Enhancing Wrist Fracture Detection with YOLO: Analysis of State-of-the-art Single-stage Detection Models

ARTICLE INFO

Keywords:
fracture detection
object detection
medical imaging
pediatric X-ray
deep learning
YOLO

ABSTRACT

Wrist fractures (specifically in the distal radius and ulna) happen frequently among children, teenagers, and young adults—often during their growth spurt. But there aren't enough radiologists or specialists to quickly and accurately review all the X-rays, especially in some areas where expert services are limited. To tackle this, researchers are looking for automated methods to detect wrist fractures using **object detection** (a type of computer vision that spots and locates items in images). Most older studies used a two-step approach called **Faster R-CNN**, but this paper focuses on **single-step (single-stage) detection methods**: YOLOv5, YOLOv6, YOLOv7, and YOLOv8. These models scan an X-ray in one go to find potential fractures. Their experiments show that these newer, single-step YOLO models do better at detecting fractures than the two-step Faster R-CNN method. Among the different versions of YOLO. **YOLOv8m** gave the **best fracture detection** (it correctly identified 92% of fractures and had a mean average precision of 95%). **YOLOv6m** did slightly better when considering all types of abnormalities (it had the highest overall sensitivity at 83% for all classes). **YOLOv8x** scored the highest overall mean average precision (77%) for all classes on a specialized paediatric wrist X-ray dataset.

1. Introduction

Wrist abnormalities are common among children, adolescents, and young adults, with fractures of the distal radius and ulna being the most prevalent, particularly during puberty. Early diagnosis and treatment of these fractures are crucial to prevent long-term complications. Digital radiography (X-ray) is the primary imaging modality used for wrist injuries, though MRI, CT, or ultrasound may be necessary when X-rays fail to provide a clear diagnosis.

Interpreting wrist radiographs requires specialized expertise, but many medical professionals lack the necessary training. Studies indicate that diagnostic errors in emergency X-ray readings can be as high as 26%, exacerbated by a global shortage of radiologists. This shortage is projected to worsen due to a growing demand for imaging studies, which increases by 5% annually, while the availability of radiology residency positions grows by only 2%.

Recent advancements in computer vision and object detection have shown promise in medical imaging, particularly for trauma X-rays. Early methods like the sliding window approach divided an image into overlapping regions, classifying each section independently. However, this method was computationally expensive. Region-based methods improved efficiency by generating and classifying only relevant object regions.

Modern object detection techniques include:

1. Single-Stage Detection (YOLO, SSD): **A fast, efficient approach that predicts bounding boxes and class probabilities in a** single pass**.**

2. Two-Stage Detection (Faster R-CNN): **A more accurate but slower approach that first generates object proposals and then refines them.**

While **two-stage detection** is widely used for detecting wrist abnormalities, **single-stage detection** has received less attention in this domain. This study evaluates the effectiveness of state-of-the-art (**SOTA**) single-stage detectors for detecting wrist abnormalities using the **GRAZPEDWRI-DX dataset**.

Wrist fractures are just one type of wrist abnormality. Other conditions, such as **Carpal Tunnel Syndrome (CTS), Ganglion Cysts, Osteoarthritis, and Tendinitis**, present distinct imaging challenges. Our objective is to detect key abnormalities such as **fractures, periosteal reactions, bone anomalies, soft tissue changes, and foreign bodies**, rather than diagnosing the underlying conditions directly.

By leveraging **deep learning-based detection models**, this study aims to enhance wrist abnormality identification, reducing reliance on expert radiologists and improving diagnostic accuracy.

1.1. Study Objective & Research Questions

The primary objective of this study is to test the effectiveness of the state-of-the-art YOLO detection models, YOLOv5, YOLOv6, YOLOv7, and YOLOv8 on a comprehensively annotated dataset "GRAZPEDWRI-DX" recently released to the public. We compare the performances of all variants within each YOLO model employed to see whether the use of a compound-scaled version of the same architecture improves its performance. Moreover, this study also investigates how effective these single-stage detection methods are in detecting fractures compared to a two-stage detection method widely used in the past. In addition to conducting object detection across multiple classes, we also evaluate the performance of a conventional CNN in binary classification, specifically in distinguishing between fractures and non-fractures. We hypothesize that fractures

in the near vicinity of the wrist in pediatric X-ray images can be detected efficiently using YOLO models proposed by ultralytics (2022), Li, Li, Jiang, Weng, Geng, Li, Ke, Li, Cheng, Nie, Li, Zhang, Liang, Zhou, Xu, Chu and Wei (2022), Wang, Bochkovskiy and Liao (2022), and ultralytics (2023) respectively.

We analyze the potential of utilizing object detection techniques in answering the following research questions (RQ):

1. To what extent do state-of-the-art YOLO object detec- tion models effectively detect fractures in the vicinity of the wrist in pediatric X-ray images?
2. In the analysis of wrist images, do the single-stage detection models outperform a two-stage detection model widely used in the past?
3. Does the use of compound scaled variants within each YOLO algorithm improve its performance in detecting fractures?
4. To what extent can the YOLO surpass conventional CNN architecture and DenseNets in terms of sensitiv- ity in fracture recognition?

1.2. Contribution

The major contributions of this article are as follows:

- A thorough performance assessment of SOTA YOLO detection models on the newly released GRAZPEDWRI- DX dataset, a large and diverse set of pediatric X-ray images. To the best of our knowledge, this is the first study of its kind.
- An in-depth comparison of the performance of various variants within each YOLO model utilized.
- Achieved state-of-the-art mean average precision (mAP) score on the GRAZPEDWRI-DX dataset.

2. Related Work

Fracture detection is a crucial aspect in the field of wrist trauma, and computer vision techniques have played a significant role in advancing the research in this area. This section provides a comprehensive overview of the existing studies on fracture detection and highlights the key findings. The studies are divided into two subheadings: "Two-stage detection" and "One-stage detection". The first subheading covers studies that have used two-stage detection techniques, while the second subheading focuses on studies that have only employed single-stage detection algorithms.

2.1. Two-stage detection

The detection of bone abnormalities, including fracture detection, has been widely studied in the literature, mainly using two-stage detection algorithms.A study applied a Faster R-CNN model with a VGG16 backbone to identify distal radius fractures in anteroposterior wrist X-ray images. The model achieved a mAP of 0.87 when tested on 1,312 images. The initial dataset consisted of 95 anteroposterior images, with and without fractures, which were then augmented for training and testing.Another study developed two Faster R-CNN models with Inception-ResNet for frontal and lateral projections of wrist images. The models were trained on 6,515 frontal projection images and 6,537 lateral projection images.

The frontal model detected 91% of fractures, with a specificity of 0.83 and a sensitivity of 0.96.The lateral model detected 96% of fractures, with a specificity of 0.86 and a sensitivity of 0.97.Both models had high AUC-ROC values, with the frontal model achieving 0.92 and the lateral model 0.93.The overall per-study specificity was 0.73, sensitivity was 0.98, and AUC was 0.89.A two-stage R-CNN method was used to achieve an average precision (AP) of 0.62 on approximately 4,000 X-ray images of arm fractures from the MURA dataset. A study introduced ParallelNet, a two-stage R-CNN network with a TripleNet backbone, for thigh fracture detection in a dataset of 3,842 thigh X-ray images. The model achieved an AP of 0.88 at an IoU threshold of 0.5.Another approach utilized a Faster R-CNN model with an anchor-based approach, combined with a multi-resolution Feature Pyramid Network (FPN) and a ResNet50 backbone. This model was tested on 2,333 X-ray images of various femoral fractures, achieving a mAP score of 0.69.

DeepWrist Pipeline: A deep learning-based pipeline called DeepWrist was developed for detecting distal radius fractures. The model was trained on a dataset of 1,946 wrist studies and evaluated on two test sets.

First test set (207 cases): Achieved AP score of 0.99.Second test set (105 challenging cases): Achieved AP of 0.64.The model generated heatmaps to indicate the probability of a fracture near the wrist but did not provide bounding boxes or polygons to clearly locate the fracture. The study was limited due to the small dataset and the disproportionate number of challenging cases. CrackNet + Faster R-CNN Approach: In this study, Radiopaedia dataset images were first classified into fracture and non-fracture categories using CrackNet. Then, Faster R-CNN was used for fracture detection on 1,052 bone images. The approach achieved:Accuracy: 0.88,Recall: 0.88,Precision: 0.89.

