Received: 18 August 2025 | Revised: 2 February 2026 | Accepted: 25 March 2026 | Published online: 10 April 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 4(4) 549–560
DOI: 10.47852/bonviewAIA62027255

BON VIEW PUBLISHING

Performance Analysis of YOLOv11-m and Related Architectures in Pediatric X-Ray Fracture Detection

Muhanad Abdul Elah Alkhalisy[1], Qusay Shihab Hamad[2, 3,*], Ali Retha Hasoon Khayeat[4] and Shahrel Azmin Suandi[5,*]

[1] *Business Informatics College, University of Information Technology and Communications, Iraq*

[2] *Department of Quality Assurance and University Performance, University of Information Technology and Communications, Iraq*

[3] *College of Engineering, Al-Farabi University, Iraq*

[4] *College of Computer Science and Information Technology, University of Kerbala, Iraq*

[5] *School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Malaysia*

Abstract: Wrist fractures are one of the most frequent fractures in children that should be diagnosed properly and immediately to avoid any further complications. Regular radiographic analysis is time-consuming, relies on qualified radiologists, and is subject to human error. There has been a lot of buzz about the potential of deep learning technologies to automate medical image analysis in recent times. The algorithms in the YOLO series are sophisticated and refined, representing some of the best work in the field. The GRAZPEDWRI-DX pediatric wrist dataset is used in this study to address the variation in the YOLOv11 detection model's inference time, recall, precision, and mean average precision (mAP). A medium-scaled variant of YOLOv5, YOLOv7, YOLOv8, YOLOv9, and YOLOv10 models were compared with YOLOv11-m. According to the experimental results, YOLOv11-m achieves the highest mAP@50–95 (0.569) for fracture detection, whereas YOLOv11-x attains the highest mAP@50–95 (0.424) across all category classes. An ablation study of YOLOv11 architectural elements was done to determine how the model optimized for medical imaging through the attention mechanism. On the other hand, the performance in fracture detection and inference time for YOLOv11 is superior to that of previous YOLO models, while it has lower computational complexity. Additionally, a performance comparison among recent YOLO-based models and the RT-DETR and Faster R-CNN models reveals variations in detection precision, recall, and accuracy, providing insight into the strengths and trade-offs of each approach. YOLOv11 could therefore make recurrent contributions and be of great value to clinical decision-making.

Keywords: YOLOv11, bone fracture, deep learning, pediatric X-ray, medical image recognition

1. Introduction

Fractures of the wrist in children are common injuries during day-to-day activities and in sports. These injuries require proper and early diagnosis to avert potential sequelae, including deformity, chronic pain, and chronic functional impairment [1]. The conventional methods for identifying wrist fractures rely heavily on radiologists' interpretation of X-ray images, which is time-consuming and prone to error, particularly in regions where qualified radiologists are scarce [2–4]. The recent breakthroughs in the field of deep learning have spawned computer-aided diagnosis (CAD) systems. Such systems have been turned into vital assets of professionals, which help in the analysis of medical images and enhance diagnostic performance and accuracy [5].

Wrist fractures in children consume a large amount of healthcare facilities such as emergency services, medical imaging, and follow-ups, as well as surgery in some instances [6]. Given the limitations of manual diagnosis, automated bone fracture detection is increasingly emphasized. Recent trends in artificial intelligence (AI) indicate that intelligent systems can be highly accurate at detecting fractures [7–9]. The YOLO algorithm is one of the most popular AI models and has become a subject of intensive

***Corresponding authors:** Qusay Shihab Hamad, Department of Quality Assurance and University Performance, University of Information Technology and Communications, Iraq and College of Engineering, Al-Farabi University, Iraq. Email: qusay@uoitc.edu.iq and Shahrel Azmin Suandi, School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Malaysia. Email: shahrel@usm.my

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

discussion in medical imaging due to its impressive balance between speed and detection accuracy [10, 11]. Recent YOLO models, particularly YOLOv8, excel at identifying fractures in X-ray images of pediatric wrists [12].

YOLO family has also been improved with time, and the recent models, YOLOv8 and YOLOv10, have made significant advancements in increasing the accuracy and speed [13]. Such developments make YOLO models more useful in bone fracture diagnosis [14]. This paper analyzes the YOLOv11 model and its variants using the GRAZPEDWRI-DX dataset [15].

The primary contributions of this work can be encapsulated as follows:

1) New use of YOLOv11: This is the first study we are aware of to evaluate the YOLOv11 architecture and its variants (n, s, m, l, x) for fracture identification in pediatrics using the full GRAZPEDWRI-DX dataset.
2) Detailed benchmarking: YOLOv11-m performance is compared with its predecessors (YOLOv5, YOLOv8, YOLOv9, YOLOv10) and with new suggested better models. The comparison presents YOLOv11 alongside the latest object detection frameworks.
3) Clinical relevance: YOLOv11-m presents the best trade-off between inference time (1.1 ms) and accuracy (mAP@50–95 = 0.569), so it is the most suitable tool to use in clinical decision support in real time, especially in emergency care scenarios.
4) Methodological rationale: This paper includes an ablation study of YOLOv11 architectural elements and how they can be optimized for medical imaging through attention and simplified design decisions.
5) The GRAZPEDWRI-DX dataset had undergone resizing of all images to a standard size of 640 × 640 pixels. This standardization brought about consistency between models as well as drastically lowering training time without compromising on the detection accuracy.

Section 2 is the beginning of the literature review in this paper. Section 3 presents the methodology, including the architecture of the YOLOv11 model, a description of the dataset, preprocessing steps, and the experimental environment setup. Section 4 presents and describes the findings and discussion of the model performance in comparison to its predecessors and finally presents the discussion of the accuracy of the Yanomamo wrist fractures detection of children using Yanomamo with the YOLOv11. Section 5 concludes our work, summarizing the key findings and outlining possible research directions to enhance future automated fracture detection research.

2. Literature Review

Computer vision has made immense contributions to the study of wrist injuries, particularly in the detection of fractures. This section provides an overview of the existing literature on fracture detection in the adult and pediatric populations, highlighting the most significant findings. There were short supplies of adult and pediatric fracture detection datasets [15]. All studies in this section were based on publicly available GRAZPEDWRI-DX data, which consists of pediatric wrist X-ray images [16]. For instance, Ahmed et al. [17] use deep neural network-based single-stage detection techniques (YOLO versions 5, 6, 7, and 8) to identify wrist fractures. Experiments have demonstrated that these models are more effective than the popular Faster Region-based Convolutional Neural Networks (R-CNN) constitute a category of machine learning models designed for computer vision, particularly in the domains of object detection and localization. method for fracture detection. The evaluations of the composite variations of all the YOLO models revealed that the most sensitive model to fracture detection was that of YOLOv8m, with a score of 0.92 and a mean average precision (mAP) of 0.95. The best recall of 0.83 by YOLOv6m was recorded in all the classes. All the classes had the highest recall of 0.83 by YOLOv6m. YOLOv8x achieved the highest mAP of 0.77 across all datasets, indicating that the model was effective at detecting pediatric wrists. Till et al. [18] employed the YOLOv7 object detection model algorithm on the pediatric wrist dataset. Data preparation, image sizing for training and testing, and model configuration were optimized. The model's attention areas were explained using explainable AI techniques, namely, Gradient Class Activation Mapping (Grad-CAM). The gains attained were up to +13.36% (mAP@0.5) and +27.01% (mAP@0.95) by model-produced gains and up to +25.51% (mAP@0.5) and +39.78% (mAP@0.95) by data preprocessing. These optimizations had a considerable impact on the mAP. Ju et al. [19] noted a new application of an improved version of the YOLOv8 model and came up with the YOLOv8+GC model to detect fractures. According to the results of the experiments, the accuracy reaches 66.32% at the union threshold of 0.5, which was greater than the 63.58% in the original model. Chien et al. [20] use the YOLOv9 algorithm to detect fractures in computer-aided diagnosis (CAD), helping physicians interpret X-ray images. The researchers augmented the data to increase the detection accuracy. The results have shown that the average precision (mAP@50–95) of 42.16% increased to 43.73% (3.7%), with the use of the YOLOv9 model versus the old model.

