image at the given starting threshold α_0, the next threshold to re-evaluate proposals would be $\alpha_1 = \alpha_0 \cdot P(X > 1|X \geq 1)$. This is a more conservative reduction in confidence thresholds, since with the prior calculation each successive threshold α_{i+1} would be approximately $16-24\%$ of α_i. With the new calculation, each α_{i+1} would be $76-84\%$ of α_i. To further investigate variants of this more moderate version, we tested cases where the decrease between each successive

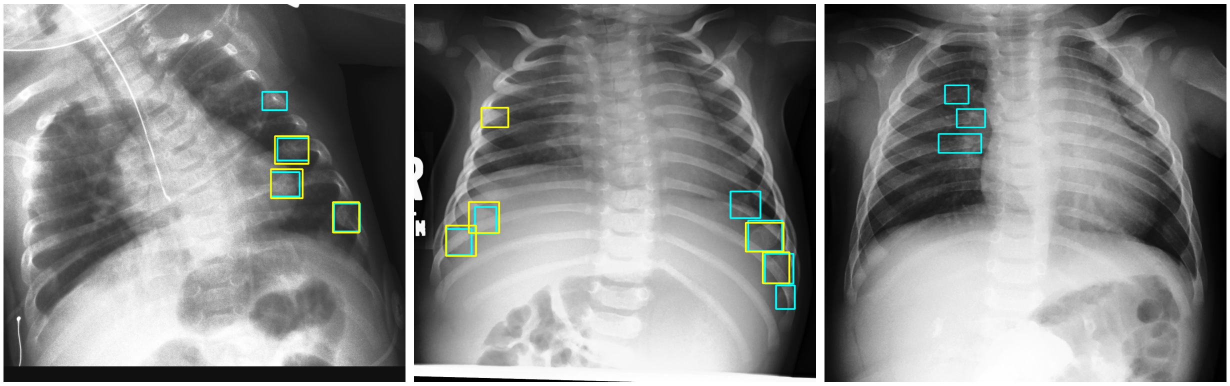

Figure 2. Example test set images with model predictions (yellow) and ground truth (teal) annotations, demonstrating examples of true positives, false positives (middle) and false negatives (all three).

α_i threshold was a constant rate γ, ranging from $10-30\%$. For example, if three rib fractures were proposed for a given starting threshold α_0, the new threshold will reduce to $\alpha_3 = \alpha_2 \cdot \gamma = \alpha_0 \cdot \gamma^3$.

Figure 1 provides a visual of the standard, posterior, and conservative schemes as well as three representative schemes with constant γ reduction. This illustrates the posterior avalanche scheme has the tendency to severely lower the decision threshold as soon as one, or especially two, rib fractures are proposed by the architectures that exceed the starting threshold α_0.

3. RESULTS

We split our fracture dataset into 75%/15%/10% training/validation/test sets, respectively. Example test set images with ground truth annotations and model predictions are presented in Fig. 2.

3.1 Inter-Reader Performance

In order to determine a baseline level of human performance to evaluate the trained CNN models against, we explored inter-reader variability among the expert radiologists and have presented a subset of these results previously.[16] An arbitrary portion of the fracture present images (N=195) were labeled by two separate board-certified radiologists. The first readers marked 4.83 fractures on average per image (range 1-20, median 4, IQR 5). Second readers marked an average of 4.76 fractures per image (range 1-27, median 4, IQR 4). Holding the first reader's boxes as pseudo-ground truth, the second readers overlapped the first reader's bounding boxes by 0.792 on average across all images. The average intersection-over-union was lower at 0.617.

Bounding boxes marked by both readers are considered to be true positives, or in concordance with one another, if the intersection-over-union was at least 0.30. With this threshold, 714 bounding boxes were true positives, 215 false positives (marked by reader 2 but not reader 1), and 227 false negatives (marked by reader 1 but not reader 2). This led to estimated expert reader-to-reader precision of 0.769, recall of 0.759, and F2 score of 0.760. See Appendix A for the derivation of the F2 score which is considered our primary target metric for this application.