Feature Ambiguity Mitigate Operator (FAMO) Model: A FAMO model, along with ResNeXt101 and Feature Pyramid Network (FPN), was used to identify fractures in a dataset of 9,040 radiographs covering various body parts (hand, wrist, pelvic, knee, ankle, foot, and shoulder). The study achieved an AP score of 0.77.Guided Anchoring (GA) for Fracture Detection: This study used Faster R-CNN with a guided anchoring (GA) method for fracture detection in hand X-ray images. The model predicted fractures using proposal regions refined by learnable and flexible anchors from the GA module. The approach was evaluated on 3,067 images, achieving an AP score of 0.71.WFD-C Ensemble Model: A dataset of wrist X-ray images from Gazi University Hospital was used to conduct 20 fracture detection experiments.

An ensemble model (WFD-C) was developed, combining five different models.A total of 26 models were evaluated, with WFD-C achieving the highest AP of 0.86.The study utilized both two-stage and single-stage detection methods:Two-stage models: Dynamic R-CNN, Faster R-CNN, SABL, and DCN models based on Faster R-CNN.Single-stage models: PAA, FSAF, RetinaNet, RegNet, SABL, and Libra.

Transfer Learning with Modified Mask R-CNN: This study used transfer learning with a modified Mask R-CNN to detect and segment fractures using two datasets:Surface crack image dataset (3,000 images)Wrist fracture dataset (315 images)The model was first trained on the surface crack dataset, then fine-tuned on the wrist fracture dataset.Achieved an average precision of 92.3% for detection and 0.78 for segmentation (0.5 IoU scale),0.79 for detection and 0.52 for segmentation (strict 0.75 IoU scale)

2.2. One-stage detection

Very few studies have been conducted demonstrating the performance of one-stage detectors in the area of wrist trauma and fracture detection. In the study by Sha, Wu and Yu (2020a), a YOLOv2 model was used to detect fractures in a dataset of 5134 spinal CT images, resulting in a mAP of
0.75. In another research by the same authors Sha, Yu and Wu (2020b), a Faster R-CNN model was applied to the same dataset, yielding an mAP of 0.73.

A recent study by Hrži'c et al. (2022) compared the per- formance of the YOLOv4 object detection model Bochkovskiy, Wang and Liao (2020) to that of the U-Net segmentation model proposed by Lindsey, Daluiski, Chopra, Lachapelle, Mozer, Sicular, Hanel, Gardner, Gupta, Hotchkiss et al. (2018) and a group of radiologists on the "GRAZPEDWRI- DX" dataset. The authors trained two YOLOv4 models for this study: one for identifying the most probable fractured object in an image and the other for counting the num- ber of fractures present in an image. The first YOLOv4 model achieved high performance, with an AUC-ROC of
0.90 and an F1-score of 0.90, while the second YOLOv4 model achieved an AUC-ROC of 0.90 and an F1-score of
0.96. These results demonstrate the superior performance of YOLOv4 in comparison to traditional methods for fracture detection.

The "GRAZPEDWRI-DX" dataset used in this study was recently published Nagy et al. (2022). The authors presented the baseline results for the dataset using the COCO pre- trained YOLOv5m variant of YOLOv5. The model was trained on 15,327 (of 20,327) images and tested on 1,000 images. They achieved a mAP of 0.93 for fracture detection and an overall mAP of 0.62 at an IoU threshold of 0.5.

In conclusion, the literature review shows that the ma- jority of studies on fracture detection have utilized the two- stage detection approach. Additionally, the datasets utilized in these studies tend to be limited in size in comparison to the dataset used in our study. This study builds upon the work of studies Hrži'c et al. (2022) and Nagy et al. (2022) by conducting a comprehensive comparative study between the state-of-the-art single-stage detection algorithms (YOLOv5, YOLOv6, YOLOv7, and YOLOv8) and a widely used two- stage model Faster R-CNN. The results of this study provide valuable insights into the performance of these algorithms and contribute to the ongoing research in the field of wrist trauma and fracture detection.

3. Material & Methods

3.1. Research Design

A quantitative (experimental) study is conducted using data from 10,643 wrist radiography studies of 6,091 unique patients collected by the Division of Paediatric Radiology, Department of Radiology, Medical University of Graz, Aus- tria. As shown in Fig. 1, the dataset was randomly partitioned into a training set of 15,245, a validation set of 4,066, and a testing set of 1016. In the following subsection, we describe various measurements used to assess the performance of the models.

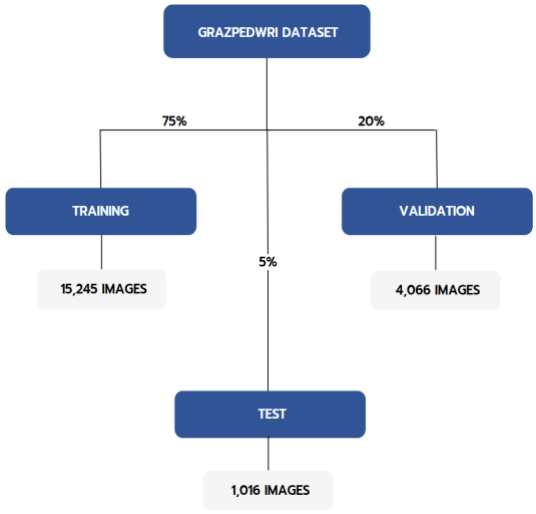


Figure 1: Dataset split into training, validation, and test sets.

3.2. Tools & Instruments

Python scripts were used to partition the dataset into training, validation, and testing sets. The deep learning framework PyTorch was used to train object detection models. To visualize, track, and compare model training, we employed the Weights and Biases (WANDB) platform. To take advantage of our system's graphical processing units (GPUs), we utilized CUDA and cuDNN. All training was performed on a Windows PC equipped with an NVIDIA GeForce RTX 2080 SUPER (with 8,192 MB of video memory), an Intel(R) Xeon(R) W-2223 CPU @ 3.60GHz processor, and 64GB of RAM. The Python version used was 3.9.13.

3.3. Deep Learning Models For Object Detection

In this study, we employed 4 single-stage detection mod- els, namely YOLOv5, YOLOv6, YOLOv7, and YOLOv8, as well as a two-stage detection model Faster R-CNN. To further optimize the performance of the single-stage models, we experimented with multiple variants of each YOLO model, ranging from 5 to 7 variants. This resulted in a total of 23 wrist abnormality detection procedures.

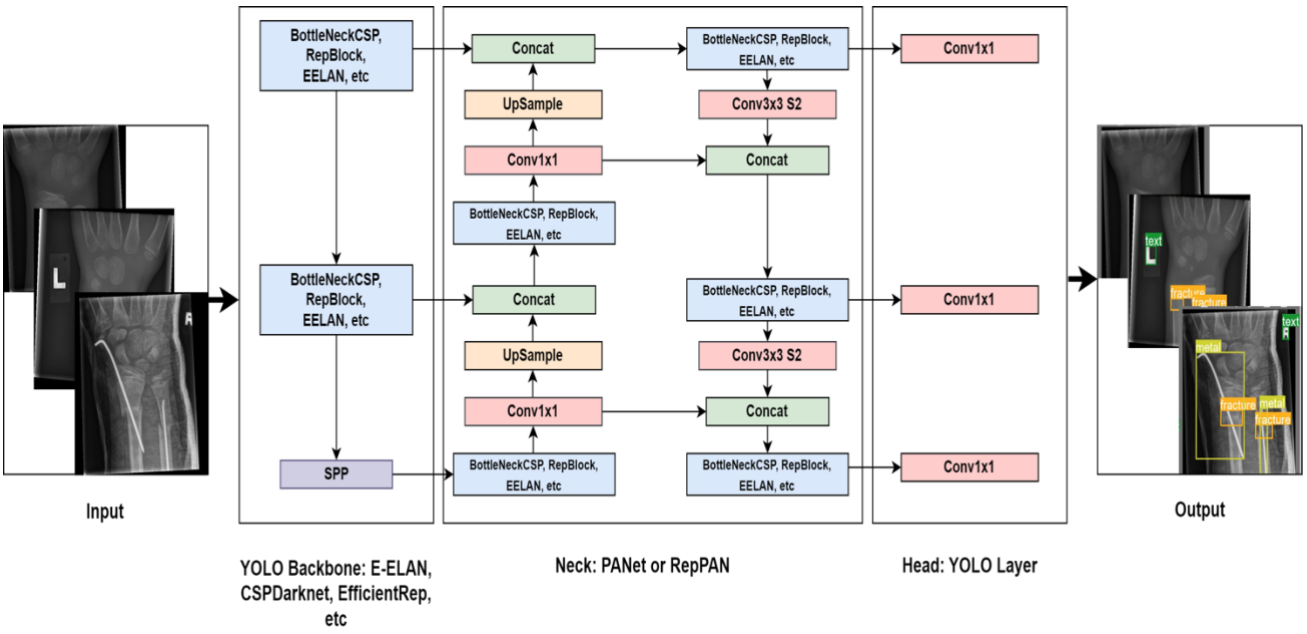


Figure 2: YOLO Architecture depicting the input, backbone, neck, head, and the output.

