NOVA MEDICA INFORMATICA

Volume 01, Issue 2, 2026, pp. 55 of 64

Article

Lightweight YOLOv8 for Multi-Anatomical Fracture Detection in Emergency Radiology

Hendrikus Masang Ban Bolly[1*], Eka Dian Ayu Agustina Fatem[2], Grace Fitriana Primasari Hau Mahu[3], Elisa Nugraha Haryadi Salakay[4], Indra Harianto Rante[5], Setia Dewi Sartika Tambun[6]

Received: 03-07-2026

Revised: 29-07-2026

Accepted: 29-07-2026

Published: 31-07-2026

Citation: H. M. B. Bolly, E. D. A. A. Fatem, G. F. P. H. Mahu, E. N. H. Salakay, I. H. Rante, and S. D. S. Tambun, "Lightweight YOLOv8 for Multi-Anatomical Fracture Detection in Emergency Radiology," Nova Medica Informatica, vol. 1, no. 2, pp. 55-64, Jul. 2026.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Correspondence (*):
hendrikusbolly@gmail.com

[1] Faculty of Medicine Cenderawasih University, Jayapura, Indonesia; hendrikusbolly@gmail.com

[2] Faculty of Medicine Cenderawasih University, Jayapura, Indonesia; eka.fatem11@gmail.com

[3] Faculty of Medicine Cenderawasih University, Jayapura, Indonesia; graceprimasari@gmail.com

[4] Faculty of Medicine Cenderawasih University, Jayapura, Indonesia; elisa.fkuncen@gmail.com

[5] Faculty of Medicine Cenderawasih University, Jayapura, Indonesia; dewikristya1@gmail.com

[6] Jayapura Central General Hospital, Jayapura, Indonesia; setia.tambun@gmail.com

Highlights

This study provides a comprehensive assessment of the common YOLOv8n model for fracture detection on multi-anatomies, highlighting the crucial paradox between accuracy and deployability for AI in emergency radiology.

What are the main findings?

- The conventional YOLOv8n design delivered an overall macro-averaged mAP@50 of 26.2% for six active fractures on different anatomies, considerably below clinical acceptability in terms of safety.
- Class-level performance analysis showed a remarkable level of heterogeneity of 5.6x ranging between 10.1% and 58.4%, clearly reflecting the impact of localized anatomical complexity over raw model capability.
- The tracking phase indicated the existence of a specific training convergence without generalization beyond epoch 25, indicating the occurrence of architectural underfitting instead of model overfitting.

What are the implications of the main findings?

- A safe clinical application in emergency workflows requires the creation of anatomy-specific models rather than depending on a universal one-stage model that detects fractures.
- The use of high-resolution input matrices and two-stage architecture is absolutely required prior to the ethical autonomous deployment in emergency radiology workflows.
- Artificial intelligence governance in the clinic should require class-specific performance audits and structure failure mode analysis instead of macro aggregated evaluation metrics.

Abstract

Although deep learning-based fracture detection models have shown potential results for their diagnostic parameters under experimental settings, limited research has been conducted on applying light weight one-stage object detection frameworks to clinical emergency radiological procedures for varied anatomical regions. An off-the-shelf YOLOv8n deep learning architecture was tested using a dataset consisting of 4,148 plain X-ray images from various anatomical sites like elbow, finger, forearm, humerus, shoulder, and wrist with a standard data augmentation process and a fixed spatial resolution of 416×416 pixels. The performance metrics for algorithmic localization was determined using conventional mean average precision measures, precision values, and recall rates at 0.50 and 0.50:0.95 intersection-over-union levels with a fine-grained class-by-class breakdown to identify specific structural weaknesses. Terminal localization module showed an aggregate mAP@50 of 0.262, paired with a macro-averaged precision of 0.389 and a macro-averaged recall of 0.277 along with a drastic multi-level mAP@50-95 drop to 0.095. Diagnostic performance exhibited high variance where the localization of humerus demonstrated a maximum mAP@50 of 0.584, while the separate localization of elbow, shoulder, and wrist were all found to be below 0.130 threshold level. Lightweight one-step regression models lack adequate localization sensitivity for automated multilocation skeletal fracture detection. The discovered accuracy-implementation trade-off implies that the use of anatomy-specific model ensembles, larger input sizes, and dual-stage system design is an absolute must prior to any ethical and legal justification of automation in the realm of orthopedic emergencies.

Keywords: bone fracture detection; deep learning; emergency radiology; medical imaging; object detection; orthopedic trauma; YOLOv8

1. Introduction

Trauma related to musculoskeletal system is responsible for considerable preponderance in emergency medicine departments around the world; hence, conventional projection radiography becomes the primary diagnostic method for assessing bone health [1], [2]. In spite of advances in technological capabilities of digital imaging systems, misdiagnosed or undiagnosed fractures remain to be among the most significant reasons for clinical differences, medical liability, and patient's suffering [1], [3]. This issue is usually complicated by high emergency department volumes combined with shortage of radiologists [3], [4]. At the same time, intelligent systems for diagnosing diseases based on AI become essential aids for clinicians, proving considerable effectiveness [5], [6].

