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

Ammar Ahmed[a], Ali Shariq Imran[b], Abdul Manaf[a], Zenun Kastrati[c] and Sher Muhammad Daudpota[a]

[a]Dept. of Computer Science, Sukkur IBA University, Sukkur, 65200, Pakistan
[b]Dept. of Computer Science, Norwegian University of Science & Technology (NTNU), Gjøvik, 2815, Norway
[c]Dept. of Informatics, Linnaeus University, Växjö, 351 95, Sweden

ARTICLE INFO

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

ABSTRACT

Diagnosing and treating abnormalities in the wrist, specifically distal radius, and ulna fractures, is a crucial concern among children, adolescents, and young adults, with a higher incidence rate during puberty. However, the scarcity of radiologists and the lack of specialized training among medical professionals pose a significant risk to patient care. This problem is further exacerbated by the rising number of imaging studies and limited access to specialist reporting in certain regions. This highlights the need for innovative solutions to improve the diagnosis and treatment of wrist abnormalities. Automated wrist fracture detection using object detection has shown potential, but current studies mainly use two-stage detection methods with limited evidence for single-stage effectiveness. This study employs state-of-the-art single-stage deep neural network-based detection models YOLOv5, YOLOv6, YOLOv7, and YOLOv8 to detect wrist abnormalities. Through extensive experimentation, we found that these YOLO models outperform the commonly used two-stage detection algorithm, Faster R-CNN, in bone fracture detection. Additionally, compound-scaled variants of each YOLO model were compared, with YOLOv8x demonstrating a fracture detection mean average precision (mAP) of 0.95 and an overall mAP of 0.77 on the GRAZPEDWRI-DX pediatric wrist dataset, highlighting the potential of single-stage models for enhancing pediatric wrist imaging.

1. Introduction

Wrist abnormalities are a common occurrence in children, adolescents, and young adults. Among them, wrist fractures such as distal radius and ulna fractures are the most common with incidence peaks during puberty Hedstrom, Svensson, Bergstrom and Michno (2010); Randsborg et al. (2013); Landin (1997); Cheng and Shen (1993). Timely evaluation and treatment of these fractures are essential to prevent life-long implications. Digital radiography is a widely used imaging modality to obtain wrist radiographs which are then interpreted by surgeons or physicians in training to diagnose wrist abnormalities. However, medical professionals may lack the specialized training to assess these injuries accurately and may rely on radiograph interpretation without the support of an expert radiologist or qualified colleagues Hallas and Ellingsen (2006). Studies have shown that diagnostic errors in reading emergency X-rays can reach up to 26% Guly (2001); Mounts, Clingenpeel, McGuire, Byers and Kireeva (2011); Er, Kara, Oyar and Unluer (2013); Juhl, Moller-Madsen and Jensen (1990). This is compounded by the shortage of radiologists even in developed countries Burki (2018); Rimmer (2017); Makary and Takacs (2022) and limited access to specialist reporting in other parts of the world Rosman (2015) posing a high risk to patient care. The shortage is expected to escalate in the upcoming years due to a growing disparity between the increasing demand for imaging studies and the limited supply of radiology residency positions. The number of imaging studies rises by an average of five percent annually, while the number of radiology residency positions only grows by two percent. Smith-Bindman, Kwan, Marlow and et al. (2019). While imaging modalities such as MRI, CT, and ultrasound can assist in the diagnosis of wrist abnormalities, some fractures may still be occult Fotiadou, Patel, Morgan and Karantanas (2011); Welling et al. (2008); Neubauer et al. (2016).

Recent advances in computer vision, more specifically, object detection have shown promising results in medical settings. Some of the positive results of detecting pathologies in trauma X-rays were recently published Adams, Henderson, Yi and Babyn (2020); Tanzi et al. (2020); Chung et al. (2018); Choi et al. (2020). Computer vision algorithms are accurate, efficient, and more importantly extremely quick to produce results compared to any radiologist or other imaging modalities currently used in practice. For example, radiology imaging delays have been found to independently contribute to longer hospital stays, as indicated by a recent study Courane, Conway, Creagh et al. (2016). In addition, a separate study Perotte, Lewin, Tambe and et al. (2018) found that creating reports from CT scans often took over three hours, with radiologists being responsible for a significant portion (42%) of the delay. The delays in obtaining clinically relevant information can have significant impacts on patients and contribute to unnecessary burdens on health systems, patients, and insurers. Computer vision algorithms can potentially address the delays associated with radiographic interpretation by providing a more efficient and prompt alternative, while still achieving comparable or even superior results.