The YOLO (You Only Look Once) algorithm, initially introduced by Redmon, Divvala, Girshick and Farhadi (2015) in 2015, is a single-stage object detection approach that uses a single pass of a convolutional neural network to make predictions about the locations of objects in an image, making it faster than other approaches to date. In 2021, YOLOv4 achieved the highest mean average precision on the MS COCO dataset while also being the fastest real- time object detection algorithm Bochkovskiy et al. (2020). Since its initial release, the algorithm has undergone several improvements, with versions ranging from v1 to v8, with each subsequent version offering smaller volume, higher speed, and higher precision. Fig. 2 illustrates the general structure of YOLO with various backbones used in this study such as CSP, VGG, and EELAN.

3.3.1. The YOLOv5 Model

The YOLO framework comprises of three components: the backbone, neck, and head. The backbone extracts image features using the CSPDarknet architecture, known for its superior performance Wang, Liao, Wu, Chen, Hsieh and Yeh (2020). We adopted the same architecture in our research. CSPDarknet involves convolution, pooling, and residual connections represented as:

$$F_i = f(F_{i-1}, W_i) + F_{i-1} \quad (1)$$

(Where F_i and F_{i-1} are feature maps at i-th and $(i-1)$- th layer respectively, W_i represents weights and biases, and $f(\cdot)$ applies convolution and pooling operations). The SPP structure is then used to extract multi-scale features from the CSPDarknet's output:

$$F_{SPP} = g(F_i) \quad (2)$$

(Where F_{SPP} denotes multi-scale feature maps, and $g(\cdot)$ per- forms the SPP operation on F_i). The neck component adopts the Path Aggregation Network (PANet) to aggregate backbone features, generating higher-level features for output layers. The head constructs output vectors containing class probabilities, objectness scores, and bounding box coordinates. YOLOv5 encompasses five model variants ("n", "s", "m", "l", and "x"), which are compound-scaled versions of the same architecture. These variants offer varying detection accuracy and performance, achieved by adjusting network depth and layer count.

3.3.2. The YOLOv6 Model

YOLOv6 features an anchor-free design and reparameterized Backbone, with VGG and CSP Backbones used in the "n" and "s" variants, and "m", "l" and "l6" variants respectively. This Backbone is referred to as EfficientRep. The Neck, named Rep-PAN, is similar to YOLOv5, but the Head is efficiently decoupled, improving accuracy and reducing computation by not sharing parameters between the classification and detection branches. The YOLOv6 includes five model variants ("n", "s", "m", "l", and "l6").

3.3.3. The YOLOv7 Model

YOLOv7 comes with several changes, including E-ELAN, which uses expand, shuffle, and merge cardinality to improve network learning without disrupting the gradient path. Other changes include Model Scaling techniques, Re-parameterization planning, and Auxiliary Head Coarse-to-Fine. Model scaling adjusts the width, depth, and resolution of a model to align with specific application requirements. YOLOv7 uses compound scaling to simultaneously scale network depth and width by concatenating layers, maintaining optimal architecture while scaling.

Re-parameterization techniques use gradient flow

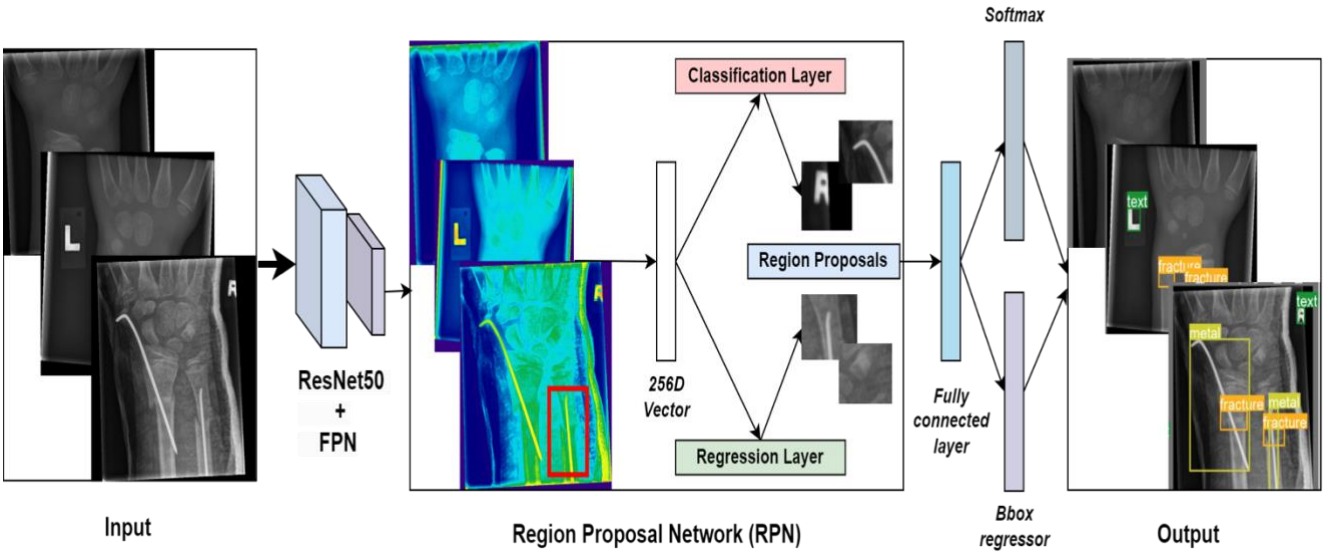


Figure 3: Faster R-CNN Pipeline.

propagation to identify modules that require averaging weights for robustness. An auxiliary head in the middle of the network improves training but requires a coarse-to-fine approach for efficient supervision. The YOLOv7 model consists of seven variants: "P5" models (v7, v7x, and v7-tiny) and "P6" models (d6, e6, w6, and e6e).

3.3.4. Faster R-CNN

The Faster R-CNN model includes a backbone, an RPN (regional proposal network), and a detection network. ResNet50 with FPN is used as the backbone for feature extraction. Anchors with variable sizes and aspect ratios are generated for each feature. The RPN selects appropriate anchor boxes using a classifier that predicts if an anchor box contains an object based on an IoU threshold of 0.5. The regressor predicts offsets for anchor boxes containing objects to fit them tightly to the ground truth labels. Finally, the RoI pooling layer converts variable-sized proposals to a fixed size to run a classifier and regress a bounding box. Fig. 3 illustrates the architecture of Faster R-CNN.

3.4. Training Details

In the experimentation of YOLO variants, standard hyperparameters were utilized. The input resolution was fixed at 640 pixels. The optimization algorithm employed was SGD with an initial learning rate $\alpha = 1\times10^{-2}$, final learning rate $\alpha_f = 1 \times 10^{-2}$ (except for YOLOv7 variants with a final learning rate $\alpha_f = 1 \times 10^{-1}$), momentum = 0.937, weight decay = 5×10^{-4}. Each variant/model underwent 100 epochs of training from scratch and was observed to converge between 90-100 epochs. Every variant was trained with a batch size of 16 except for the "P6" variants of YOLOv7 namely (d6, e6, w6, e6e) which were trained with a batch size of 8 due to computational constraints.

With Faster R-CNN, the only difference was the learning rate of $\alpha = 1 \times 10^{-3}$, momentum of 0.9 and weight decay of 5×10^{-4}. All other parameters were the same as YOLO variants. As with YOLO models, the selection of these parameters is not deliberate, they are the default settings.

All binary classifiers were trained for a maximum of 100 epochs using a batch size of 64. The learning rate was set at 1×10^{-3}. The Adam optimization algorithm guided the training process. Input images were standardized to a resolution of 224 pixels.