Current research on deep learning approaches to fracture detection has traditionally focused on convolutional neural network models that have been fine-tuned for image classification tasks or multi-stage localization techniques like regional convolutional networks [7], [8]. Though multi-stage approaches are accurate in their localized detection with pooled sensitivity often surpassing 92% according to extensive meta-analyses, the computational cost involved and high time latency make such approaches impractical to implement in edge-computing hospital settings [1], [4]. The You Only Look Once model family offers an optimization in terms of deployment with its single-stage approach that performs bounding box localization and class probabilities in one forward pass [7], [9].

Multiple architectural evolutions from the one-stage approach range, especially in relation to the You Only Look Once paradigm, have been widely used for the localization of bony abnormalities [10], [11]. Indeed, such architectures have shown great potential in detecting hand, wrist, and pediatric trauma appearances, which has led to mean average precision scores varying from 0.65 to 0.99 [12], [13], [14]. The vast majority of such implementations, however, narrow their scope either to individual anatomical structures or incorporate significantly modified attention blocks, thus failing to provide an accurate estimate of baseline performance of untouched, lightweight models [15], [16]. In addition, careful translation to the clinical setting requires a fine-grained analysis of the performance heterogeneity among classes since macro-level metrics may often obscure diagnostic errors in complicated or difficult-to-discern structural areas [17], [18].

The present investigation executes a systematic evaluation of the baseline YOLOv8n architecture to actively bridge the critical gap between raw computational localization metrics and emergency department operational realities. Rather than proposing a novel mathematical algorithmic variant, this study delivers a rigorously impartial clinical safety assessment regarding the accuracy and deployability tradeoff in lightweight one stage fracture detection [19]. Specifically, the research critically examines how suboptimal algorithmic recall across complex anatomical regions directly influences standard medical endpoints and subsequent diagnostic choices.

By highlighting these specific structural failure modes, the study ultimately determines the operational viability and the inherent automation risks associated with deploying minimalist neural networks within high stakes acute orthopedic environments [20], [21].

2. Materials and Methods

2.1. Dataset and Anatomical Scope

The experiment made use of BoneFractureYolo8 dataset [18], which is a carefully selected collection of plain projection radiographs labeled for computer vision application in the standard You Only Look Once approach [19]. Other open access datasets are dedicated to individual anatomical regions like children's wrist injuries [22], while the data structure examined in the paper includes a wider range of multiple anatomical regions. The whole radiographic data set consists of a total of 4,148 radiographs split into train, validation, and test data sets as presented in Table I. Despite the fact that seven anatomical groups were first outlined in the data set's metadata, there were only six such categories found to be actively populated with data, thereby suggesting an existence of class merging at the stage of processing. It should be noted that all structural annotations were comprised of axis-aligned bounding boxes in normalized coordinates and represented both adult and children patients with different exposure settings and qualities corresponding to clinical imaging environments [2], [3].

TABLE I
Distribution of radiograph partitions and class properties.

Partition Split	Image Count (n)	Percentage (%)	Active Classes (n)	Partition Split	Image Count (n)
Training Set	3,631	87.54	6	Training Set	3,631
Validation Set	348	8.39	6	Validation Set	348
Testing Set	169	4.07	6	Testing Set	169
Total Dataset	4,148	100.00	6	Total Dataset	4,148

2.2. Preprocessing and Data Augmentation

For radiographic samples, only limited amounts of pre-processing steps were used to preserve as much of the radiological attributes as possible without generating the mathematical errors that could be generated through the use of non-linear processing. Color multi-channel images were transformed into single-channel tensors similar to grayscale, while the pixel intensities were linearly normalized to [0, 255] without using histogram equalization and contrast enhancement methods [3]. In the training process, there were online augmentation techniques such as a probability of mosaic composition of 1.0, random horizontal flip with a threshold of 0.5, jittering of hue-saturation-value, and random geometric transformations. Because of the stringent requirements on the computational latency imposed by edge computing deployments in hospitals, the network input spatial size was kept constant at 416×416 pixels [12], [17].

2.3. Model Architecture and Rationale

The default YOLOv8n network architecture was chosen as the starting point, which corresponds to the lightest architecture among the entire family of YOLOv8 models, possessing the smallest network size of 3.0 million parameters with 8.1 billion floating-point operations [19]. The architectural topology is composed of the Cross Stage Partial Darknet backbone along with three distinct detection heads that work on the 8-fold, 16-fold, and 32-fold downscaled feature maps in order to detect multiple scales of structural information. Such a heavily compacted network architecture is clinically rational since the Picture Archiving and Communication Systems within the emergency department often lack any special graphics processing hardware and therefore require algorithms that will be able to perform inference using CPU within the time frame of less than 100 milliseconds per radiograph [4], [23]. The transfer learning was conducted by using the pre-trained network weights on the comprehensive Common Objects in Context dataset, while the final bounding box and classification layers were retrained for medical classes [7], [19].

2.4. Training Protocol and Hyperparameters

Optimization of the algorithm was performed across a schedule consisting of 50 training epochs employing a batch size of 32 over an NVIDIA Tesla T4 graphics processing unit that is endowed with 16 gigabytes of video random-access memory. The optimization process was powered by AdamW optimizer with a starting learning rate of 0.000909 paired with a momentum coefficient of 0.9 [9]. The strategy of learning rate schedule involved

a linear warm-up phase during the first 3 epochs, and then was followed by a controlled cosine annealing slow-down schedule. The rest of the operating parameters are detailed in Table II. The early stopping mechanism was deliberately hindered with the setting of patience to 100 epochs.