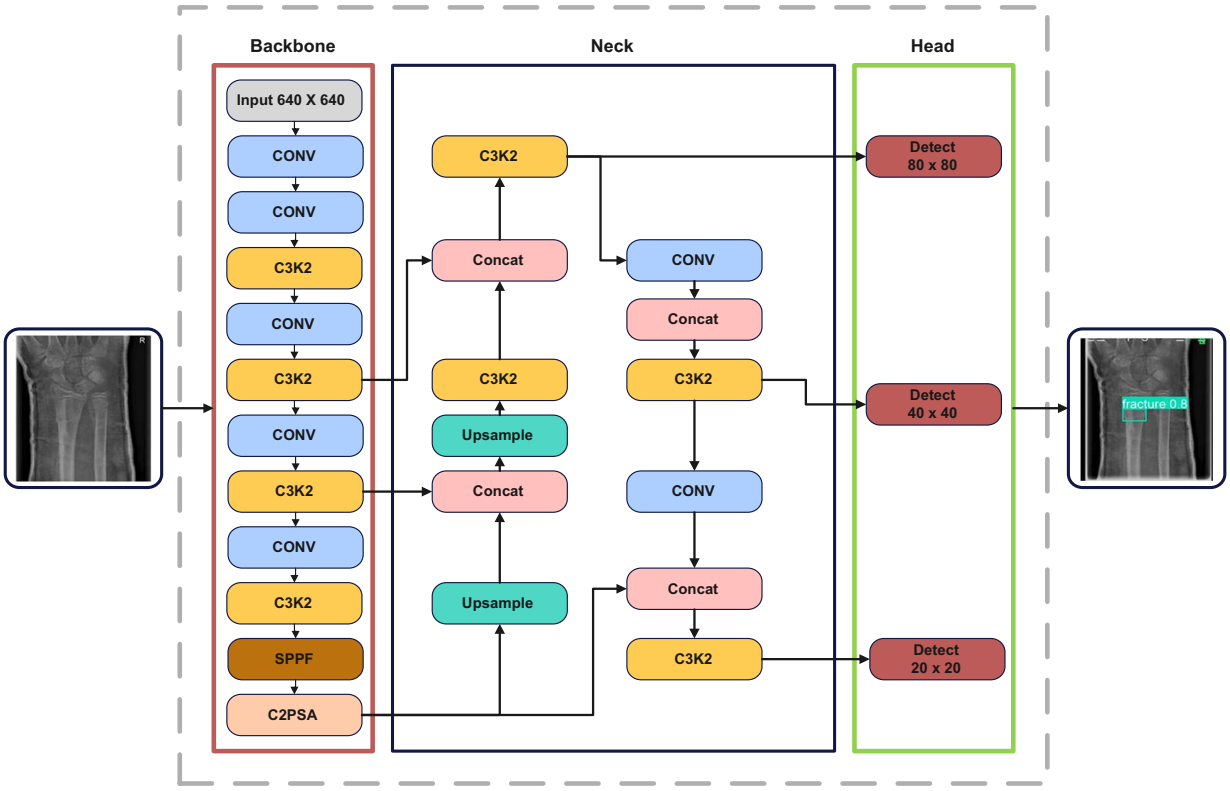
Ju et al. [19] propose four iterations of feature context excitation within the YOLOv8 model. Both versions have distinct features: contextual excitation, global context, gather excitation, and squeeze excitation. Results indicate that the improved model increases the mAP@50 by 65.78% to 66.32%. This model is better than other versions and superior to all state-of-the-art models that have existed to date. Ahmed et al. [21] evaluate various versions of the YOLOv10 model to assess its effectiveness in identifying pediatric wrist fractures. This paper investigates how changes in model complexity affect detection performance by enhancing the model architecture and adopting a dual-label mapping strategy. The model trained on the data achieved a mAP@50–95 of 51.9, better than any YOLOv9 model trained on the same dataset so far (43.3), an improvement of 8.6.

3. Methodology

Precision is important in a wide range of activities and usages; even slight mistakes can have a significant impact. A particular task is medical image analysis during diagnostic work, since proper, reliable outcomes are extremely important and required. The patient care, course of treatment, and health outcomes are directly affected by the decisions [22].

The final model is YOLOv10 that is aimed at detecting simplification and acceleration due to its lightweight design [22]. Meanwhile, YOLOv11, in turn, is an extension of YOLOv8, with more powerful modules for feature extraction and attention mechanisms to enhance detection performance [23]. YOLOv11 is better structured than YOLOv8, extracting features more effectively and minimizing parameter count [24]. YOLOv11 also incorporates various refinements, including the addition of the C3K2 [25] block composition to the C2f block of the neck [26], which is more computationally efficient and provides higher processing speed. Modification is performed using two convolutional functions with small kernel sizes, which are faster to compute and take less

Figure 1
YOLOv11 architecture



time to infer at the cost of insignificant accuracy, and a rational decision-making process capable of assisting in the accurate detection of notable and obscure abnormalities in wrist X-rays or wrist MRIs [27]. The YOLOv11 workflow is shown in Figure 1.

YOLO models consist of three components: backbone, neck, and head. YOLOv11 follows the same structure.

Backbone: The backbone is significantly improved by adding a Cross-Stage Partial (CSP) block with kernel size C3k2 (2f) instead of the C2f used in previous designs [28]. The C3k2 block is a lighter CSP bottleneck. Instead of making one big convolution, two smaller convolutions are used [29]. C3k2 implies that there is a smaller size of the kernel ("k2), which enables the model to be faster without any loss in performance. Use the SPPF block (Spatial Pyramid Pooling-Fast), for example, in YOLOv8, to pool spatially at different scales [25]. Multi-scale features are obtained by applying different max pooling layers to the image inputs. Meanwhile, to improve the identification of crucial areas in the image, the Cross-Stage Professional that houses the spatial attention module block adds an attention module immediately after the Spatial Pyramid Pooling-Fast block [26]. The application can be effectively used for complex tasks, such as healthcare, where accuracy and precision are important.

Neck: The main task of the neck in YOLO is to merge feature maps at various scales and pass them to the head blocks via several Conv layers and concatenation operations [30]. To improve the speed and performance of feature aggregation, YOLOv11 uses the C3k2 block rather than the C2f block in the neck [31]. YOLOv11 builds on the model to pay more attention to critical visual features by adding spatial attention in the form of a C2PSA block, an option not present in its predecessor, YOLOv8 [32]. Placing a C2PSA block on the neck helps to increase the accuracy of the detection, particularly that of smaller or partially blocked objects, which increase the local feature representation [33].

Head: The head produces the final predictions of the YOLOv1 model [34], as in earlier iterations. It assigns classes to objects, quantifies their objectness (how similar they are to a feature of interest), and predicts their bounding boxes. YOLOv11 has several C3k2 blocks that process and refine feature maps effectively [35]. At the head, several pathways have C3k2 blocks rather than a CSP bottleneck. These blocks reveal visible characteristics at various scales and levels. This block helps to minimize the overhead of trainable parameters [34]. To further improve the feature maps, Yolov11 incorporated a CBS (Convolution–BatchNorm–Silu) [36]. After every block of C3k2, CBS plays an essential role in feature extraction and detection. It guarantees the successful transfer of feature maps to the next layer for bounding-box detection and classification [37].