3.5. Evaluation Metrics: mAP

For the evaluation of object detection, a common way to determine if the predicted location of an object was correct is to find in *Intersection over Union (IoU)*. It is defined as the ratio of the intersection of the predicted and the ground truth bounding box over the union of the predicted and ground truth bounding box. A visual illustration of *IoU* is presented in Fig. 4. Given the set of predicted bounding boxes *A* for a given image, and the set of ground truth bounding boxes *B* for the same image. The IoU can be computed as:

$$IoU(A, B) = \frac{A \cap B}{A \cup B}; \qquad \text{where } A, B \in [0, 1] \tag{3}$$

Commonly, if the $IoU > 0.5$, we classify the detection as true positive, otherwise, it is classified as false positive. Given *IoU*, we can compute the number of true positives *TP* and false positives *FP* and compute the Average precision *AP* for each object class *c* as follows:

$$AP(c) = \frac{TP(c)}{TP(c) + FP(c)} \tag{4}$$

Finally, after computing *AP* for each object class, we compute the Mean Average Precision *mAP* which is an average of *AP* across all classes *C* under consideration. *mAP* is given

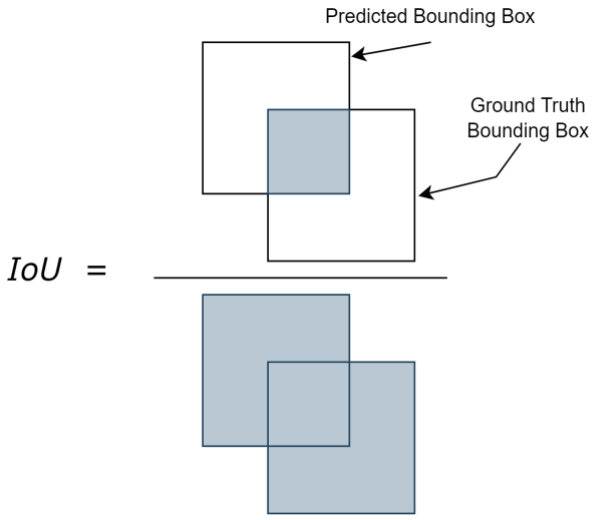


Figure 4: Visual illustration of Intersection over Union (*IoU*).

mAP is given as:

$$mAP = \frac{1}{C}\sum_{c=1} AP(c) \quad (5)$$

mAP is the metric that quantifies the performance of object detection algorithms. Thus, the metric $mAP_{0.5}$ indicates *mAP* for $IoU > 0.5$. This is the *IoU* threshold we will be using to make our assessments of the detection models.

3.5.1. Sensitivity

Sensitivity, in the context of our model, pertains to its capacity to accurately recognize true detections among all positive detections within the dataset. Specifically, it gauges the model's ability to correctly identify the presence of a fracture or abnormality. We prioritize this metric due to the potential consequences of false negatives in wrist trauma cases. Failure to detect fractures is a frequent reason for differences in diagnosis between the initial interpretation of X-ray images and the final analysis conducted by certified radiologists. The calculation for sensitivity is as follows:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \quad (6)$$

3. Dataset

The dataset used in this study is called GRAZPEDWRI-DX *for machine learning* presented by the authors in Nagy et al. (2022) and is publicly made available to encourage computer vision research. The dataset contains pediatric wrist radiograph images in PNG format of 6,091 patients (mean age 10.9 years, range 0.2 to 19 years; 2,688 females, 3,402 males, 1 unknown), treated at the Division of Paediatric Radiology, Department of Radiology, Medical University of Graz, Austria. The dataset includes a total of 20,327 wrist images covering lateral and posteroanterior projections. The radiographs were acquired over the span of 10 years between 2008 and 2018 and have been comprehensively annotated between 2018 and 2020 by expert radiologists and various medical students. The annotations were validated by three experienced radiologists as the X-ray images were annotated. This process was repeated until a consensus was met between the annotations and interpretations from three radiologists. We choose to use this dataset in our study for the following reasons:

1. The dataset is quite large consisting of 20,327 labeled and tagged images, making it suitable for various computer vision algorithms
2. To our knowledge, there are no related pediatric datasets publicly available, with others featuring only binary labels or not as comprehensively labeled as the one we use.
3. To the best of our knowledge, this is the first comprehensive study of the recently released GRAZPEDWRI-DX dataset using state-of-the-art computer vision models YOLOv5, v6, v7 and v8.
4. It contains diverse images of the early stages of bone growth and organ formation in children. Studying the wrist at this stage offers unique insights into the diagnosis, treatment, and prevention of anomalies that are not possible when studying adult wrists.

4.1. Analysis of Objects in the Dataset

The dataset includes a total of 9 objects: periosteal reaction, fracture, metal, pronator sign, soft tissue, bone anomaly, bone lesion, foreign body, and text. The object "text" is present in all X-ray images and is used to identify the side of the body (right or left hand) on which the X-ray was taken. The number of objects in the dataset is shown in Table 1. The table clearly indicates that the object "fracture" has the most common occurrence in wrist X-rays of GRAZPEDWRI-Dataset. The class "periosteal reaction" has the second largest occurrence followed by the third largest class "metal". Meanwhile, the classes "bone anomaly", "bone lesion", and "foreign body" have the lowest occurrence. Note that this table shows how many X-ray images contain a particular object and not the number of times an object is labeled in the dataset. Additionally, a histogram is shown in Fig. 5 visually shows the class distribution.

Table 1
Class Distribution

Abnormality	Instances	Ratio
Boneanomaly	192	0.94%
Bonelesion	42	0.21%
Foreignbody	8	0.04%
Fracture	13550	66.6%
Metal	708	3.48%
Periostealreaction	2235	11.0%
Pronatorsign	566	2.78%
Softtissue	439	2.16%

In Table 2, we show the number of images in which a particular anomaly occurs only once, twice, or multiple

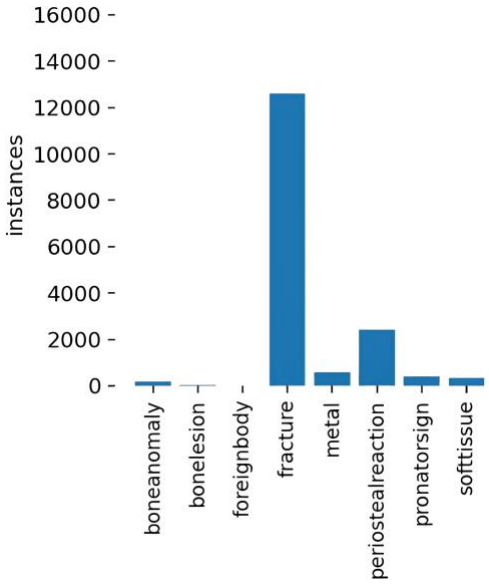


Figure 5: Histogram of Class Distribution.

times. The column "Total" represents the total number of images in which a particular anomaly is present.

Table 2
Object Occurrences

Abnormality	Zero	One	Two	More	Total
Fracture	6777	9212	4137	201	13550
Boneanomaly	20135	42	24	126	192
Bonelesion	20285	11	8	23	42
Foreignbody	20319	0	0	8	8
Metal	19620	347	219	141	707
Periostealreaction	18092	1273	885	77	2235
Pronatorsign	19761	456	71	39	566
Softtissue	19888	221	82	136	439

4. Results & Discussion

This section presents a comprehensive analysis of the performance of various models for wrist abnormality detection on the GRAZPEDWRI-DX dataset. A total of 23 detection procedures were conducted using different variants of each YOLO model and a two-stage detection model (Faster R-CNN) on a test set consisting of 1016 randomly selected samples. The performance of each model was evaluated using metrics such as precision, recall, and mean average precision (mAP). We begin by providing a detailed analysis of the variants within each YOLO model. Next, we select the best-performing variant from each YOLO model based on the highest mAP score obtained for the fracture class, as well as across all classes. Finally, we compare these variants to determine the overall best-performing model and evaluate its performance against Faster R-CNN.