3.2 Base Network Performance

RetinaNet achieved a precision score of 0.913 and YOLOv5 achieved 0.855, both well in excess of the inter-reader performance. However, recall values for both are much lower than the inter-reader performance, only reaching 0.502 and 0.541, respectively. This demonstrates that while the basic architectures have relatively few false positives, but they end up missing half of all rib fractures present in the test set.

Table 2. RetinaNet and YOLOv5 results applying the standard, fixed decision threshold, and avalanche schemes. γ represents the constant rate reduction between each decision threshold in the avalanche scheme.

RetinaNet				YOLOv5			
Scheme	Precision	Recall	F2	Scheme	Precision	Recall	F2
Standard	**0.913**	0.502	0.552	Standard	**0.855**	0.541	0.584
Posterior	0.100	**0.760**	0.328	Posterior	0.597	**0.659**	**0.646**
Conservative	0.574	0.677	**0.653**	Conservative	0.733	0.611	0.632
$\gamma = 0.15$	0.409	0.664	0.590	$\gamma = 0.15$	0.725	0.598	0.620
$\gamma = 0.20$	0.282	0.576	0.477	$\gamma = 0.20$	0.730	0.520	0.551

3.3 Avalanche Decision

Figure 3 shows F2 score performance of the RetinaNet model as the initial decision threshold α_0 changes, using the "standard" approach of a constant threshold to get all proposals (blue dashes with circles), the posterior distribution avalanche scheme (black dashes with triangles), the conservative scheme (red dashed with squares), and constant γ reduction schemes with $\gamma \in [0.10, 0.15, 0.20, 0.25, 0.30]$. This figure influenced decisions for the γ values to use in the avalanche scheme tests below.

Based on the starting threshold α_0 that obtained the highest F2 score for $\gamma = 0.15$ and $\gamma = 0.20$ in Fig. 3, we chose $\alpha_0 = 0.55$ and $\alpha_0 = 0.75$ for the constant reduction avalanche schemes, respectively, to test alongside the posterior and conservative decision schemes. Applying these schemes to RetinaNet and YOLOv5 detection architectures shows improved recall performance at the expense of reduced precision, shown in Tab. 2. While normal RetinaNet attained a recall of 0.502, each avalanche scheme applied improves recall performance to between 0.576 and 0.760. However, precision drops from a high of 0.913 to as low as 0.100 across all models, leading to drops in F2 performance in all but the conservative scheme. With this scheme the F2 value increases by 0.101 to 0.653.

In contrast, applying the avalanche schemes to YOLOv5 does not incur as dramatic changes in precision as RetinaNet. Recall performance increases more modestly than the avalanche schemes on RetinaNet, but drops in precision are much smaller between 0.122 and 0.258. This causes a F2 score improvement in all schemes by at least 0.039. This is possibly due to the YOLOv5 architectures' tendency to be more reserved in bounding box proposals as compared to RetinaNet; for instance, the stock RetinaNet model proposed nearly 2,500 boxes whereas stock YOLOv5 proposed only 310 for a single image. Thus, the reduction in the acceptance threshold applies much more heavily to RetinaNet, causing the severe decline in precision versus YOLOv5. This