3.1. Dataset

The dataset used in this study was GRAZPEDWRI-DX, published by Nagy et al. [15]. These data were gathered by the Department of Pediatric Diagnostic Radiology in the Medical University of Graz, Austria, by a quantitative study based on 10,643 X-ray examinations (wrist) and 6091 unique patient cases. The dataset comprises 20,327 images, 74,459 labels, and 67,771 objects annotated across 9 classes. Table 1 presents the number of instances per class. As can be seen from the table, the most common objects in the dataset are "fracture" and "text." This dataset is highly imbalanced, with the imbalance mainly reflected in the number of class instances and classes. Figure 2 presents a dataset histogram of class frequency.

For the experiments, the dataset is split into 70% (14,301), 20% (3997), and 10% (2029) for training, validation, and testing, respectively.

Table 1
Instances of objects for each class

Class name	Instances no.
Boneanomaly	276
bonelesion	45
foreignbody	8
fracture	18090
metal	818
periostealreaction	3453
Pronatorsign	567
softtissue	464
text	23722

3.2. Experiment

3.2.1. Environment setup

All experiments took place on the Google Colab Pro+ cloud service, utilizing an NVIDIA A100-SXM4-40GB GPU, 85 GB of RAM, and a 256 GB disk. They were used and executed with YOLOv11 Ultralytics version 8.3, which represents the official release of the YOLOv11 architecture [38], Python 3.11.1, torch 2.5.1, and CUDA version 12.4.

3.2.2. Matrices used

Each model's effectiveness was assessed using precision, recall, and mAP at 0.5 (mAP@50) and 0.5:0.95 (mAP@50–95) Intersection over Union (IoU) thresholds.

1) Intersection over Union

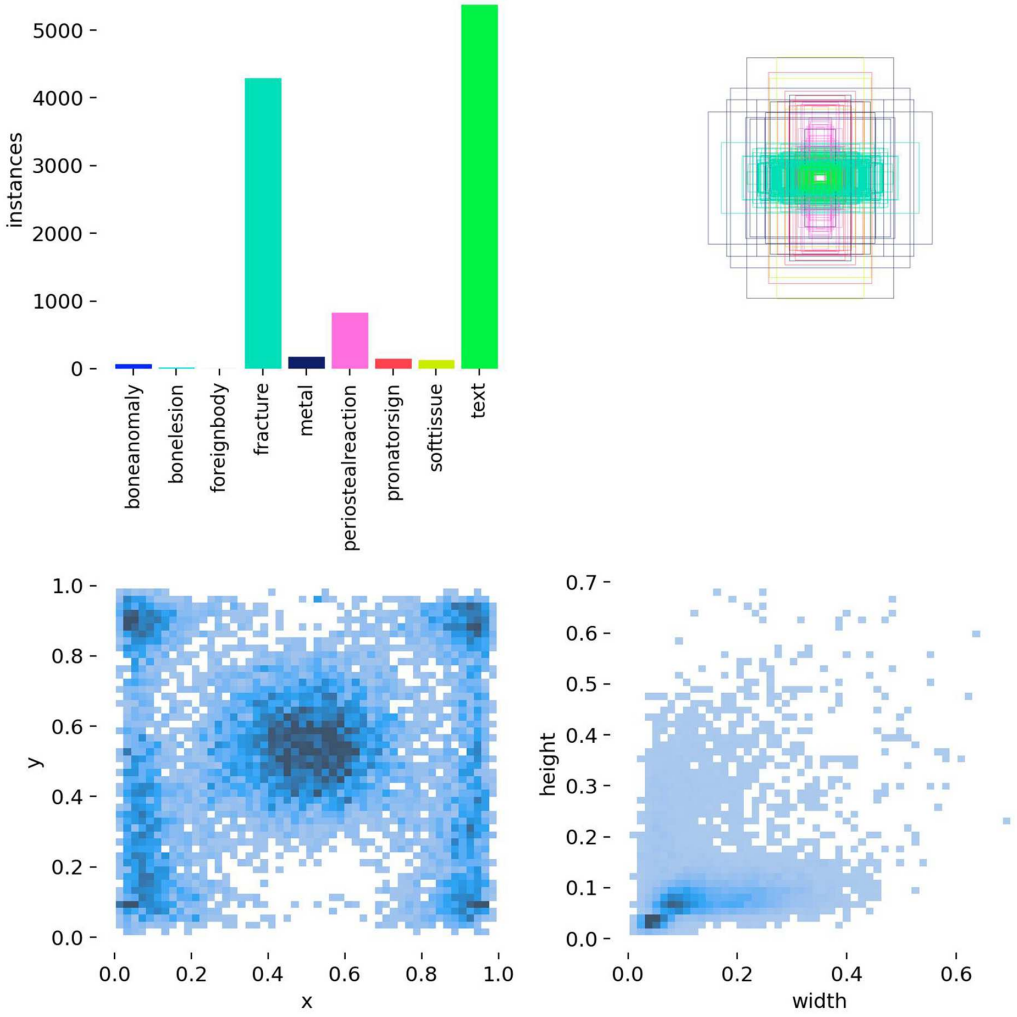
In computer vision object detection, IoU is a standard metric for assessing the detection accuracy [32]. IoU is used during YOLO training to associate ground-truth objects with anchor boxes. When evaluating the model, YOLO predicts multiple bounding boxes for the same object. Non-maximum suppression is applied to these overlapping boxes using the IoU metric to eliminate redundant boxes. If two predicted boxes have IoUs exceeding a specified threshold, the box with the higher confidence score is retained or dropped [32, 39]. The following equation represents the IoU:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{1}$$

Where:

The overlap area refers to the shared space within the actual and projected boundaries; the union area includes the entire region covered by both bounding boxes.

Figure 2
Class histogram distribution



2) Mean Average Precision

mAP measures how effectively a model predicts objects by combining precision and recall across multiple confidence thresholds [40]. Precision evaluates the proportion of true positives among all identified objects, whereas recall measures the proportion of true positives out of all actual objects in an image. These metrics are computed as follows:

$$Precision = \frac{True\ Positives}{True\ Positives\ + False\ Positives} \tag{2}$$

$$Recall = \frac{True\ Positives}{True\ Positives\ + False\ Negatives} \tag{3}$$

The mAP is determined by dividing the chosen object class's average precision (AP) by the total number of classes. The equation below computes the mean of AP values:

$$mAP = \frac{\sum AP_i}{N} \tag{4}$$

Where (N) in the equation is the number of object classes.

3.2.3. Training details

YOLOv11 has five variants: YOLOv11-n, YOLOv11-s, YOLOv11-m, YOLOv11-l, and YOLOv11-x. All experimentation on these variants and other YOLO predecessors used in this work is done using default settings and standard hyperparameters. Table 2 details all training parameters used in this work.

Every model utilized the "GRAZPEDWRI-DX" dataset detailed in the earlier section. Instead of training only at 640 × 640, we adopted multi-scale training (e.g., randomly resizing inputs between 480 and 800 during training). This helps the model generalize better to variable input sizes and capture both fine fracture lines (at high resolution) and overall structure (at low resolution).

There is a significant class imbalance in the GRAZPEDWRI-DX dataset, with fracture instances predominating in the label distribution. The effects of imbalance are somewhat mitigated by YOLO's anchor-based detection and confidence-thresholding, even though no explicit rebalancing technique (such as focal loss or oversampling) was used. However, minority classes showed lower recall, indicating a significant drawback. To increase the resilience of minority-class detection, future research will investigate focal loss, class-aware sampling, and synthetic augmentation.