The results of YOLOv5 variants are presented in Table 3 and 4, showing the performance of the variants across all classes and on the fracture class, respectively. All values are

Table 3
YOLOv5 Results (Across All Classes)

Model variant	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5n	0.77	0.52	0.59	0.34
YOLOv5s	0.75	0.66	0.65	0.38
YOLOv5m	0.80	0.62	0.69	0.44
YOLOv5l	0.76	0.61	0.68	0.43
YOLOv5x	0.73	0.64	0.69	0.45

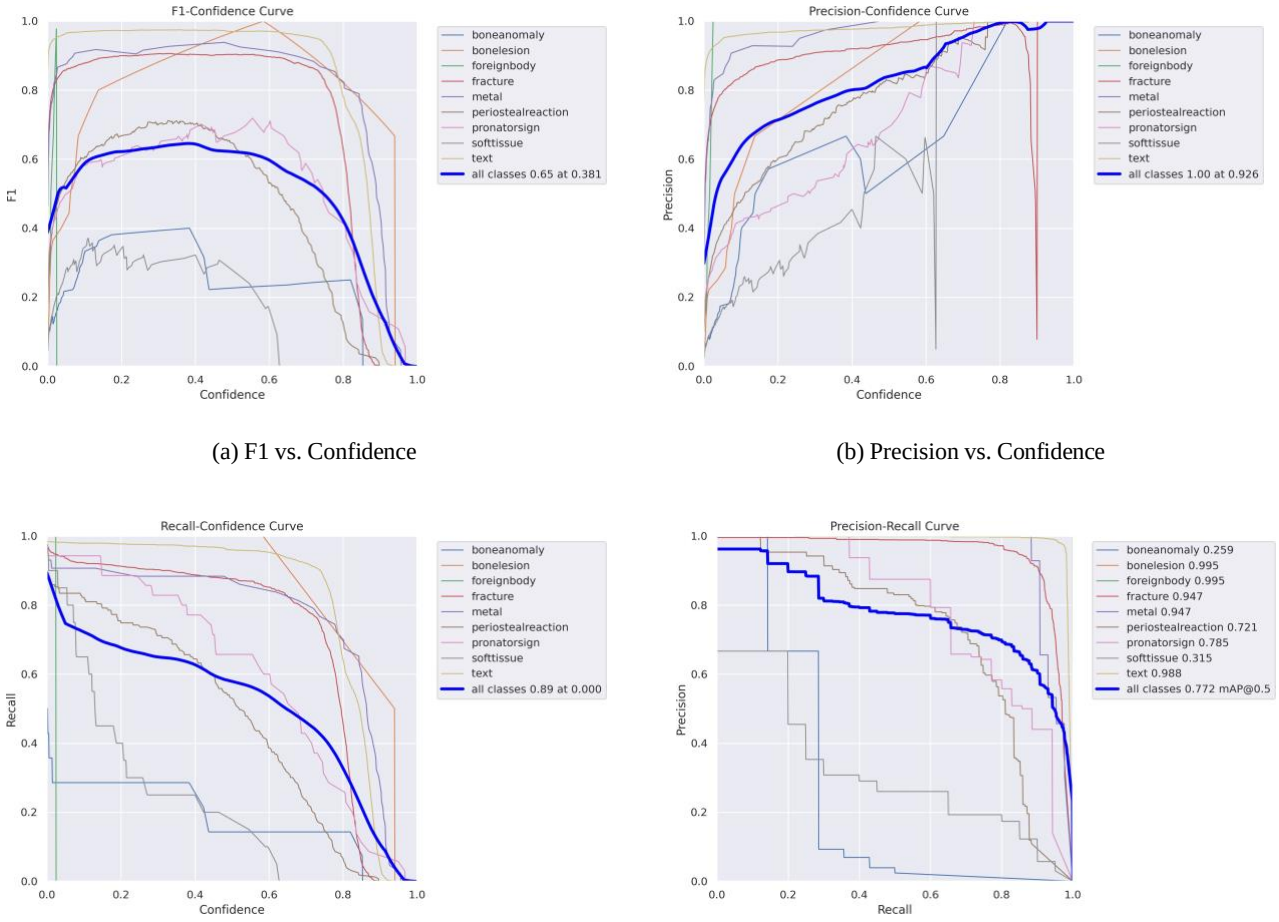
Table 4
YOLOv5 Results (Fracture Class)

Model variant	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5n	0.87	0.91	0.94	0.54
YOLOv5s	0.89	0.91	0.95	0.56
YOLOv5m	0.91	0.90	0.94	0.56
YOLOv5l	0.92	0.90	0.95	0.57
YOLOv5x	0.91	0.90	0.95	0.57

rounded to two decimal places. The results show that the fractures were detected with the highest mAP of 0.95 at IoU = 0.5, with a precision of 0.92, and a recall of 0.90 by the YOLOv5 variant, YOLOv5l. Additionally, the performance of YOLOv5l appears to be satisfactory across all classes with the mAP score of 0.68 at IoU = 0.5. The variant YOLOv5x seems to perform just as well in terms of mAP obtained for the fracture class. In terms of overall performance across all classes, the highest mAP score achieved was 0.69 by the two YOLOv5 variants "m" and "x". The highest precision obtained across all classes is 0.80 by the variant "m", while the highest recall achieved was 0.66 by the variant "s". It can also be observed from the results shown in Table 4 that as the complexity of the architecture in YOLOv5 increases, its performance improves.

Table 5 displays the mAP scores of all YOLOv5 variants at an IoU threshold of 0.5 for all classes present in the GRAZPEDWRI-DX dataset. It is worth noting that these mAP scores are particularly significant as they are calculated at an IoU threshold of 0.5, which is a commonly used threshold in object detection evaluations. These scores are crucial indicators of the performance of the YOLOv5 variants on the GRAZPEDWRI-DX dataset and provide valuable insights into their abilities to detect objects within the various classes present in the GRAZPEDWRI-DX dataset. Upon examination of the Table 5, it can be seen that almost all variants of YOLOv5 demonstrate the capability to detect classes that are in the minority, such as bone anomaly, bone lesion, and foreign body, with considerably good mAP scores as seen in Table 5. For instance, despite the limited number of instances of the class "Bonelesion" (only 42, as shown in Table 1), the four variants of YOLOv5 ("s", "m," "l," and "x") are able to correctly detect it in all instances where it occurs, with the mAP score of 1.00.

Table 6 and 7 present the results of YOLOv6 variants, showcasing their performance on all classes and the fracture class, respectively. Variants "n", "s", and "m" achieved the highest mAP of 0.94 at an IoU threshold of 0.5 for detecting



(a) F1 vs. Confidence

(b) Precision vs. Confidence

(c) Recall vs. Confidence

(d) Precision vs. Recall

Figure 6: Performance analysis curves (YOLOv8x)

Table 5
YOLOv5 mAP@0.5 Scores (For All Classes)

Model variant	Boneanomality	Bonelesion	Foreignbody	Fracture	Metal	Periostealreaction	Pronatorsign	Softtissue	Text
YOLOv5n	0.31	0.57	0.00	0.94	0.88	0.66	0.74	0.21	0.99
YOLOv5s	0.31	1.00	0.00	0.95	0.91	0.75	0.74	0.25	0.99
YOLOv5m	0.33	1.00	0.33	0.94	0.92	0.69	0.75	0.25	0.99
YOLOv5l	0.34	1.00	0.25	0.95	0.90	0.71	0.75	0.19	0.99
YOLOv5x	0.37	1.00	0.33	0.95	0.92	0.71	0.77	0.19	0.99

fractures. Variants "n", "m", and "l" displayed the highest precision for the fracture class with a value of 0.94, while variant "s" had the highest recall of 0.89. In terms of overall performance across all classes, the highest mAP score of 0.64 at an IoU threshold of 0.5 was obtained by variants "m" and "l", with variant "l" achieving the highest precision of 0.60 and variant "m" having the highest recall of 0.83.

Table 8 illustrates that YOLOv6 variants, similar to YOLOv5 variants, exhibit the ability to detect minority classes. However, Table 7 reveals that, unlike YOLOv5, as the complexity of the model increases from variant "m" to "l" and then to "l6", the mAP score decreases, indicating that complexity beyond variant "m" results in decreased performance. This trend is also observed in Table 6, where increasing complexity from variant "l" to "l6" results in decreased performance across all classes.