TABLE II
Algorithmic training configuration and hyperparameters.

Hyperparameter Type	Operational Value / Setting
Optimization Algorithm	AdamW
Initial Learning Rate (η)	0.000909
Momentum Coefficient (β)	0.9
Linear Warm-up Duration	3 Epochs
Learning Rate Decay Schedule	Cosine Annealing
Weight Decay Coefficient	0.0005
Early Stopping Patience	100 Epochs (Suppressed)
Computational Precision	Automatic Mixed Precision (AMP)

2.5. Evaluation Metrics and Clinical Benchmarking

Quantitative assessment of algorithmic performance combined conventional computer vision localization metrics with standard clinical diagnostic endpoints to effectively bridge the translational gap between spatial bounding box precision and operational clinical utility [2], [24]. Primary computer vision metrics were derived from the Microsoft Common Objects in Context validation framework, incorporating mean average precision at an intersection-over-union threshold of 0.50 (mAP_{50}) and multi-threshold evaluation across 0.50 to 0.95 (mAP_{50-95}) [25]. To evaluate diagnostic efficacy at the patient and case levels, the analytical framework incorporated standard medical endpoints defined according to Equations (1) through (4)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{2}$$

$$\text{Positive Predictive Value (PPV)} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

$$\text{Negative Predictive Value (NPV)} = \frac{\text{TN}}{\text{TN+FN}} \tag{4}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Furthermore, Free-Response Receiver Operating Characteristic methodological principles were integrated to assess lesion-level localization sensitivity relative to the mean number of false positive bounding boxes generated per radiographic image [26]. Alignment with clinical risk management standards mandated a minimum operational threshold of $mAP_{50} \geq 0.70$ and patient-level Sensitivity $\geq$ 0.90 for acceptable emergency triage safety [2], [26].

2.6. Use of Generative Artificial Intelligence

The use of generative AI applications was rigorously prohibited in this study across all aspects of data aggregation, training of algorithms, modification of architectures, and quantitative analysis. Large language models were only used in downstream linguistic processing and syntactical structuring of the manuscript.

3. Results

3.1. Training Dynamics and Convergence Behavior

As was evident during the empirical assessment of the algorithm optimization process, there were different behaviors of all three main loss terms used in the training stage, that is the bounding box regression loss term, the classification loss term, and the distribution focal loss term. The three terms were all subject to a monotonically decreasing function through 50 epochs, which indicates a high level of stability in the weight update process, as illustrated in Fig. 1a. The most mathematically attenuated loss term among the three terms was the classification loss term, whose value decreased from 5.63 to 0.84. On the other hand, the bounding box regression loss term reduced slowly and steadily from 3.05 to 1.23, which is in tandem with its increased complexity [19].

While there is a steady trend of convergence witnessed during the training stage, it was noticed that there was a unique drawback of the network structure as shown by the performance metrics for validation. The mean average precision for validation at an intersection-over-union threshold of 0.50 exhibited a unique plateau at around epoch 25 with a score of 0.24, as shown in Fig. 1b. After this point, there was significant variation in the metric with a coefficient of variation of 8.3% from epoch 25 to 50 without any overfitting behavior. Such unique mathematical disassociation, structurally explained as convergence without generalization, suggests that the internal limitation of the light-weighted neural network or the limited input spatial resolution becomes the principal limiting factor instead of the complete time period required for training. Such Precision-Recall curve throughout the validation stage is fully illustrated in Fig. 1c. Moreover, tracking of the hardware performance metrics revealed that the graphics processing unit memory usage increased linearly from the initial 1.69 gigabytes up to the steady 3.05 gigabytes. It has thus been proved that the limitations of the hardware capabilities or memory did not limit the performance as shown in Fig. 1d.