Table 2
Training parameters

Hyperparameter	Value
Input size	640 × 640
Optimizer algorithm	SGD
Learning_rate (Initial)	0.01
Learning_rate (Final)	0.01
Momentum	0.937
Weight decay	0.0005
Epoch	100
batch size	32
multi_scale	True

3.3. Architectural rationale

Fracture lines in pediatric radiographs exhibit low contrast and fine discontinuity, making multi-scale feature aggregation and spatial attention important. The C3k2 block is more localized and has low computation redundancy and is spatially resolved with very restrictive inference times. Moreover, by integrating spatial attention modules, the network is able to concentrate on regions of anatomy of interest, which is required to decrease the quantity of fractures overlooked in high-noise X-ray images [21].

In order to support the point about the importance of methodological novelty and to emphasize the advantages of YOLOv11, we have performed an ablation study to assess the effect of basic architectural characteristics. We assessed the role of attention mechanisms by creating a modified version of YOLOv11-m, named No-SA-YOLOv11-m, generated in the absence of the spatial attention module (disabled). Table 3 presents the ablation analysis of the spatial attention module in YOLOv11-m for fracture detection in the GRAZPEDWRI-DX dataset.

The results in Table 3 indicate that integrating spatial attention modules enhances localization and recall strength, especially at more stringent IoU thresholds (mAP@50–95), suggesting that the architectural optimizations made in YOLOv11 are directly associated with better fracture detection results, not a nonsignificant difference in architecture.

4. Results and Discussion

Evaluation of the performance of the YOLOv11 models required an adequate evaluation, including a comparatively simple comparative analysis. To determine which model will prevail in terms of accuracy and performance, the models were compared against four of their immediate predecessors: YOLOv5, YOLOv8, YOLOv9, and YOLOv10. In this evaluation, the same GRAZPEDWRI-DX dataset was used as input for each model to determine which YOLO architecture identified more abnormalities in the dataset's wrist images.

4.1. YOLOv11 variant performance analysis

Table 4 presents the performance of all classes across various variants of YOLOv11. The findings revealed that YOLOv11-x performed best among the others, with the highest mAP (mAP@50–95) of 0.424. YOLOv11-l had lower mAP@50–95 (0.405) but higher recall (0.580), making it suitable for high-sensitivity applications.

Table 3
Ablation study evaluating the contribution of spatial attention in YOLOv11-m for the fracture class

Model-variant	Spatial attention	Precision(P)	Recall(R)	mAP@50	mAP@50–95
YOLOv11-m	Enabled	0.850	0.904	0.939	0.569
No-SA-YOLOv11-m	Disabled	0.836	0.886	0.931	0.552

Table 4
YOLOv11 variant evaluation on all classes

Model-variant	Precision(P)	Recall(R)	mAP@50	mAP@50–95	Params (M)
YOLOv11-n	0.642	0.539	0.567	0.370	2.5
YOLOv11-s	0.782	0.519	0.564	0.368	9.4
YOLOv11-m	0.620	0.534	0.618	0.414	20.0
YOLOv11-l	0.626	0.580	0.623	0.405	25.3
YOLOv11-x	0.693	0.566	0.620	0.424	56.9

Interestingly, YOLOv11-s was high in precision (0.782) but low in the power to reduce false positives, whereas recall was lower than in both YOLOv11-x and YOLOv11-l.

In terms of model size, the most parameters are in YOLOv11-x (56.9M), whereas the lightest is YOLOv11-n (2.5M). The recall and precision of smaller models are lower than those of larger models, indicating a trade-off between performance and model size. Using a YOLOv11 variant is a trade-off between accuracy and recall, determined by usage requirements.

4.2. YOLOv11 performance on fracture class

Table 5 is an evaluation of the performance of the YOLOv11 variants in terms of fracture classes. Fracture detection performance was much higher than the overall class rating, with all model variants achieving mAP@50 values above 0.93, ranging from 0.939 to 0.941. YOLOv11-m achieves the highest recall (0.904) and mAP@50–95 (0.569), indicating good overall performance in fracture detection across different IoU thresholds. On the other hand, YOLOv11-s was the most accurate variant (0.902) and mAP@50 (0.941), suggesting it is the most effective, as it limits the number of false positives.

YOLO11-n and YOLO11-m have the shortest inference time (1.1 ms), which is appropriate to use in real-time, whereas YOLO11-x is highly accurate but requires more time to be inferred (3.7 ms). The mAP@50–95 value indicates a small but significant increase in YOLOv11-m (0.569), supporting its ability to detect fractures across different intersections at various union (IoU) levels.

4.3. Comparison with YOLO predecessor

Based on Table 5 results, which show that YOLOv11-m is superior to other variants in detecting fractures across varying IoU thresholds, the performance of YOLOv11-m was inspected by quantifying the performance of other YOLO predecessors, such as YOLO versions 5,8,9, and 10 (m) variants, as shown in Tables 6 and 7.

As shown in Table 6, YOLOv11-m achieved a remarkable mAP@50–95 (0.414), higher than YOLOv5-m (0.368), YOLOv8-m (0.392), YOLOv9-m (0.397), and YOLOv10-m (0.382). This improvement reflects the generalizability and accuracy of YOLOv11-m among various wrist abnormalities. However, it has a moderate trade-off in precision (0.620) and recall (0.534) compared to YOLO8-m, which has a slightly higher recall (0.619) but lower mAP scores. Though YOLOv11-m's precision (0.620) was lower than YOLOv5-m's (0.727), indicating that YOLOv11-m provides improved overall detection accuracy, its false positive rate may be slightly higher. Regarding model size, YOLOv11-m has 20.0M parameters, which are compact and entail much lower computational cost than YOLOv5-m (41.2M) with mAP 0.325.

YOLOv11-m showed competitive performance for fracture detection, with mAP@50 of 0.939, as shown in Table 6. This performance is comparable to YOLOv8-m (0.939) and is very close to that of YOLOv9-m (0.943) and YOLOv5-m (0.941). Only YOLOv11-m achieved the highest mAP@50–95 score (0.569), indicating consistent performance across various thresholds. Its precision (0.850) was somewhat lower than that of YOLOv5-m (0.916) and YOLOv8-m (0.879), but its high recall (0.904) indicates that more fractures are correctly detected, reducing the risk of false negatives. Moreover, from an efficiency perspective, YOLO11-m has the shortest inference time (1.1 ms), which is substantially faster than its predecessors and remains competitive in accuracy.

Though the total changes in mAP@50–95 are modest, they are clinically significant in high-sensitivity applications such as fracture detection, where even small increases in recall can reduce the risk of false diagnoses. Previous research has demonstrated that even small increases in model mAP in medical imaging can result in statistically significant reductions in diagnostic error when implemented at scale.

The clinical implications of a false negative in diagnosing fractures in pediatrics are more severe than those of a false positive. Loss of fractures may lead to incomplete healing, growth difficulties, delayed treatment, and long-term functional difficulties. In this way, pediatric detection systems must be characterized by sensitivity and recall to reduce missed injuries. YOLOv11-m shows high recall (0.904), demonstrating its strong ability to detect actual fracture cases. False positives can be effectively addressed through radiologist review, especially in assisted diagnostic

Table 5
YOLOv11 variant evaluation on fracture class

Model-variant	Precision(P)	Recall(R)	mAP@50	mAP@50–95	Inference (ms)
YOLOv11-n	0.885	0.883	0.932	0.551	1.1
YOLOv11-s	0.902	0.883	0.941	0.565	1.3
YOLOv11-m	0.850	0.904	0.939	0.569	1.1
YOLOv11-l	0.887	0.891	0.934	0.559	2.4
YOLOv11-x	0.877	0.903	0.939	0.566	3.7

Table 6
Performance comparison of YOLOv11-m and its predecessor for all classes

Model-variant	Precision(P)	Recall(R)	mAP@50	mAP@50–95	Params (M)
YOLOv5-m	0.727	0.542	0.568	0.368	41.2
YOLOv8-m	0.663	0.619	0.608	0.392	25.9
YOLOv9-m	0.681	0.554	0.601	0.397	20.1
YOLOv10-m	0.656	0.495	0.562	0.382	16.4
YOLOv11-m	0.620	0.534	0.618	0.414	20.0