The performance of YOLOv7 variants on both across classes and the fracture class is presented in Tables 9 and 10, respectively. The results indicate that the second variant of the YOLOv7 model exhibits the highest mean average precision (mAP) of 0.94 at an intersection over union (IoU) threshold of 0.5, with a precision of 0.86 and recall of 0.91 for detecting fractures. This variant also demonstrates superior performance across all classes with a mAP of 0.61

Table 6
YOLOv6 Results (Across All Classes)

Model variant	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv6n	0.50	0.73	0.51	0.31
YOLOv6s	0.51	0.82	0.62	0.37
YOLOv6m	0.59	0.83	0.64	0.36
YOLOv6l	0.60	0.80	0.64	0.41
YOLOv6l6	0.49	0.77	0.52	0.31

Table 7
YOLOv6 Results (Fracture Class)

Model variant	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv6n	0.94	0.86	0.94	0.55
YOLOv6s	0.92	0.89	0.94	0.54
YOLOv6m	0.94	0.87	0.94	0.55
YOLOv6l	0.94	0.87	0.93	0.53
YOLOv6l6	0.91	0.86	0.92	0.53

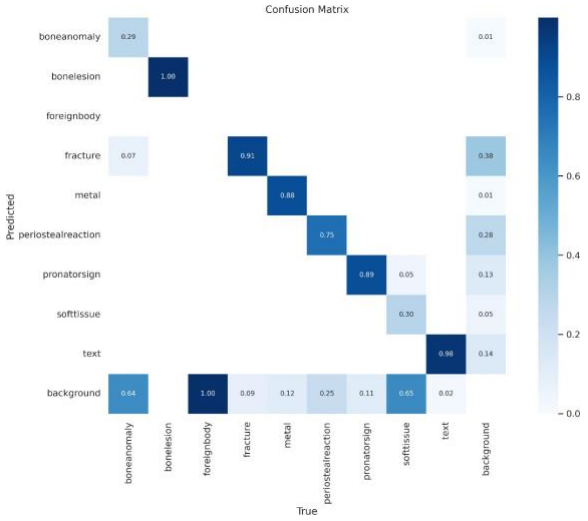


Figure 7: Confusion Matrix (YOLOv8x).

at an IoU of 0.5, a precision of 0.79, and a recall of 0.54. The variant YOLOv7x seems to perform just as well in terms of mAP obtained for the fracture class but has a lower mAP score compared to the second variant across all classes. Additionally, it can be observed from our experiments that, in contrast to YOLO5, increasing the complexity of the YOLOv7 architecture, in terms of depth and number of layers, hurts its performance in detecting wrist abnormalities. The only exception to this trend is the increase in performance observed when comparing the smaller variant "YOLOv7-Tiny" to the slightly larger variant "YOLOv7". The "YOLOv7-Tiny" achieved mAP of 0.5 at IoU=0.5, but the "YOLOv7" variant showed an improvement of 0.11 across all classes. Additionally, when focusing on the specific class of fractures, an improvement of 0.01 in the mAP score was observed, suggesting that there is an optimal balance of complexity and performance for this model. The decline in performance for YOLOv7's "P6" models, specifically "W6", "E6", "D6", and "E6E", compared to the "P5" models may be attributed to the reduced image resolution. However, the results across all classes indicate that even with this resolution, the performance of "P6" models either decreases or does not improve at all.

It is worth noting that rare classes such as bone anomaly, bone lesion, and foreign body have a very low mAP score and are sometimes not detected at all, as shown in Table 11. However, the second variant of YOLOv7 is the only variant able to detect all the minority classes such as "bone anomaly", "bone lesion", and "foreign body".

Tables 12 and 13 show the performance of YOLOv8 model variants across all classes and on the fracture class, respectively. The YOLOv8 variant "YOLOv8x" achieved the highest mAP of 0.95 for fracture detection at an IoU threshold of 0.5, with a precision of 0.91 and a recall of 0.89. Additionally, it demonstrated superior overall performance across all classes with a mAP of 0.77 at an IoU threshold of 0.5. Table 14 also shows that all YOLOv8 variants demonstrated good performance in detecting all classes, including minority classes, except the "foreign body" class not being detected by the small and the medium variants. The results suggest that using compound-scaled variants of the YOLOv8 architecture generally improves performance, except for a decrease in mAP scores across all classes when moving from the variant "s" to a medium variant "m", with a decrease of 0.09 in Table 13.

The results of the experimental evaluation using the two-stage detector Faster R-CNN are presented in Table 15. The table shows the mean Average Precision (mAP) scores obtained for each class individually as well as the overall mAP across all classes. The results indicate that all variants of the YOLO model outperform Faster R-CNN by a significant margin. This is supported by the fact that the mean mAP score of every YOLO variant was found to be higher than that of Faster R-CNN, both for fracture detection and overall performance across all classes. These findings suggest that the single-stage detection algorithm, YOLO, is a more effective model for this task. Moreover, Faster R-CNN does not seem to exhibit the ability to detect the classes in minority such as "bone anomaly", "bone lesion", and "foreign body".

Figures 8 and 9 provide an overview of the mAP scores obtained for fracture class as well as across all classes by all YOLO variants and Faster R-CNN. In applications where false positives are costly, a model with high precision may be preferable, while in situations where missing detections are costly, a model with high recall may be more desirable. The mean Average Precision (mAP) serves as a comprehensive measure of the model's performance. Therefore, we selected the best-performing variant within each YOLO model based on the highest mAP achieved for the fracture class and overall performance across all classes. We have also compared their mAP scores to each other as well as with that of the Faster R-CNN model, as illustrated in Table 16. We also evaluated the performance of all variants, including Faster R-CNN, on a challenging image containing multiple objects

Table 8
YOLOv6 mAP@0.5 Scores (For All Classes)

Model variant	Boneanomality	Bonelesion	Foreignbody	Fracture	Metal	Periostealreaction	Pronatorsign	Softtissue	Text
YOLOv6n	0.10	0.01	0.00	0.94	0.84	0.73	0.76	0.29	0.98
YOLOv6s	0.15	0.54	0.33	0.94	0.91	0.72	0.76	0.21	0.98
YOLOv6m	0.10	0.10	1.00	0.94	0.87	0.75	0.76	0.31	0.98
YOLOv6l	0.10	0.10	1.00	0.93	0.93	0.71	0.77	0.25	0.98
YOLOv6l6	0.13	0.10	0.00	0.92	0.90	0.67	0.76	0.22	0.98

Table 9
YOLOv7 Results (Across All Classes)

Model variant	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv7-Tiny	0.59	0.52	0.50	0.28
YOLOv7	0.79	0.54	0.61	0.39
YOLOv7x	0.68	0.49	0.53	0.32
YOLOv7-W6	0.53	0.43	0.44	0.24
YOLOv7-E6	0.56	0.38	0.40	0.21
YOLOv7-D6	0.50	0.45	0.42	0.24
YOLOv7-E6E	0.75	0.45	0.44	0.25

Table 10
YOLOv7 Results (Fracture Class)

Model variant	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv7-Tiny	0.79	0.91	0.93	0.53
YOLOv7	0.86	0.91	0.94	0.55
YOLOv7x	0.85	0.90	0.94	0.54
YOLOv7-W6	0.72	0.89	0.90	0.50
YOLOv7-E6	0.55	0.84	0.83	0.44
YOLOv7-D6	0.62	0.89	0.89	0.49
YOLOv7-E6E	0.74	0.89	0.91	0.52

of interest, including 2 fractures, 3 periosteal reactions, 1 metal, and 1 text. The bounding box estimates for these objects from each variant and Faster R-CNN are illustrated in Fig. 10.

It is clear from Table 16 that the variant "YOLOv8x" of YOLOv8 is the best-performing variant out of all the variants employed in this study. The results presented in this study using the variant "YOLOv8x" represent a significant improvement upon the ones originally presented in Nagy et al. (2022) for the fracture class. In that paper, the model variant "YOLOv5m" trained on COCO weights achieved a mean average precision (mAP) score of 0.93 for fracture detection and an overall mAP score of 0.62 at an IoU threshold of 0.5. In contrast, the results obtained in this study demonstrate a higher mAP score of 0.95 for fracture detection and an overall mAP of 0.77 at an IoU threshold of 0.5. Fig. 6a, 6b, 6c, and 6d present the F1 versus Confidence, Recall versus Confidence, Precision versus Confidence, and Precision versus Recall curves, respectively, for the variant "YOLOv8x" across all classes. These curves provide a visual representation of the model's performance on different confidence intervals and allow for a more thorough evaluation of its capabilities. The F1 versus Confidence curve shows the relationship between the model's F1 score, which is a measure of the balance between precision and recall, and the confidence of its predictions. The Recall versus Confidence curve illustrates the model's ability to correctly identify objects, while the Precision versus Confidence curve demonstrates the proportion of correct predictions made by the model. The Precision versus Recall curve shows the trade-off between the model's precision and recall, with higher precision typically corresponding to lower recall and vice versa. Additionally, a confusion matrix 7 is shown for the variant "YOLOv8x".