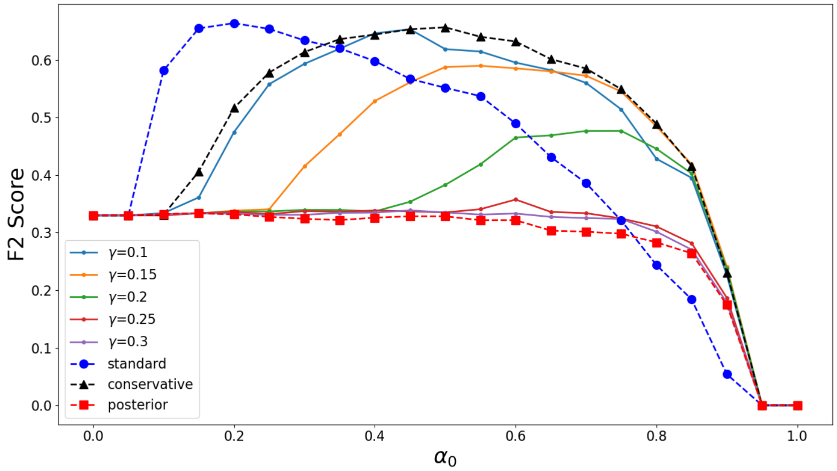

Figure 3. F2 scores for various avalanche decision schemes (as described in Sec. 2.3) applied on RetinaNet across all possible starting decision thresholds α_0.