ORCID(s):

Object detection has emerged as a powerful tool for identifying abnormalities in X-ray images. Its ability to locate and classify various objects within an image has made it a valuable asset in the diagnosis and treatment of various medical conditions. In recent years, significant progress has been made in the development of object detection algorithms, leading to their widespread adoption in the medical community. An earlier approach called the sliding window approach Lampert, Blaschko and Hofmann (2008) for object detection involved dividing an image into a grid of overlapping regions and then classifying each region as containing the object of interest or not. There are several disadvantages of this approach, one of them being that it is computationally expensive as a large number of regions need to be classified. To address these issues, region-based methods were invented. The introduction of Region-based Convolutional Neural Network (R-CNN) Girshick, Donahue, Darrell and Malik (2013b) was the first breakthrough in the application of region-based methods. The main idea behind these methods was to generate candidate object regions and classify only those regions as containing the object of interest or not.

Another method developed as an improvement over the sliding window approach was the single-stage detection method which has gained popularity in recent years due to its efficiency and good performance. This approach uses a single forward propagation through the network to predict bounding boxes and class probabilities, eliminating the need to generate candidate object regions, and making it faster than region-based approaches. While two-stage detection generates candidate regions in the first stage and refines them in the second stage at the cost of speed, single-stage detection provides a balance between speed and accuracy by predicting final results in a single pass through the network.

Two-stage detection has been the most widely used approach for detecting wrist abnormalities in recent years. However, there has been limited research on the effectiveness of single-stage detectors in detecting various abnormalities in the wrist, including fractures. In this study, we focus on the effectiveness of single-stage detectors in detecting wrist abnormalities, more specifically, we focus on the capabilities of recent versions of the YOLO algorithm (v5, v6, v7, and v8). Additionally, this study is unique in its use of a large, comprehensively annotated dataset called GRAZPEDWRI-DX presented in a recent publication Nagy, Janisch, Hržić, Sorantin and Tschauner (2022). The characteristics and complexity of the dataset are discussed in section 4

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" Nagy et al. (2022) recently released to the public. We compare the performances of all variants within each YOLO model employed in this study 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. We hypothesize that fractures in the near vicinity of the wrist in pediatric X-ray images can be detected efficiently using YOLOv5, YOLOv6, YOLOv7, and YOLOv8 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 prove our hypothesis using the comprehensively annotated GRAZPEDWRI-DX dataset.

The general objective of the study is to use the GRAZPEDWRI-DX dataset to 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 detection models effectively detect fractures in the vicinity of the wrist in pediatric X-ray images?
2. In the analysis of wrist images, do the state-of-the-art 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 in the vicinity of the wrist in pediatric X-ray images?

1.2. Contribution

The major contributions of this article are as follows:

- A thorough performance assessment of state-of-the-art YOLO detection models (YOLOv5, YOLOv6, YOLOv7, and YOLOv8) 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, including compound, medium, and smaller-scaled versions.
- Achieved state-of-the-art mean average precision (mAP) score on the (GRAZPEDWRI-DX dataset).
- A detailed analysis of the performance of single-stage detection models in comparison to the two-stage detection model widely used in the literature, namely, Faster R-CNN.

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. For instance, In a study by Yahalomi, Chernofsky and Werman (2018), a Faster R-CNN model utilizing Visual Geometry Group (VGG16) was applied to identify distal radius fractures in anteroposterior wrist X-ray images. The model achieved a mAP of 0.87 when tested on a set of 1,312 images. It should be noted that the initial dataset consisted of only 95 anteroposterior images, with and without fractures, which were then augmented for training as well as for testing.

Thian, Li, Jagmohan, Sia, Chan and Tan (2019) developed two separate Faster R-CNN models with Inception-ResNet for frontal and lateral projections of wrist images. The models were trained on 6,515 and 6,537 images of frontal and lateral projections, respectively. 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 a high area under the receiver operating characteristic curve (AUC-ROC) values, with the frontal model having 0.92 and the lateral model having 0.93. The overall per-study specificity was 0.73, sensitivity was 0.98, and AUC was 0.89.

Guan, Zhang, Yao, Wang and Wang (2020) used a two-stage R-CNN method to achieve an average precision (AP) of 0.62 on approximately 4,000 X-ray images of arm fractures in musculoskeletal radiographs, MURA dataset. Wang, Yao, Zhang, Guan, Wang and Zhang (2021) developed a two-stage R-CNN network called ParallelNet, with a TripleNet backbone network, for fracture detection in a dataset of 3,842 thigh fracture X-ray images, achieving an AP of 0.88 at an Intersection over Union (IoU) threshold of 0.5.

Qi, Zhao, Shi, Zuo, Zhang, Long, Wang and Wang (2020) used a Faster R-CNN model with an anchor-based approach, combined with a multi-resolution Feature Pyramid Network (FPN) and a ResNet50 backbone network. They tested the model on 2333 X-ray images of different types of femoral fractures and obtained a mAP score of 0.69.