Our study found that the relationship between the complexity of a YOLO model and its performance is not always linear. Our results on the GRAZPEDWRI-DX dataset revealed that the performance of YOLO models did not consistently improve with increasing complexity, except for YOLOv5 and YOLOv8.

5. Conclusion & Future Work

In this study, we aimed to evaluate the performance of state-of-the-art single-stage detection models, specifically YOLOv5, YOLOv6, YOLOv7, and YOLOv8, in detecting wrist abnormalities and compare their performances against each other and the widely used two-stage detection model Faster R-CNN. Additionally, the analysis of the performance of all variants within each YOLO model was also provided. The evaluation was conducted using the recently released GRAZPEDWRI-DX Nagy et al. (2022) dataset, with a total of 23 detection procedures being carried out. The findings of our study demonstrated that YOLO models outperform the commonly used two-stage detection model, Faster R-CNN, in both fracture detection and across all classes present in the GRAZPEDWRI-DX dataset.

Furthermore, an analysis of YOLO models revealed that the YOLOv8 variant "YOLOv8x" achieved the highest mAP across all classes of wrist abnormalities in the GRAZPEDWRI-DX dataset, including the fracture class, at an IoU threshold of 0.5. We also discovered that the relationship between the complexity of a YOLO model, as measured by the use of compound-scaled variants within each YOLO model, and its performance is not always linear. Specifically, our analysis of the GRAZPEDWRI-DX dataset revealed that the performance of YOLO variants did not consistently improve with increasing complexity, except for YOLOv5 and YOLOv8. Some variants were successful in detecting minority classes while others were not. These results contribute to understanding the relationship between the

Table 11
YOLOv7 mAP@0.5 Scores (For All Classes)

Model variant	Boneanomality	Bonelesion	Foreignbody	Fracture	Metal	Periostealreaction	Pronatorsign	Softtissue	Text
YOLOv7-Tiny	0.13	0.00	0.00	0.93	0.88	0.69	0.69	0.14	0.99
YOLOv7	0.20	0.33	0.33	0.94	0.95	0.76	0.71	0.25	0.99
YOLOv7x	0.17	0.10	0.00	0.94	0.90	0.72	0.70	0.24	0.99
YOLOv7-W6	0.00	0.00	0.00	0.90	0.88	0.57	0.46	0.14	0.98
YOLOv7-E6	0.00	0.00	0.00	0.83	0.81	0.42	0.40	0.11	0.98
YOLOv7-D6	0.00	0.00	0.00	0.89	0.87	0.53	0.34	0.14	0.99
YOLOv7-E6E	0.01	0.00	0.00	0.91	0.88	0.60	0.43	0.12	0.99

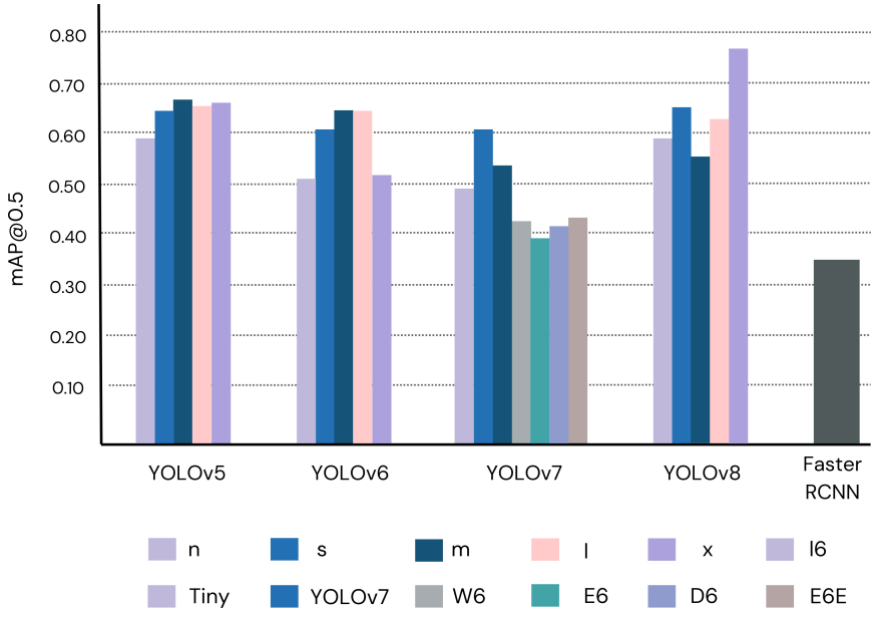


Figure 8: mAP Scores (Across All Classes).

Table 12
YOLOv8 Results (Across All Classes)

Model variant	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8n	0.73	0.58	0.59	0.36
YOLOv8s	0.72	0.63	0.65	0.39
YOLOv8m	0.60	0.60	0.56	0.36
YOLOv8l	0.74	0.60	0.62	0.41
YOLOv8x	0.79	0.64	0.77	0.53

Table 13
YOLOv8 Results (Fracture Class)

Model variant	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8n	0.87	0.88	0.93	0.55
YOLO8s	0.87	0.91	0.94	0.56
YOLOv8m	0.84	0.92	0.95	0.57
YOLOv8l	0.92	0.90	0.95	0.57
YOLO8x	0.91	0.89	0.95	0.57

complexity of YOLO models and their performance, which is important for guiding the development of future models. Our study highlights the potential of single-stage detection algorithms, specifically YOLOv5, YOLOv6, YOLOv7, and YOLOv8, for detecting wrist abnormalities in clinical settings. These algorithms are faster than their two-stage counterparts, making them more practical for emergencies commonly found in hospitals and clinics. Additionally, the study's results indicate that single-stage detectors are highly accurate in detecting wrist abnormalities, making them a promising choice for clinical use.

While this research was conducted, YOLOv8 was the most recent version. The results of this study can serve as a benchmark for evaluating the performance of future models for wrist abnormality detection, as further improvements to either YOLOv8 or future versions of YOLO may surpass the results obtained in this study. It is worth noting that this study didn't explore the entire hyperparameter space

Table 14
YOLOv8 mAP@0.5 Scores (For All Classes)

Model variant	Boneanomality	Bonelesion	Foreignbody	Fracture	Metal	Periostealreaction	Pronatorsign	Softtissue	Text
YOLO8n	0.20	0.50	0.11	0.93	0.91	0.71	0.70	0.26	0.99
YOLOv8s	0.27	1.00	0.00	0.94	0.93	0.71	0.76	0.21	0.99
YOLOv8m	0.19	0.27	0.00	0.95	0.96	0.73	0.80	0.18	0.99
YOLOv8l	0.22	0.55	0.10	0.95	0.97	0.72	0.79	0.26	0.99
YOLOv8x	0.26	1.00	1.00	0.95	0.95	0.72	0.79	0.32	0.99

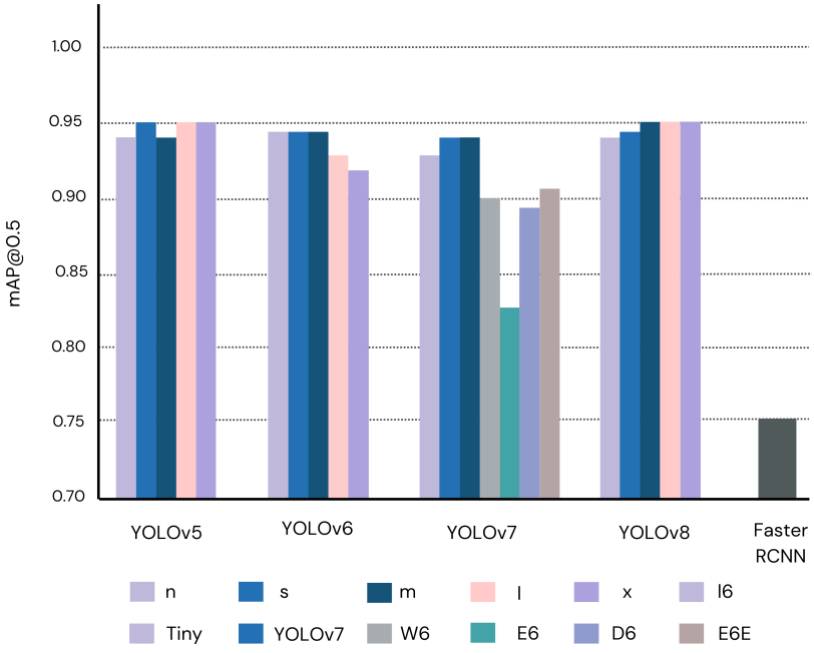


Figure 9: mAP Scores (Fracture Class).