Table 7
Comparison of performance between YOLOv11-m and its predecessor for the fracture class

Model-variant	Precision(P)	Recall(R)	mAP@50	mAP@50–95	Inference (ms)
YOLOv5-m	0.916	0.887	0.941	0.562	1.8
YOLOv8-m	0.879	0.909	0.939	0.561	1.7
YOLOv9-m	0.867	0.906	0.943	0.566	2.2
YOLOv10-m	0.866	0.899	0.933	0.560	1.9
YOLOv11-m	0.850	0.904	0.939	0.569	1.1

Table 8
Comparison of performance between YOLOv11-m and other improved (m) versions from articles for the fracture class

Model	Precision(P)	Recall(R)	mAP@50	mAP@50–95
YOLOv11-m-GAM [41]	0.731	0.601	0.618	0.383
YOLOv11-m-RES-GAM [41]	0.623	0.639	0.637	0.394
YOLOv11-m (Ours)	0.620	0.534	0.618	0.414

Table 9
Comparison of performance between YOLOv11-m and reported performance numbers from recent papers

Model	Precision(P)	mAP@50	mAP@50–95
YOLOv8+GC [19]	0.660	0.663	0.428
YOLOv8+ECA [42]	0.898	0.643	0.416
ASC-YOLO [43]	0.655	0.611	0.402
RT-DETR [43]	0.624	0.533	0.340
Faster R-CNN [38]	0.730	0.350	0.575
YOLOv11-m (Ours)	0.620	0.618	0.414

workflows, despite its slightly lower precision compared to earlier models. This trade-off aligns with clinical standards, in which expert interpretation complements rather than replaces decision-support systems, making it suitable for emergency department triage and initial screening, given the need for rapid identification of potential fractures.

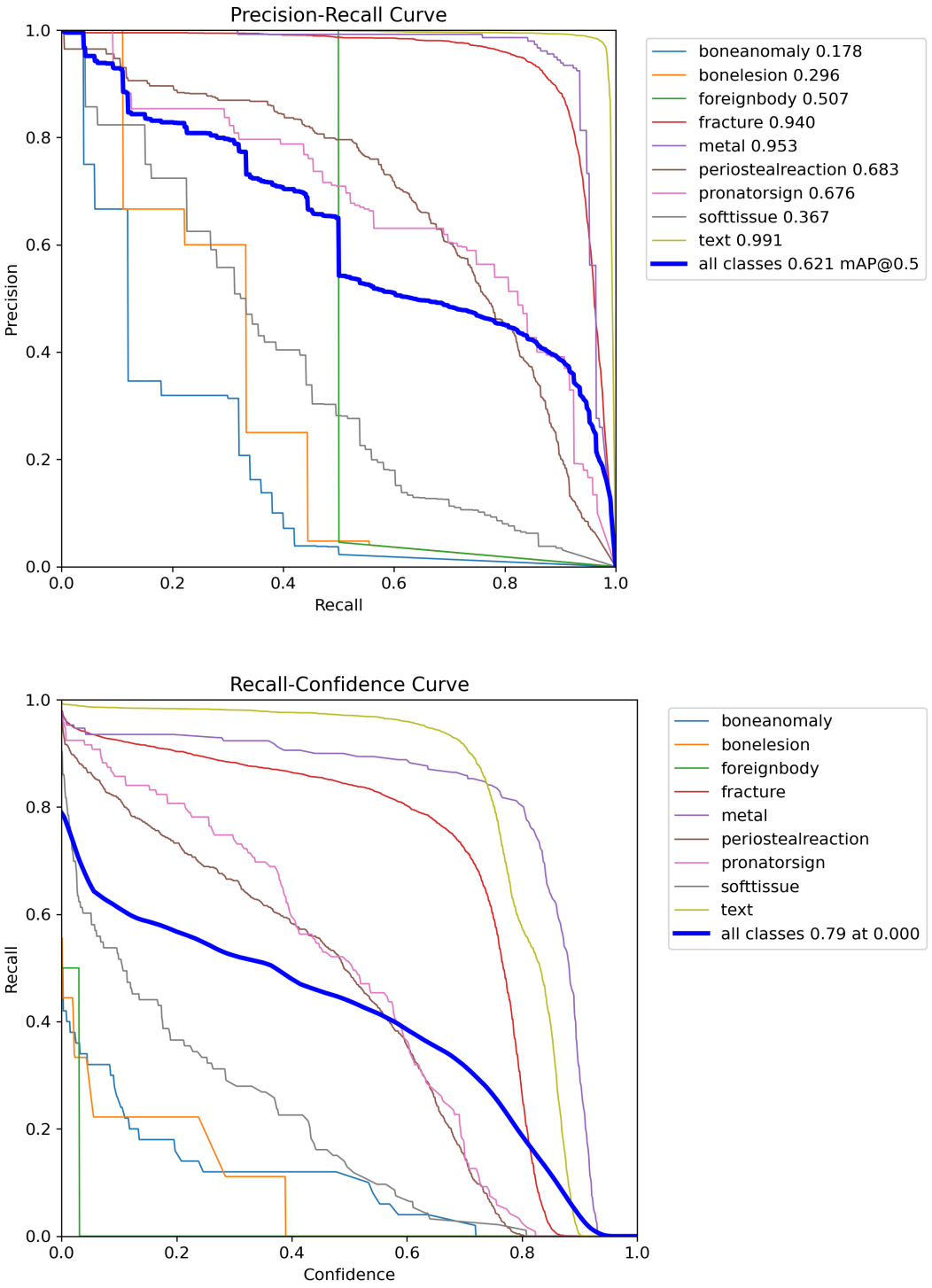
Table 8 presents a comparative analysis of the YOLO11-m model and several enhanced (m) versions for wrist fracture detection on the GRAZPEDWRI-DX dataset. Results show that integrating attention mechanisms (e.g., GAM and RES-GAM) improves recall and mAP, with YOLO11-m-RES-GAM achieving the highest mAP@50 (0.637). The G-YOLOv11-m model achieves the highest precision (0.768). Overall, enhancements to YOLO11-m can improve detection accuracy, especially in recall and localization.

Table 9 presents a comparative analysis of the YOLO11-m model and other reported performance numbers from recent papers for wrist fracture detection on the GRAZPEDWRI-DX dataset. Figure 3 illustrates the precision–recall and recall–confidence curves for both overall and individual classes utilizing the YOLOv11-m on the test set. In contrast, Figure 4 shows the loss curves for different YOLOv11-m variants during training, plotted as recall against confidence at various thresholds. A precision–recall curve helps visualize the change in precision and recall values at different classification thresholds. It helps us optimize different metrics using our use case by finding the best threshold value. Additionally, Figure 5 presents the confusion matrix for the YOLOv11-m variant.

The YOLOv11-m confusion matrix, as shown in Figure 5, shows classification performance across all dataset classes. While misclassifications mostly occur within minority classes due to dataset imbalance, the model's high true-positive rate for the fracture class indicates strong sensitivity.

Figure 6 shows the inference results on the YOLOv11-m variant with their corresponding ground truth. The results show the most promising fracture detection performance, with

Figure 3
YOLOv11-m precision–recall, recall–confidence curves



high confidence levels and the ability to differentiate different pathologies.

Figure 7 presents Grad-CAM visualizations for representative pediatric wrist X-ray images using the YOLOv11-m model. The overall alterations in mAP@50–95 are relatively unimportant, but they hold clinical importance in high-sensitivity uses of the system, as fracture detection, where even minor increases in recall can decrease the risk of false diagnoses. Past studies have shown that even minor changes in model mAP in medical imaging may lead to statistically significant diagnostic error reduction when used at scale.