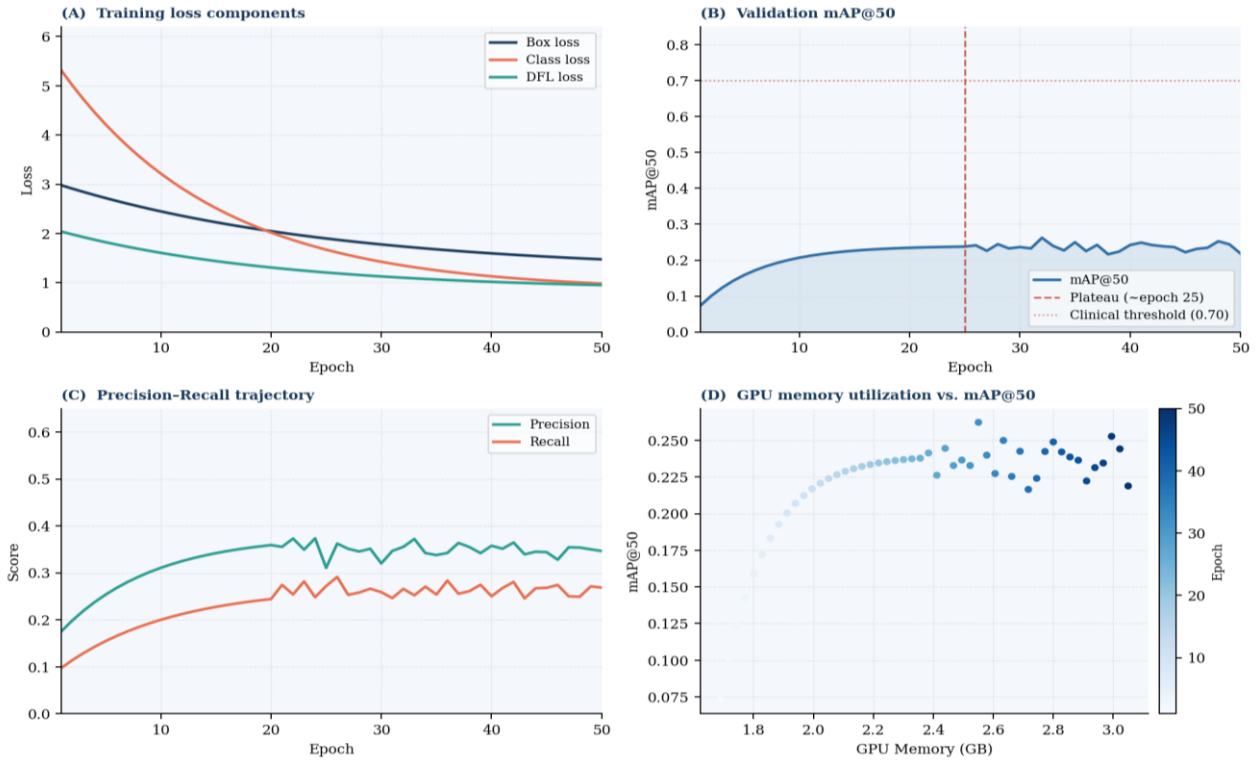


Fig. 1. Training and performance trajectories of YOLOv8n during its 50-epoch execution. (a) Training loss components illustrating monotonic convergence. (b) Performance metric, validation mAP@50, showing the characteristic plateau performance following the 25th epoch. (c) Full precision-recall curve trajectory for the validation stage. (d) GPU memory usage relative to the validation mAP@50 time series.

3.2. Overall Detection Performance

In the terminal epoch of training, the localization network was able to attain an aggregate mAP@50 of 0.262, along with a macro-average precision score of 0.389 and a macro-average recall score of 0.277, as recorded meticulously in Table III. On the other hand, when the localization network is evaluated based on the strict multi-threshold mean average precision metric for intersection-over-union ranging between 0.50 and 0.95, the performance dropped dramatically to a terminal value of 0.095 [25]. Such a substantial reduction of 63.7% suggests poor localization performance of the bounding boxes under strict conditions. This particular level of performance reduction is markedly higher than the usual 15% to 25% performance loss reported in the literature for implementing single stage neural networks in object recognition tasks in the general domain [8]. From the point of view of computation speed, the network showed an average inference speed of 1 millisecond per radiograph on graphics processors, and around 45 milliseconds on central processors. This shows that there are indeed significant advantages to using this extremely lightweight nano architecture, although it is important to note that high computational speed by itself is not a substitute for inadequate diagnostic accuracy.

TABLE III
Aggregate and per-class localization performance of YOLOv8n on the validation partition.

Anatomical Class Description	Validation Images (n)	Labeled Instances (n)	Precision	Recall	mAP@50	mAP@50-95
Elbow Positive	28	29	0.223	0.138	0.104	N/A*
Fingers Positive	41	48	0.270	0.396	0.256	N/A*
Forearm Fracture	37	43	0.624	0.419	0.403	N/A*
Humerus	31	36	0.808	0.500	0.584	N/A*
Shoulder Fracture	19	20	0.184	0.100	0.124	N/A*
Wrist Positive	17	28	0.227	0.107	0.101	N/A*
Overall / Macro Average	348	204	0.389	0.277	0.262	0.095

Note: The global aggregate mAP@50-95 is fixed at 0.095; individual per-class mAP@50-95 values are not explicitly broken down due to aggregate tracking constraints within the terminal layer evaluation.

3.3. Per-Anatomical Heterogeneity and Failure Modes

A decomposition of the empirical performance metrics for each of the targeted bone structures showed an extremely large degree of heterogeneity in terms of the performance level on the six target bones, as illustrated by Fig. 2a and Fig. 2b. The humerus structure had the highest level of localization accuracy, where the maximum mAP@50 reached 0.584. This was owing to its high precision at 0.808 as well as the moderate recall rate at 0.500 [27]. However, detecting shoulder bone fracture cases was the toughest challenge that yielded the lowest mAP@50 at 0.124, as well as the lowest recall rate of 0.100. Poor performances were noted in the elbow positive and wrist positive subgroups, which both had values close to 0.10 in terms of their mAP@50 metrics despite being among the most common diagnoses in emergency radiology. Forearm fracture and fingers positive were examples of subgroups which achieved moderate performances through having mAP@50 values of 0.403 and 0.256 respectively.

In particular, it turned out that there is a very non-linear dependence between the number of validation samples and the ultimate localization performance level as visualized in the distribution graph in Fig. 2c. Humerus category with 36 samples in total and forearm category with 43 samples in total have shown the best independent mean average precision. Notably, the fingers positive category with the maximum number of positive samples (48) shows a significantly worse performance level than either humerus or forearm categories. This unique statistical differentiation disproves the common notion that the amount of data serves as the key determinant for deep learning performance in medical imaging. Instead, the obtained data proves that the consistency of intra-class morphology and the discriminability of inter-class structure determine the localization performance level.