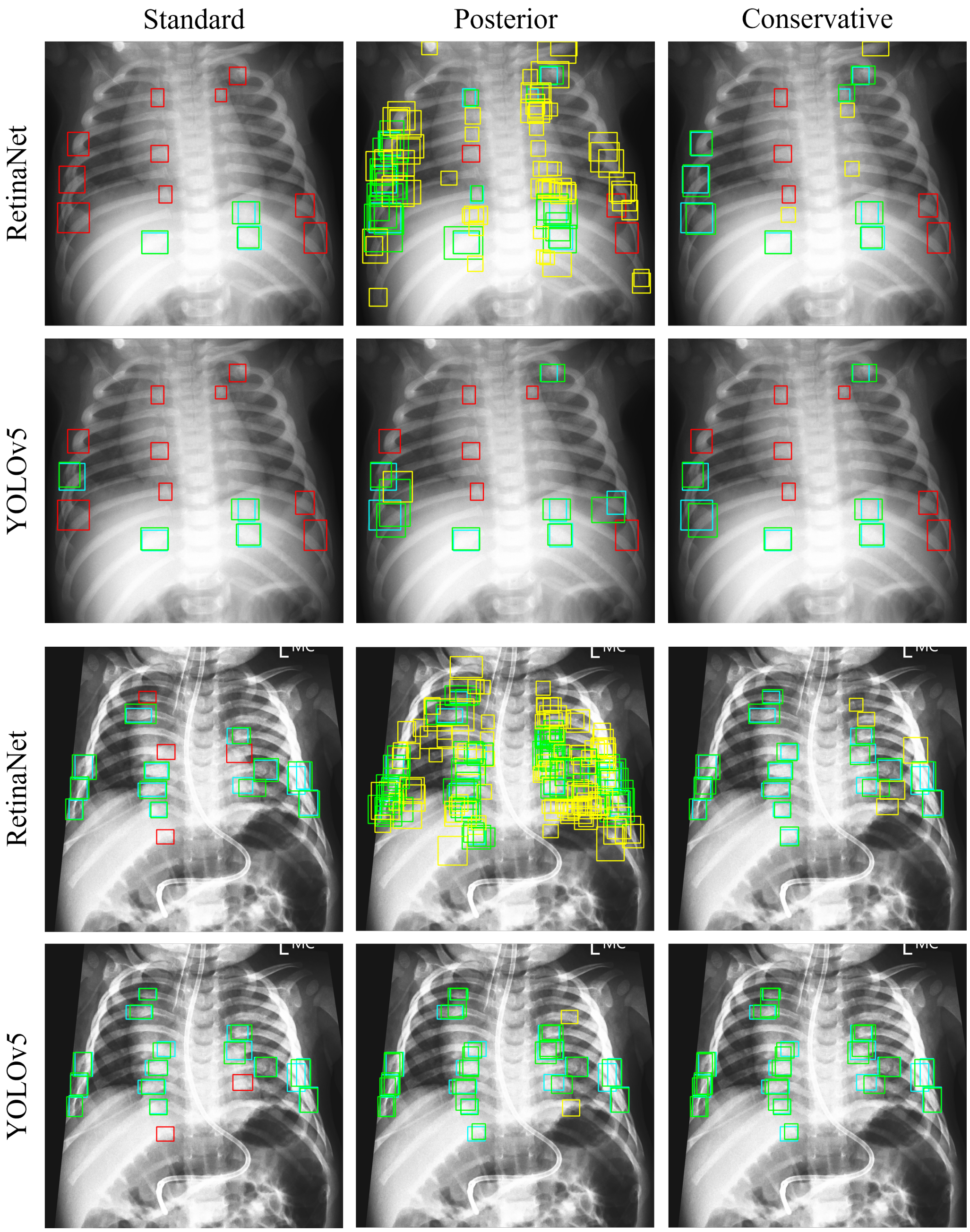

Figure 4. Representative images from two patients with the standard inference as well as posterior and conservative avalanche schemes applied to both RetinaNet and YOLOv5. Ground truth boxes are represented in teal (if a corresponding model prediction exists) or red (no matched prediction). Bounding box proposals from the trained networks are shown in green (true positive match with ground truth) or yellow (false positive).

phenomenon is easily seen in Fig 4, where the images for RetinaNet with the posterior avalanche scheme is littered with proposed objects (many false positives), whereas the corresponding YOLOv5 images contain much fewer additional bounding box proposals as compared to the standard scheme.

4. CONCLUSION

In this paper, we use a curated dataset of pediatric radiographs to perform localized detection of rib fractures using the RetinaNet and YOLOv5 object detection architectures. To our knowledge, we are among the first to adapt these deep learning object detectors for rib fracture detection in pediatric radiographs. To improve the capabilities of the two stock architectures, we implemented a novel avalanche decision scheme inspired by realistic domain knowledge that applies a decision threshold that decreases as a function of the number of accepted objects in each image. If this work is extended to additional deep learning architectures, care must be taken on which scheme to apply to each deep learning model as some models are can be more sensitive (more liberal with bounding box proposals) as demonstrated with performance trade-offs with RetinaNet. However, it can also be seen that many bounding box proposals from RetinaNet using the posterior scheme overlap; therefore an easy extension to this work would be to implement a non-maximum suppression to reduce purely overlapping bounding boxes to mitigate the effect of the posterior scheme on these models with more numerous proposals. Overall, while precision reduced moderately, these proposed avalanche schemes offered higher recall and ultimately higher F2 scores suggesting that they are viable approaches for this application.

ACKNOWLEDGMENTS

This work is supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the NIH under Award Number R21HD097609. We would like to thank the radiologists that contributed their time and expertise to provide this work with ground truth rib fracture annotations.

APPENDIX A. F2 SCORE

A common evaluation metric used in classification and detection tasks is F1 score, which is the harmonic mean between precision and recall.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

This is a specific implementation of a general F_β score where both precision and recall are considered equally as important in the final calculation. The constant term β can be changed to weight precision or recall β-times more than the other. For instance, if one wants recall to carry β-times as much importance as precision in the calculation, the equation becomes

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{2}$$

If instead one desired to weight precision β-times more than recall, the denominator would change to $(\text{precision} + (\beta^2 \cdot \text{recall}))$. For our evaluation, we are placing a larger emphasis on recall over precision, as we want the deep learning models to be liberal with predicting potential fracture locations even if the predictions actually do not contain fractures. In the long term, this tool is intended to serve as a reading aid for radiologists to flag suspicious regions; the final determination of the presence or absence of fractures will be determined by the radiologist aided by the model output. For these reasons, we are predominantly evaluating all models by F2 score and placing twice as much weight on recall as precision, i.e., setting $\beta = 2$ in Eq. (2).