In the case of YOLOv11, it turns out that there is a significant trade-off between competence and efficiency in the detection process. According to the mAP@50–95 values, YOLOv11-x and YOLOv11-m have quite good performance at different thresholds of IoU.

Figure 4
YOLOv11-m loss curves

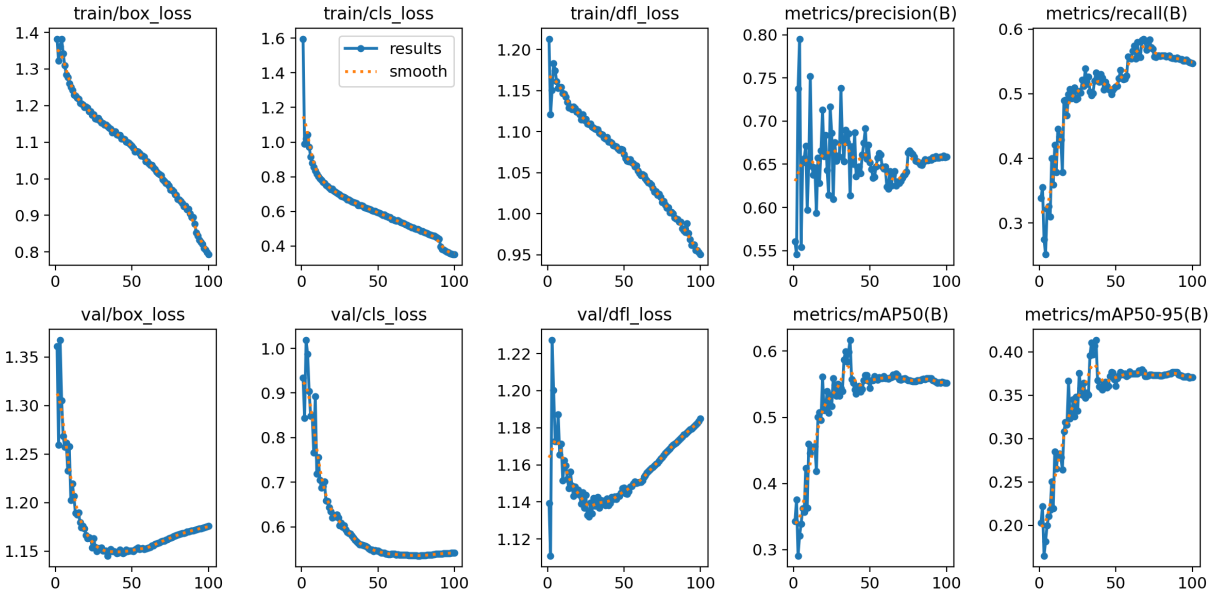


Figure 5
YOLOv11-m confusion matrix

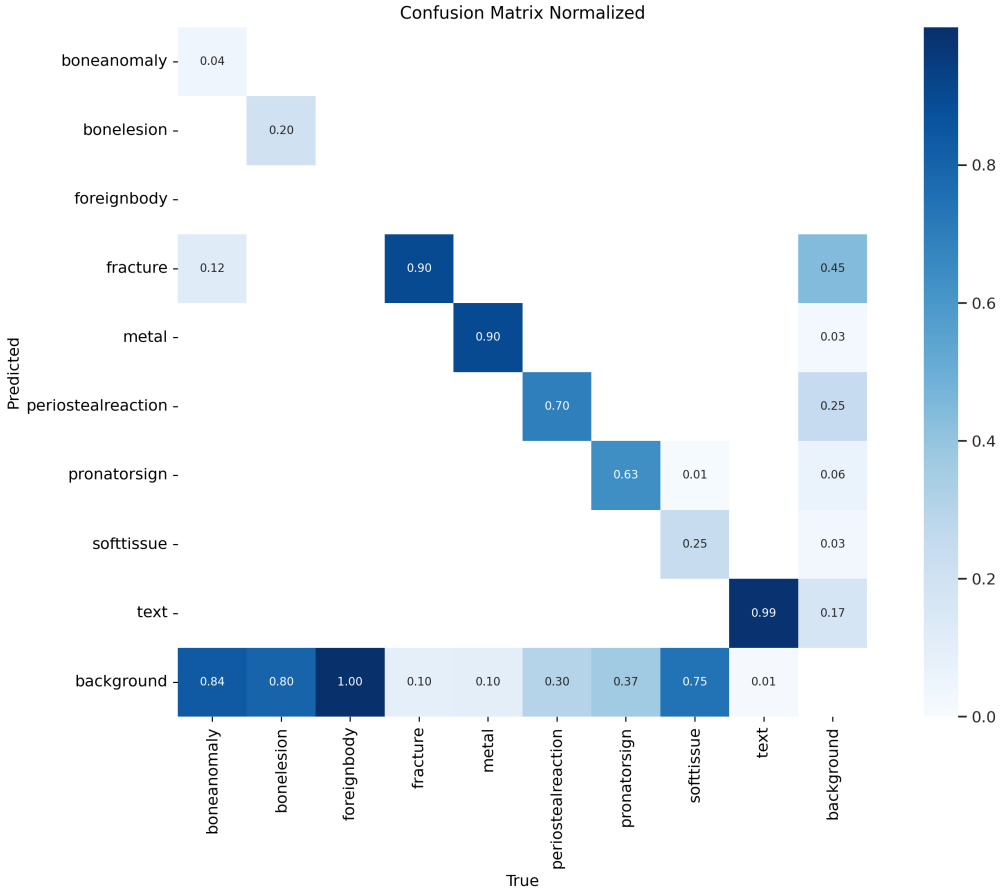


Figure 6
Inference of YOLOv11-m variant

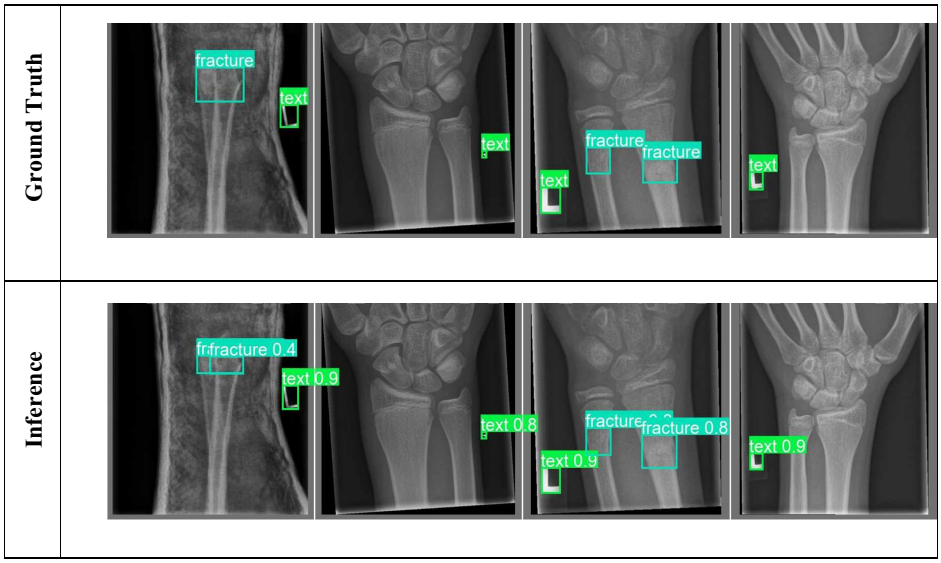
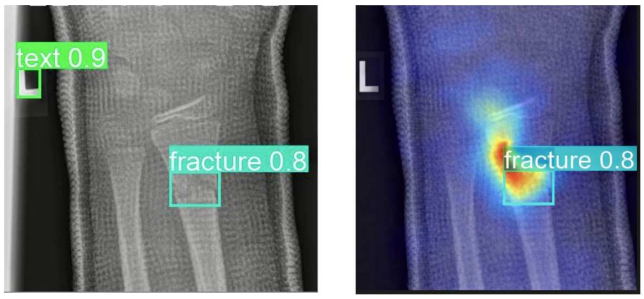