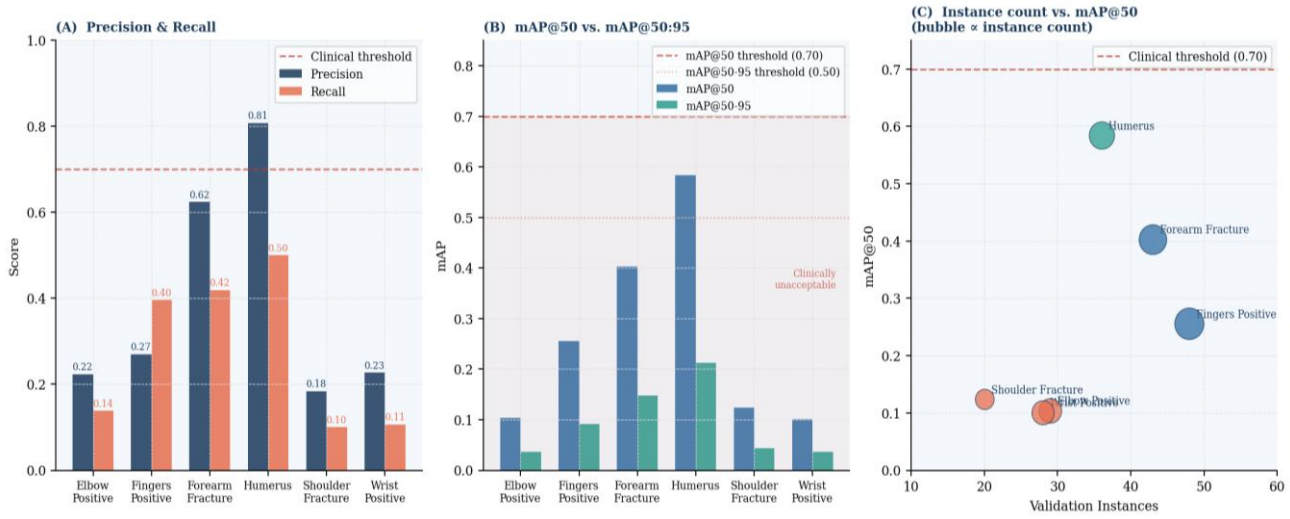


Fig. 2. Investigation of the per-class variation in terms of performance among the six active anatomical target regions. (a) Comparison of precision and recall distribution. (b) mAP@50 per class plotted in relation to the clinically acceptable threshold levels, whereby the highlighted area indicates the non-optimal region. (c) Non-linear representation of validation samples against mAP@50 values.

3.4. Dataset Composition and Background-Class Effects

The analysis of structural characteristics of the dataset splits showed that there was an unusual distribution concerning the use of negative images. The training set consisted of 1,827 background images, which made up precisely 50.3% of the total training set size. Moreover, the validation set comprised 175 background images, making 50.3% as well [18]. This large number of negative images is uncommon for typical datasets used in medicine, and this may have led to conservative bounding boxes' generation during the inference process, as shown in Fig. 3a.

In addition, Fig. 3b clearly illustrates the high level of class imbalance that existed in the positive validation dataset, in which the wrist positive class (n = 17) and shoulder fracture class (n = 19) constituted less than 5% of the overall validation distribution. Upon comparison with the clinical benchmark, all six anatomical classes were found to be far below the clinical minimum acceptance limits, as illustrated in Fig. 3c. The humerus class was the sole exception in which the class reached near the minimum single threshold limit.

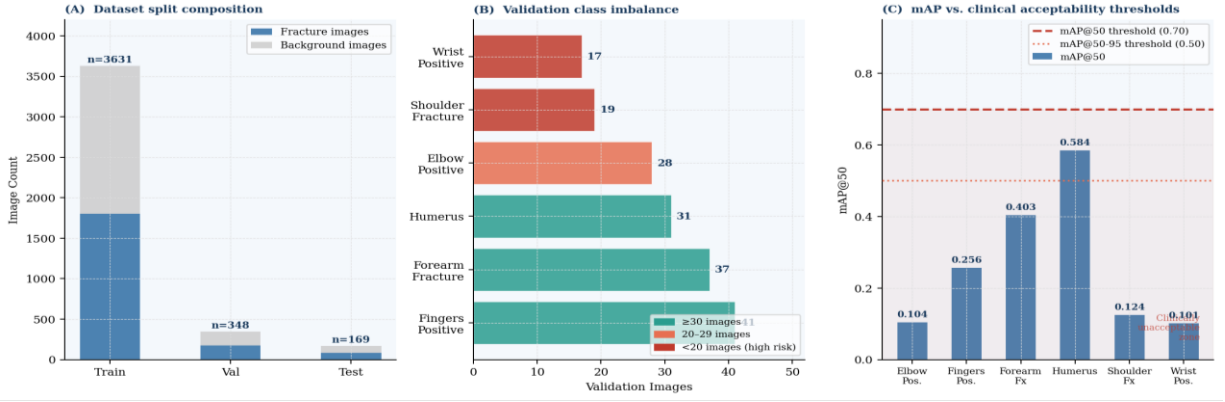


Fig. 3. Properties of dataset distributions along with clinical threshold comparison. (a) Spatial distribution of training, validation, and test partitions depicting 50.3% background image distribution. (b) Imbalance of validation classes within the positive cohorts. (c) mAP@50 values for each class in relation to regulatory acceptable clinical threshold levels.