REFERENCES

[1] Kelly, C., Street, C., and Building, M. E. S., "Child Maltreatment 2019," *Child Maltreatment* , 306 (2019).

[2] Kemp, A. M., Dunstan, F., Harrison, S., Morris, S., Mann, M., Rolfe, K., Datta, S., Thomas, D. P., Sibert, J. R., and Maguire, S., "Patterns of skeletal fractures in child abuse: systematic review," *BMJ* **337**, a1518 (Oct. 2008). Publisher: British Medical Journal Publishing Group Section: Research.

[3] Darling, S. E., Done, S. L., Friedman, S. D., and Feldman, K. W., "Frequency of intrathoracic injuries in children younger than 3 years with rib fractures.," *Pediatric radiology* **44**(10), 1230–1236 (2014).

[4] Merten, D. F., Radkowski, M. A., and Leonidas, J. C., "The abused child: a radiological reappraisal.," *Radiology* **146**, 377–381 (Feb. 1983). Publisher: Radiological Society of North America.

[5] Rosenkrantz, A. B., Hughes, D. R., and Duszak, R., "The U.S. Radiologist Workforce: An Analysis of Temporal and Geographic Variation by Using Large National Datasets," *Radiology* **279**, 175–184 (Apr. 2016). Publisher: Radiological Society of North America.

[6] Zhang, B., Jia, C., Wu, R., Lv, B., Li, B., Li, F., Du, G., Sun, Z., and Li, X., "Improving rib fracture detection accuracy and reading efficiency with deep learning-based detection software: a clinical evaluation," *The British Journal of Radiology* **94**, 20200870 (Feb. 2021). Publisher: The British Institute of Radiology.

[7] Zhou, Q.-Q., Wang, J., Tang, W., Hu, Z.-C., Xia, Z.-Y., Li, X.-S., Zhang, R., Yin, X., Zhang, B., and Zhang, H., "Automatic Detection and Classification of Rib Fractures on Thoracic CT Using Convolutional Neural Network: Accuracy and Feasibility," *Korean Journal of Radiology* **21**, 869–879 (July 2020).

[8] Haitaamar, Z. N. and Abdulaziz, N., "Detection and Semantic Segmentation of Rib Fractures using a Convolutional Neural Network Approach," in [*2021 IEEE Region 10 Symposium (TENSYMP)*], 1–4 (Aug. 2021). ISSN: 2642-6102.

[9] Jin, L., Yang, J., Kuang, K., Ni, B., Gao, Y., Sun, Y., Gao, P., Ma, W., Tan, M., Kang, H., Chen, J., and Li, M., "Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet," *eBioMedicine* **62**, 103106 (Dec. 2020).

[10] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal Loss for Dense Object Detection," *arXiv:1708.02002 [cs]* (Feb. 2018). RetinaNet.

[11] Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., NanoCode012, TaoXie, Kwon, Y., Michael, K., Changyu, L., Fang, J., V, A., Laughing, tkianai, yxNONG, Skalski, P., Hogan, A., Nadar, J., imyhxy, Mammana, L., AlexWang1900, Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M. T., Marc, albinxavi, fatih, oleg, and wanghaoyang0106, "ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support," (Oct. 2021).

[12] Henon, Y., "pytorch-retinanet," (2018).

[13] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., "Pytorch: An imperative style, high-performance deep learning library," in [*Advances in Neural Information Processing Systems 32*], Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., eds., 8024–8035, Curran Associates, Inc. (2019).

[14] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You only look once: Unified, real-time object detection," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (June 2016).

[15] Redmon, J. and Farhadi, A., "Yolov3: An incremental improvement," (2018).

[16] Gadgeel, G., Burkow, J., Perez, F., Junewick, J., Zbojniewicz, A., Otjen, J., and Alessio, A., "Evaluation of inter-reader reproducibility for detection and labeling of pediatric rib fractures on radiographs," in [*International Pediatric Radiology Congress (abstract)*], (October 2021).