Figure 7
Grad-CAM visualizations for pediatric wrist fracture using the YOLOv11-m model



5. Conclusion

This study determined the performance of YOLOv11 and its variants in detecting abnormalities and fractures on the pediatric wrist using the GRAZPEDWRI-DX dataset. The results of the experiments indicated that the performance measures, such as inference time, recall, precision, and mAP for YOLOv11, were comparable to those of the best predecessors, 5-m, 8-m, 9-m, and 10-m. The best achieved of mean average precision (mAP@50–95) of the YOLOv11 variants was seen in YOLOv11-m (0.569) and YOLOv11-x (0.424) across all classes. YOLOv11 shows superior performance in fracture detection and inference times compared to earlier YOLO models, indicating improved detection capabilities. The comparative findings also reveal that certain architectural additions, including attention processing and residual connections, can remarkably enhance the level of precision as well as maintain a balance between recall and precision. The size-to-accuracy ratio of the model indicates that YOLOv11 can be applied to satisfy the needs of any application; smaller variants perform better in real time, whereas larger ones deliver better accuracy when it comes to high-precision tasks, for example, medical image analysis. YOLOv11-m can be an effective intervention in clinical practice, as it can be used as a triage tool in emergency rooms, and suspected fractures can be identified as soon as the images are taken. The recollection of high plays a crucial role in the diagnostics of children, in which unidentified fractures may cause either growth retardation or chronic deformity. The system is meant to assist, not to substitute for radiologists, because they will be given precedence in suspicion cases, which will undergo expedited reviews. To ensure safety, robustness, and generalizability, they would require future multicenter validation and regulatory approval, for example, FDA clearance or CE marking, before deployment. Future work may optimize the YOLOv11 architecture to enhance recall while maintaining precision, particularly for thorough class recognition. Additionally, investigating hybrid methods that integrate attention mechanisms or feature fusion techniques could enhance performance.

Ethical Statement

The authors declare that this study did not require formal ethical approval because it utilizes the GRAZPEDWRI-DX dataset, a publicly available, de-identified secondary dataset originally published by Nagy, E., Janisch, M., Hržić, F., Sorantin, E., and Tschauner, S. (2022). A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning. Scientific data, 9(1), 222. https://doi.org/10.1038/s41597-022-01328-z, reference number [15]. As the dataset consists of fully anonymized pediatric X-ray images that do not contain personally identifiable information, this research is exempt from Institutional Review Board or ethics committee approval.

Conflicts of Interest

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

Data Availability Statement

The data that support the findings of this study are openly available at https://doi.org/10.1038/s41597-022-01328-z, reference number [15].

Author Contribution Statement

Muhanad Abdul Elah Alkhalisy: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Qusay Shihab Hamad:** Validation, Investigation, Writing – review & editing, Visualization, Supervision, Project administration. **Ali Retha Hasoon Khayeat:** Validation, Investigation, Writing – review & editing. **Shahrel Azmin Suandi:** Writing – review & editing, Supervision.

References

[1] Hedström, E. M., Svensson, O., Bergström, U., & Michno, P. (2010). Epidemiology of fractures in children and adolescents: Increased incidence over the past decade: A population-based study from northern Sweden. *Acta Orthopaedica*, *81*(1), 148–153. https://doi.org/10.3109/17453671003628780

[2] The Diagnostic Error in Medicine 12th Annual International Conference. (2019). *Diagnosis*, *6*(4), eA1–eA96. https://doi.org/10.1515/dx-2019-0075

[3] Hamad, Q. S., Samma, H., Suandi, S. A., & Saleh, J. M. (2022). Study of VGG-19 depth in transfer learning for COVID-19 X-Ray image classification. In *Proceedings of the 11th International Conference on Robotics, Vision, Signal Processing and Power Applications: Enhancing Research and Innovation through the Fourth Industrial Revolution*, 930–935. https://doi.org/10.1007/978-981-16-8129-5_142

[4] Hamad, Q. S., Samma, H., & Suandi, S. A. (2023). Feature selection of pre-trained shallow CNN using the QLESCA optimizer: COVID-19 detection as a case study. *Applied Intelligence*, *53*(15), 18630–18652. https://doi.org/10.1007/s10489-022-04446-8

[5] Li, L., Liu, Z., Huang, H., Lin, M., & Luo, D. (2019). Evaluating the performance of a deep learning-based computer-aided diagnosis (DL-CAD) system for detecting and characterizing lung nodules: Comparison with the performance of double reading by radiologists. *Thoracic Cancer*, *10*(2), 183–192. https://doi.org/10.1111/1759-7714.12931

[6] Williams, B. A., Palumbo, N. E., Phillips, S. A., & Blakemore, L. C. (2020). What they want-Caregiver and patient immobilization preferences for pediatric buckle fractures of the wrist. *The Iowa Orthopaedic Journal*, *40*(1), 83–90.

[7] Guermazi, A., Tannoury, C., Kompel, A. J., Murakami, A. M., Ducarouge, A., Gillibert, A., . . . , & Hayashi, D. (2022). Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*, *302*(3), 627–636. https://doi.org/10.1148/radiol.210937

[8] Kalmet, P. H., Sanduleanu, S., Primakov, S., Wu, G., Jochems, A., Refaee, T., . . . , & Poeze, M. (2020). Deep learning in fracture detection: A narrative review. *Acta Orthopaedica*, *91*(2), 215–220. https://doi.org/10.1080/17453674.2019.1711323

[9] Duron, L., Ducarouge, A., Gillibert, A., Lainé, J., Allouche, C., Cherel, N., . . . , & Feydy, A. (2021). Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: A multicenter cross-sectional diagnostic study. *Radiology*, *300*(1), 120–129. https://doi.org/10.1148/radiol.2021203886

[10] Prameswari, F., Octafiani, H., & Haryanto, T. (2024). You Only Look Once (YOLOv8) for fish species detection. In *IOP Conference Series: Earth and Environmental Science, 1359*(1), 012023. https://doi.org/10.1088/1755-1315/1359/1/012023

[11] Ragab, M. G., Abdulkadir, S. J., Muneer, A., Alqushaibi, A., Sumiea, E. H., Qureshi, R., . . . , & Alhussian, H. (2024). A comprehensive systematic review of YOLO for medical object detection (2018 to 2023). *IEEE Access*, *12*, 57815–57836. https://doi.org/10.1109/ACCESS.2024.3386826

[12] Meza, G., Ganta, D., & Gonzalez Torres, S. (2024). Deep learning approach for arm fracture detection based on an improved YOLOv8 algorithm. *Algorithms*, *17*(11), 471. https://doi.org/10.3390/a17110471

[13] Sundaresan Geetha, A., Alif, M. A. R., Hussain, M., & Allen, P. (2024). Comparative analysis of YOLOv8 and YOLOv10 in vehicle detection: Performance metrics and model efficacy. *Vehicles*, *6*(3), 1364–1382. https://doi.org/10.3390/vehicles6030065

[14] Jeon, Y.-D., Kang, M.-J., Kuh, S.-U., Cha, H.-Y., Kim, M.-S., You, J.-Y., . . . , & Yoon, D.-K. (2024). Deep learning model based on You Only Look Once algorithm for detection and visualization of fracture areas in three-dimensional skeletal images. *Diagnostics*, *14*(1), 11. https://doi.org/10.3390/diagnostics14010011

[15] Nagy, E., Janisch, M., Hržić, F., Sorantin, E., & Tschauner, S. (2022). A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning. *Scientific Data*, *9*(1), 222. https://doi.org/10.1038/s41597-022-01328-z