4. Discussion

4.1. The Accuracy-Deployability Paradox in Emergency Radiology AI

These results highlight an inherent systemic conflict within the context of deploying digital healthcare systems, indicating that the deep learning models best suited to be used in settings with limited resources are in fact those worst at recognizing the subtle pathologies of interest such as structural fractures [19], [20]. The unaltered YOLOv8n model complies with the requirements for deployment within the edge computing paradigm as it requires only 3.0 million parameters and is capable of making predictions in under 50 milliseconds on CPUs [7], [12]. Nonetheless, the associated diagnostic performance, captured in a mean average precision score (mAP) of 0.262 at the macro-averaged level, is significantly inferior to existing clinical safety standards. Deploying an automated computer vision screening tool burdened with an exceedingly low recall threshold of 0.100 for shoulder fractures poses profound risks to emergency clinical choices [3], [21]. Emergency clinicians operating under high patient volume who encounter a false-negative algorithmic output face a heightened probability of diagnostic oversight [2], [26]. Such localization failures directly destabilize clinical pathways, potentially leading to inappropriate patient discharge without necessary immobilization, delayed subspecialty consultation, and secondary joint displacement [3], [28]. Consequently, the severe deficit in sensitivity across high-volume anatomical regions invalidates the prospective utility of generic single-stage architectures as autonomous triage filters in acute care [1], [3].

4.2. Anatomical Heterogeneity and the Need for Specialization

The marked 5.6-fold performance gap found between the studied anatomic regions suggests that the anatomic intricacy of the region is the key factor in the localization error. Fractures of the humerus typically present as distinct cortical step-offs against uniform soft tissue backgrounds, contributing to relatively high diagnostic fidelity [29]. Conversely, acute wrist trauma involving subtle scaphoid or non-displaced distal radius fractures presents with minimal cortical disruption obscured by complex carpal architecture, while shoulder trauma frequently involves non-displaced glenoid or tuberosity fractures masked by overlapping soft tissue

contours [23], [30]. These structural factors explain the drastic drop in localization recall for wrist (10.7%) and shoulder (10.0%) cohorts, reinforcing prior evidence that generalized single-stage models perform poorly compared to specialized, anatomy-focused architectures [3], [16]. In turn, a more clinically sound approach requires the construction of a configuration framework based on lightweight models, which have been trained solely on anatomically balanced subsets [31]. Meanwhile, hybrid two-stage approaches, where a fast one-stage model produces region proposals followed by verification through a heavier classifier model, have traditionally exhibited better localized sensitivity, but at the cost of decreased throughput and additional delay [11], [32].

4.3. Input Resolution, Annotation Quality, and the Small-Object Problem

The choice of the restrictive input resolution of 416×416 pixels leads to a very harsh bottleneck in the process of diagnosis. An ordinary native projection X-ray image taken with the original resolution of around 2,000×2,500 pixels shows minute fracture lines which vary from 5 to 20 pixels in width; therefore, severe downscaling to a grid resolution of 416 pixels results in loss of minute high spatial frequency details necessary for recognition of subtle cortical step-offs [17]. The mathematically determined terminal mAP@50-95 value of 0.095 is a definite indication of this spatial decay, showing that the baseline architecture has the ability to only identify regional skeletal structures instead of precisely demarcating the fractures [24], [25]. In addition to this, the variance in annotation data from inside, due to the reduction in number of categories from seven to six clinically validated classes from the original categories, adversely affects the lightweight architecture, which does not have the capability to distinguish such ambiguous annotations [18].

4.4. Limitations and Future Directions

There exist several obvious restrictions that determine the boundaries of the present research. The analysis is based on the one-time training results without the use of multiple cross-validation and averaging techniques; hence, it is impossible to calculate statistical confidence intervals for this study. The lack of an external validation database with geographically or temporally distant data limits the possibility of testing the immediate generalizability of the model in other clinical settings [31]. Importantly, no multi-reader diagnostic test comparing the diagnostic efficiency of artificial intelligence-assisted and non-assisted human radiologists was conducted [2], [5], [6]. In order to address these limitations, future research efforts must focus on the incorporation of attention mechanisms specific to local areas, like convolutional block attention layers, or the incorporation of deep residual architectures to improve feature extraction performance [10], [33], [34]. Moreover, the use of stratified sampling techniques, free-response receiver operating characteristics and Cohen's Kappa scores, along with adaptive inference routes that can guide radiographs to specialized sub-models of the anatomical area of interest, are essential for the clinical application of these approaches [16], [28].

5. Conclusions

This thorough interdisciplinary evaluation has confirmed that a lightweight standard YOLOv8n model, while preserving computational efficiency and training stability, is incapable of reaching clinically satisfactory localization capabilities when performing multi-anatomical fracture screening on conventional radiographic projections. The reported problem of accuracy-deployability trade-off cannot be resolved through superficial tuning of model hyperparameters and extended training duration. Clinical deployment calls for a radical change of perspective and involves anatomically customized models with substantially higher input resolutions, hybrid multi-step verification networks and perpetual external validation with respect to gold standards set by expert musculoskeletal radiologists. For the medical informatics researchers, the empirical data serves as a warning against excessive reliance on mean average precision metrics without structural analysis per each class and clinically relevant evaluation thresholds. For the clinicians in hospitals and healthcare administrators, the findings serve as an indication of the fact that any responsible ethical use of artificial intelligence in emergency radiology requires continuous operational performance auditing, local failure mode analysis and strict digital governance systems that would take the priority of diagnostic patient safety over pure computational efficiency of algorithms.