Table 15
Faster R-CNN mAP@0.5 Scores (Across All Classes

Abnormality	mAP@0.5
Boneanomaly	0.00
Bonelesion	0.00
Foreignbody	0.00
Fracture	0.75
Metal	0.78
Periostealreaction	0.54
Pronatorsign	0.10
Softtissue	0.03
Text	0.96
All	0.35

Table 16
mAP@0.5 Scores For Best Performing Model Variants

Model	Fracture	All
YOLOv5x	0.95	0.69
YOLOv6m	0.94	0.64
YOLOv7	0.94	0.61
YOLOv8x	**0.95**	**0.77**
Faster R-CNN	0.75	0.35

and finding the best hyperparameters for each YOLO model may improve wrist abnormality detection performance on the dataset. Computational limitations restricted the input resolution to 640 pixels, but higher resolutions could further improve performance. The study showed that the models had difficulty detecting "bone anomaly", "bone lesion", and "foreign body" due to low instances of these classes, so increasing their instances through augmentation or image generation could enhance performance. Additionally, the performance of classification models could also be assessed by exploring the dataset for pure classification tasks without object localization.

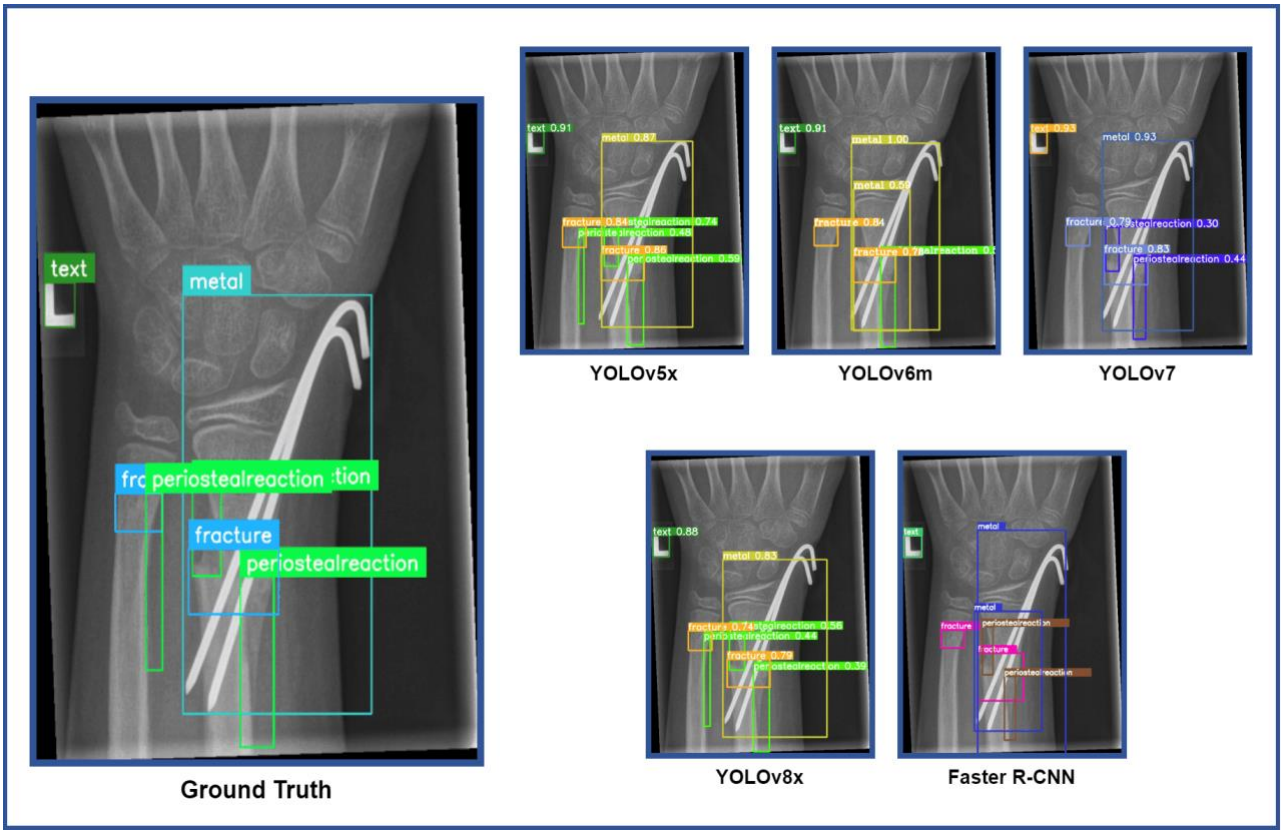


Figure 10: Bounding box estimated by YOLO variants and Faster R-CNN.

References

Adams, S.J., Henderson, R.D.E., Yi, X., Babyn, P., 2020. Artificial intelligence solutions for analysis of x-ray images. Canadian Association of Radiologists journal = Journal l'Association canadienne des radiologistes 846537120941671.

Bochkovskiy, A., Wang, C., Liao, H., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv.org URL: https://arxiv.org/abs/2004.10934.

Burki, T.K., 2018. Shortfall of consultant clinical radiologists in the uk. Lancet Oncol 19.

Cheng, J., Shen, W., 1993. Limb fracture pattern in different pediatric age groups: a study of 3350 children. J Orthop Trauma 7, 15–22. doi:10.1097/00005131-199302000-00004.

Choi, J.W., et al., 2020. Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. Investigative radiology 55, 101–110.

Er, E., Kara, P., Oyar, O., Unluer, E., 2013. Overlooked extremity fractures in the emergency department. Ulus Travma Acil Cerrahi Derg 19, 25–28.

Fotiadou, A., Patel, A., Morgan, T., Karantanas, A.H., 2011. Wrist injuries in young adults: the diagnostic impact of ct and mri. Eur J Radiol 77, 235–239. doi:10.1016/j.ejrad.2010.06.029.

Guan, B., Zhang, G., Yao, J., Wang, X., Wang, M., 2020. Arm fracture detection in x-rays based on improved deep convolutional neural network. Computer and Electrical Engineering 81, 106530.

Guly, H., 2001. Diagnostic errors in an accident and emergency department. Emerg Med J 18, 263–269.

Hallas, P., Ellingsen, T., 2006. Errors in fracture diagnoses in the emergency department: Characteristics of patients and diurnal variation. BMC Emerg Med 6.

Hardalaç, F., Uysal, F., Peker, O., Çiçeklidağ, M., Tolunay, T., Tokgöz, N., Kutbay, U., Demirciler, B., Mert, F., 2022. Fracture detection in wrist x-ray images using deep learning-based object detection models. Sensors 22, 1285. doi:10.3390/s22031285.

Hedstrom, E.M., Svensson, O., Bergstrom, U., Michno, P., 2010. Epidemiology of fractures in children and adolescents. Acta Orthopaedica 81, 148–153.

Hrži'c, F., et al., 2022. Fracture recognition in paediatric wrist radiographs: An object detection approach. Mathematics 10, 2939. doi:10.3390/math10162939.

Joshi, D., Singh, T., Joshi, A., 2022. Deep learning-based localization and segmentation of wrist fractures on x-ray radiographs. Neural Computing and Application 34, 19061–19077. doi:10.1007/s00521-022-07510-z.

Juhl, M., Moller-Madsen, B., Jensen, J., 1990. Missed injuries in an orthopaedic department. Injury 21, 110–112.

Lampert, C., Blaschko, M., Hofmann, T., 2008. Beyond sliding windows: object localization by efficient subwindow search, in: 2008 IEEE conference on computer vision and pattern recognition, IEEE. pp. 1–8.

Landin, L.A., 1997. Epidemiology of children's fractures. Journal of Pediatric Orthopaedics B 6, 79–83.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., Zhang, B., Liang, Y., Zhou, L., Xu, X., Chu, X., Wei, X., 2022. Yolov6: A single-stage object detection framework for industrial applications. arXiv arXiv:2209.02976.

Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., et al., 2018. Deep neural network improves fracture detection by clinicians. Proceedings of the National Academy of Sciences 115, 11591–11596. URL: https://www.pnas.org/content/115/47/11591, doi:10.1073/pnas.1807792115,