[16] Dibo, R., Galichin, A., Astashev, P., Dylov, D. V., & Rogov, O. Y. (2024). DeepLOC: Deep learning-based bone pathology localization and classification in wrist x-ray images. In *11th International Conference on Analysis of Images, Social Networks and Texts*, 199–211. https://doi.org/10.1007/978-3-031-54534-4_14

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

[18] Till, T., Tschauner, S., Singer, G., Lichtenegger, K., & Till, H. (2023). Development and optimization of AI algorithms for wrist fracture detection in children using a freely available dataset. *Frontiers in Pediatrics*, *11*, 1291804. https://doi.org/10.3389/fped.2023.1291804

[19] Ju, R.-Y., Chien, C.-T., Lin, C.-M., & Chiang, J.-S. (2024). Global context modeling in YOLOv8 for pediatric wrist fracture detection. In *2024 International Symposium on Intelligent Signal Processing and Communication Systems*, 1–5. https://doi.org/10.1109/ISPACS62486.2024.10869064

[20] Chien, C.-T., Ju, R.-Y., Chou, K.-Y., & Chiang, J.-S. (2024). YOLOv9 for fracture detection in pediatric wrist trauma X-ray images. *Electronics Letters*, *60*(11), e13248. https://doi.org/10.1049/ell2.13248

[21] Ahmed, A., Imran, A. S., Kastrati, Z., Daudpota, S. M., Ullah, M., & Noor, W. (2024). Learning from the few: Fine-grained approach to pediatric wrist pathology recognition on a limited dataset. *Computers in Biology and Medicine*, *181*, 109044. https://doi.org/10.1016/j.compbiomed.2024.109044

[22] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. In *Proceedings of the 38th International Conference on Neural Information Processing*, 107984–108011.

[23] Bumbaca, S., & Borgogno-Mondino, E. (2025). On the minimum dataset requirements for fine-tuning an object detector

for arable crop plant counting: A case study on maize seedlings. *Remote Sensing*, *17*(13), 2190. https://doi.org/10.3390/rs17132190

[24] Jima, W. D., Desta, S. F., Tarekegn, T. A., Gebremedhin, G. S., Gutema, A. B., & Debelee, T. G. (2025). Tsetse fly detection and sex classification model enrichment employing YOLOv8 and YOLO11 architecture. *Applied AI Letters*, *6*(3), e70004. https://doi.org/10.1002/ail2.70004

[25] Xue, Z., Lin, H., & Wang, F. (2022). A small target forest fire detection model based on YOLOv5 improvement. *Forests*, *13*(8), 1332. https://doi.org/10.3390/f13081332

[26] Zayani, H. M., Kachoukh, A., Ghodhbani, R., Khediri, N., Abd-Elkawy, E. H., Ammar, I., ..., & Saidani, T. (2025). Cat breed classification with YOLOv11 and optimized training. *Engineering, Technology & Applied Science Research*, *15*(2), 21652–21657. https://doi.org/10.48084/etasr.10218

[27] Xue, Z., Kong, L., Wu, H., & Chen, J. (2025). Fire and smoke detection based on improved YOLOV11. *IEEE Access*, *13*, 73022–73040. https://doi.org/10.1109/ACCESS.2025.3564434

[28] Shen, P., Mei, K., Cao, H., Zhao, Y., & Zhang, G. (2025). LDDFSF-YOLO11: A lightweight insulator defect detection method focusing on small-sized features. *IEEE Access*, *13*, 90273–90292. https://doi.org/10.1109/ACCESS.2025.3569970

[29] Wang, Q., & Liu, Z. (2025). FEFM-YOLO11: Underwater object detection algorithm based on improved YOLO11. *Neural Processing Letters*, *57*(5), 82. https://doi.org/10.1007/s11063-025-11805-2

[30] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A review of Yolo algorithm developments. *Procedia Computer Science*, *199*, 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135

[31] Bi, J., Li, K., Zheng, X., Zhang, G., & Lei, T. (2025). SPDC-YOLO: An efficient small target detection network based on improved YOLOv8 for drone aerial image. *Remote Sensing*, *17*(4), 685. https://doi.org/10.3390/rs17040685

[32] Gong, M., Wang, D., Zhao, X., Guo, H., Luo, D., & Song, M. (2021). A review of non-maximum suppression algorithms for deep learning target detection. In *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, *11763*, 1176332. https://doi.org/10.1117/12.2586477

[33] Zhao, Q., & Zhu, J. (2025). An Improved YOLOv11 architecture with multi-scale attention and spatial fusion for fine-grained residual detection. *Results in Engineering*, *27*, 107061. https://doi.org/10.1016/j.rineng.2025.107061

[34] Huang, J., Wang, K., Hou, Y., & Wang, J. (2025). LW-YOLO11: A lightweight arbitrary-oriented ship detection method based on improved YOLO11. *Sensors*, *25*(1), 65. https://doi.org/10.3390/s25010065

[35] He, L., Zhou, Y., Liu, L., & Ma, J. (2024). Research and application of YOLOv11-based object segmentation in intelligent recognition at construction sites. *Buildings*, *14*(12), 3777. https://doi.org/10.3390/buildings14123777

[36] Cao, X., Zhang, X., Hou, Y., Lu, Z., Liu, R., & Qin, W. (2024). Small object detection in drone scenes based on improved YOLO. In *2024 9th International Conference on Image, Vision and Computing*, 153–156. https://doi.org/10.1109/ICIVC61627.2024.10837527

[37] Tong, L., Fan, C., Peng, Z., Wei, C., Sun, S., & Han, J. (2024). WTBD-YOLOV8: An improved method for wind turbine generator defect detection. *Sustainability*, *16*(11), 4467. https://doi.org/10.3390/su16114467

[38] Wang, R., Zhao, J., Liu, X., Tang, X., Shi, Y., & Wei, L. (2024). *AdvYOLO: Advanced YOLOv8 application for bone pathology localization and classification in wrist X-ray images*. Research Square. https://doi.org/10.21203/rs.3.rs-4051336/v1

[39] Wang, W., Li, X., Lyu, X., Zeng, T., Chen, J., & Chen, S. (2023). Multi-attribute NMS: An enhanced non-maximum suppression algorithm for pedestrian detection in crowded scenes. *Applied Sciences*, *13*(14), 8073. https://doi.org/10.3390/app13148073

[40] Aishwarya, N., Prabhakaran, K. M., Debebe, F. T., Reddy, M. S. S. A., & Pranavee, P. (2023). Skin cancer diagnosis with YOLO deep neural network. *Procedia Computer Science*, *220*, 651–658. https://doi.org/10.1016/j.procs.2023.03.083

[41] Tariq, M., & Choi, K. (2025). YOLO11-driven deep learning approach for enhanced detection and visualization of wrist fractures in X-ray images. *Mathematics*, *13*(9), 1419. https://doi.org/10.3390/math13091419

[42] Amirouche, F., Prosper, A. M., & Mzeihem, M. (2025). Enhancing pediatric distal radius fracture detection: Optimizing YOLOv8 with advanced AI and machine learning techniques. *BMC Medical Imaging*, *25*(1), 316. https://doi.org/10.1186/s12880-025-01669-2

[43] Du, S., & Wei, Y. (2025). ASC-YOLO: Multi-scale feature fusion and adaptive decoupled head for fracture detection in medical imaging. *Applied Sciences*, *15*(16), 9031. https://doi.org/10.3390/app15169031

How to Cite: Alkhalisy, M. A. E., Hamad, Q. S., Khayeat, A. R. H., & Suandi, S. A. (2026). Performance Analysis of YOLOv11-m and Related Architectures in Pediatric X-Ray Fracture Detection. *Artificial Intelligence and Applications*, *4*(4), 549–560. https://doi.org/10.47852/bonviewAIA62027255