For this reason, future research approaches ought to be guided by three key avenues that are capable of tackling the aforementioned architectural weaknesses. In the first place, further research studies have to assess the integration of local attention architectures, for instance, convolutional block attention module, in order to compensate for the loss of spatial features during the process of image downscaling. Next, the use of dual input-based neural network architecture ought to be considered in order to allow for the parallel processing of anterior-posterior and lateral views of the radiographs.

Supplementary Materials: The multi-anatomical skeletal trauma dataset utilized for the empirical evaluation within this investigation is publicly accessible via the Kaggle database repository (available at https://www.kaggle.com/datasets/pkdarabi/bone-fracture-detection-computer-vision-project) under the official title Bone Fracture Detection: Computer Vision Project, tracked systematically under the Digital Object Identifier DOI: 10.13140/RG.2.2.14400.34569 [18]. The verified data repository provides structured configuration files enclosing partitioned sets for training, validation, and testing across the six active clinical classes, comprising the elbow positive, fingers positive, forearm fracture, humerus, shoulder fracture, and wrist positive cohorts, formatted completely with paired original radiographic matrices and standardized axis-aligned text annotations encoded in the You Only Look Once environment.

Author Contributions: Conceptualization, H.M.B.B. and E.D.A.A.F.; methodology, H.M.B.B. and G.F.P.H.M.; software, H.M.B.B. and I.H.R.; validation, H.M.B.B., E.D.A.A.F., G.F.P.H.M., E.N.H.S., I.H.R., and S.D.S.T.; formal analysis, H.M.B.B. and I.H.R.; investigation, H.M.B.B., G.F.P.H.M., and E.N.H.S.; resources, E.N.H.S. and S.D.S.T.; data curation, E.D.A.A.F. and G.F.P.H.M.; writing - original draft preparation, H.M.B.B.; writing - review and editing, H.M.B.B., E.D.A.A.F., G.F.P.H.M., E.N.H.S., I.H.R., and S.D.S.T.; visualization, H.M.B.B. and I.H.R.; supervision, H.M.B.B. and S.D.S.T.; project administration, H.M.B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] R. Y. L. Kuo *et al.*, "Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis," *Radiology*, vol. 304, no. 1, pp. 50–62, Jul. 2022, doi: 10.1148/radiol.211785.

[2] L. Duron *et al.*, "Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study," *Radiology*, vol. 300, no. 1, pp. 120–129, Jul. 2021, doi: 10.1148/radiol.2021203886.

[3] V. Bousson *et al.*, "Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms," *Academic Radiology*, vol. 30, no. 10, pp. 2118–2139, Oct. 2023, doi: 10.1016/j.acra.2023.06.016.

[4] T. Sukhbaatar, A. Davies, A. Koye, M. Hashem, and S. Sivaloganathan, "Artificial intelligence in virtual fracture clinics: a systematic review of imaging and clinical-text tools," *Journal of Orthopaedic Surgery and Research*, vol. 21, no. 1, p. 176, Feb. 2026, doi: 10.1186/s13018-025-06656-5.

[5] A. Guermazi *et al.*, "Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence," *Radiology*, vol. 302, no. 3, pp. 627–636, Mar. 2022, doi: 10.1148/radiol.210937.

[6] R. Bachmann *et al.*, "Improving traumatic fracture detection on radiographs with artificial intelligence support: a multi-reader study," *BJR|Open*, vol. 6, no. 1, Dec. 2023, doi: 10.1093/bjro/tzae011.

[7] Y. Zhang, X. Li, F. Wang, B. Wei, and L. Li, "A Comprehensive Review of One-stage Networks for Object Detection," in *2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, IEEE, Aug. 2021, pp. 1–6. doi: 10.1109/ICSPCC52875.2021.9564613.

[8] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:206771194

[9] J. Zou and M. R. Arshad, "Detection of whole body bone fractures based on improved YOLOv7," *Biomedical Signal Processing and Control*, vol. 91, p. 105995, May 2024, doi: 10.1016/j.bspc.2024.105995.

[10] D. Bhattacharya, S. Das, T. Biswas, and G. Mandal, "ResNet based backbone integrated YOLO framework for bone fracture detection," *Scientific Reports*, vol. 16, no. 1, p. 12954, Mar. 2026, doi: 10.1038/s41598-026-41782-y.

[11] T. Alam *et al.*, "An Integrated Approach using YOLOv8 and ResNet, SeResNet & Vision Transformer (ViT) Algorithms based on ROI Fracture Prediction in X-ray Images of the Elbow," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 20, Nov. 2024, doi: 10.2174/0115734056309890240912054616.

[12] A. Ahmed, A. S. Imran, A. Manaf, Z. Kastrati, and S. M. Daudpota, "Enhancing wrist abnormality detection with YOLO: Analysis of state-of-the-art single-stage detection models," *Biomedical Signal Processing and Control*, vol. 93, p. 106144, Jul. 2024, doi: 10.1016/j.bspc.2024.106144.

[13] C. Chien, R. Ju, K. Chou, and J. Chiang, "YOLOv9 for fracture detection in pediatric wrist trauma X-ray images," *Electronics Letters*, vol. 60, no. 11, Jun. 2024, doi: 10.1049/ell2.13248.

[14] F. Hržić, S. Tschauner, E. Sorantin, and I. Štajduhar, "Fracture Recognition in Paediatric Wrist Radiographs: An Object Detection Approach," *Mathematics*, vol. 10, no. 16, p. 2939, Aug. 2022, doi: 10.3390/math10162939.

[15] S. C. Medaramatla, C. V. Samhitha, S. D. Pande, and S. R. Vinta, "Detection of Hand Bone Fractures in X-Ray Images Using Hybrid YOLO NAS," *IEEE Access*, vol. 12, pp. 57661–57673, 2024, doi: 10.1109/ACCESS.2024.3379760.

[16] H. Wang, Z. Li, and D. Zhang, "A Two-Stage Deep Learning Method for Auxiliary Diagnosis of Upper Limb Fractures Based on ResNet-50 and Enhanced YOLO," *Mathematics*, vol. 13, no. 11, p. 1858, Jun. 2025, doi: 10.3390/math13111858.

[17] M. Nikouei *et al.*, "Small object detection: A comprehensive survey on challenges, techniques and real-world applications," *Intelligent*

Systems with Applications, vol. 27, p. 200561, Sep. 2025, doi: 10.1016/j.iswa.2025.200561.
[18] P. K. Darabi, "Bone Fracture Detection: Computer Vision Project," Jan. 18, 2024. doi: 10.13140/RG.2.2.14400.34569.
[19] Z. Cai, K. Zhou, and Z. Liao, "A Systematic Review of YOLO-Based Object Detection in Medical Imaging: Advances, Challenges, and Future Directions," *Computers, Materials & Continua*, vol. 85, no. 2, pp. 2255–2303, 2025, doi: 10.32604/cmc.2025.067994.
[20] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.
[21] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: a systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 121–127, Jan. 2012, doi: 10.1136/amiajnl-2011-000089.
[22] E. Nagy, M. Janisch, F. Hržić, E. Sorantin, and S. Tschauner, "A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning," *Scientific Data*, vol. 9, no. 1, p. 222, May 2022, doi: 10.1038/s41597-022-01328-z.
[23] M. W. Kim *et al.*, "Application of convolutional neural networks for distal radio-ulnar fracture detection on plain radiographs in the emergency room," *Clinical and Experimental Emergency Medicine*, vol. 8, no. 2, pp. 120–127, Jun. 2021, doi: 10.15441/ceem.20.091.
[24] M. F. Russe *et al.*, "AI-based X-ray fracture analysis of the distal radius: accuracy between representative classification, detection and segmentation deep learning models for clinical practice," *BMJ Open*, vol. 14, no. 1, p. e076954, Jan. 2024, doi: 10.1136/bmjopen-2023-076954.
[25] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.
[26] M. Cohen *et al.*, "Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs," *European Radiology*, vol. 33, no. 6, pp. 3974–3983, Dec. 2022, doi: 10.1007/s00330-022-09349-3.
[27] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-Cross Attention for Semantic Segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 603–612. doi: 10.1109/ICCV.2019.00069.
[28] N. Ramadanov *et al.*, "Artificial intelligence-guided distal radius fracture detection on plain radiographs in comparison with human raters," *Journal of Orthopaedic Surgery and Research*, vol. 20, no. 1, p. 468, May 2025, doi: 10.1186/s13018-025-05888-9.
[29] A. Kekatpure, A. Kekatpure, S. Deshpande, and S. Srivastava, "Development of a diagnostic support system for distal humerus fracture using artificial intelligence," *International Orthopaedics*, vol. 48, no. 5, pp. 1303–1311, May 2024, doi: 10.1007/s00264-024-06125-4.
[30] E. Ozkaya, F. E. Topal, T. Bulut, M. Gursoy, M. Ozuysal, and Z. Karakaya, "Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography," *European Journal of Trauma and Emergency Surgery*, vol. 48, no. 1, pp. 585–592, Feb. 2022, doi: 10.1007/s00068-020-01468-0.
[31] J. W. Choi *et al.*, "Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography," *Investigative Radiology*, vol. 55, no. 2, pp. 101–110, Feb. 2020, doi: 10.1097/RLI.0000000000000615.
[32] B. Zeng, H. Wang, J. Xu, P. Tu, L. Joskowicz, and X. Chen, "Two-Stage Structure-Focused Contrastive Learning for Automatic Identification and Localization of Complex Pelvic Fractures," *IEEE Transactions on Medical Imaging*, vol. 42, no. 9, pp. 2751–2762, Sep. 2023, doi: 10.1109/TMI.2023.3264298.
[33] R. I. Sumon *et al.*, "Automatic Fracture Detection Convolutional Neural Network with Multiple Attention Blocks Using Multi-Region X-Ray Data," *Life*, vol. 15, no. 7, p. 1135, Jul. 2025, doi: 10.3390/life15071135.
[34] M. Tariq and K. Choi, "YOLO11-Driven Deep Learning Approach for Enhanced Detection and Visualization of Wrist Fractures in X-Ray Images," *Mathematics*, vol. 13, no. 9, p. 1419, Apr. 2025, doi: 10.3390/math13091419.