Raisuddin, Vaattovaara, Nevalainen and et al. (2021) developed a deep learning-based pipeline called DeepWrist for detecting distal radius fractures. The model was trained on a dataset of 1946 wrist studies and was evaluated on two test sets. The first test set, comprising 207 cases, resulted in an AP score of 0.99, while the second test set, comprising 105 challenging cases, resulted in an AP of 0.64. The model generated heatmaps to indicate the probability of a fracture near the vicinity of the wrist but did not provide a bounding box or polygon to clearly locate the fracture. The study was limited by the use of a small dataset with a disproportionate number of challenging cases.

Ma and Luo (2021) in their study, first classified the images in the Radiopaedia dataset into the fracture and non-fracture categories using CrackNet. After this, they utilized Faster R-CNN for fracture detection on the 1052 bone images in the dataset. With an accuracy of 0.88, a recall of 0.88, and a precision of 0.89, they demonstrated the usefulness of the proposed approach. Wu, Yan, Liu, Yu, Geng, Wu, Han, Guo and Gao (2021) applied a Feature Ambiguity Mitigate Operator model along with ResNeXt101 and a FPN to identify fractures in a collection of 9040 radiographs of various body parts, including the hand, wrist, pelvic, knee, ankle, foot, and shoulder. They accomplished an AP of 0.77.

Xue, Yan, Luo, Zhang, Chaikovska, Liu, Gao and Yang (2021) proposed a guided anchoring method (GA) for fracture detection in hand X-ray images using the Faster R-CNN model, which was used to forecast the position of fractures using proposal regions that were refined using the GA module's learnable and flexible anchors. They evaluated the method on 3067 images and achieved an AP score of 0.71.

Hardalaç, Uysal, Peker, Çiçeklidağ, Tolunay, Tokgöz, Kutbay, Demirciler and Mert (2022) conducted 20 fracture detection experiments using a dataset of wrist X-ray images from Gazi University Hospital. To improve the results, they developed an ensemble model by combining five different models, named WFD-C. Out of the 26 models evaluated for fracture detection, the WFD-C model achieved the highest average precision of 0.86. This study utilized both two-stage and single-stage detection methods. The two-stage models employed were Dynamic R-CNN, Faster R-CNN, and SABL and DCN models based on Faster R-CNN. Meanwhile, the single-stage models used were PAA, FSAF, RetinaNet and RegNet, SABL, and Libra.

Joshi, Singh and Joshi (2022) employed transfer learning with a modified Mask R-CNN to detect and segment fractures using two datasets: a surface crack image dataset of 3000 images and a wrist fracture dataset of 315 images. They first trained the model on the surface crack dataset and then fine-tuned it on the wrist fracture dataset. They achieved an average precision of 92.3% for detection and 0.78 for segmentation on a 0.5 scale, 0.79 for detection, and 0.52 for segmentation on a strict 0.75 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 performance 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 number 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 majority 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, Austria. 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.

3.2. Study Dimensions

The following dimensions are used to facilitate the interpretation of results:

- Abnormality-(*ab*): The object detection models were evaluated on their ability to detect different types of abnormalities in the radiographic images.
- Fracture-(*f*): The object detection models were also evaluated on their ability to effectively detect fractures in the radiographic images.
- Recall-(*r*): The proportion of positive instances that were correctly detected by the model. The calculation of recall is TP / (TP + FN), where TP represents the number of true positive detections and FN the number of false negative detections.

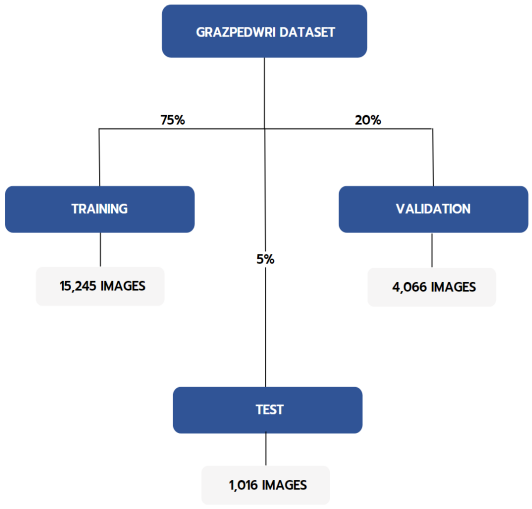


Figure 1: Dataset split into training, validation, and test sets.

- Precision-(*p*): The proportion of positive detections that were actually positive instances. It is calculated by dividing the number of true positive detections (TP) by the sum of true positives and false positives (incorrect detections) represented as TP / (TP + FP), where FP stands for false positive detections.
- Mean Average Precision-(*mAP*): mAP is a performance metric used to evaluate an object detection model with an intersection over union (IoU) threshold of 0.5. It's a widely adopted evaluation method for object detection models as it takes into account both precision and recall.

3.3. 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.

3.4. Deep Learning Models For Object Detection

In this study, we employed 4 single-stage detection models, 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 variations of each YOLO model, ranging from 5 to 7 variants. This resulted in a total of 23 wrist abnormality detection procedures.

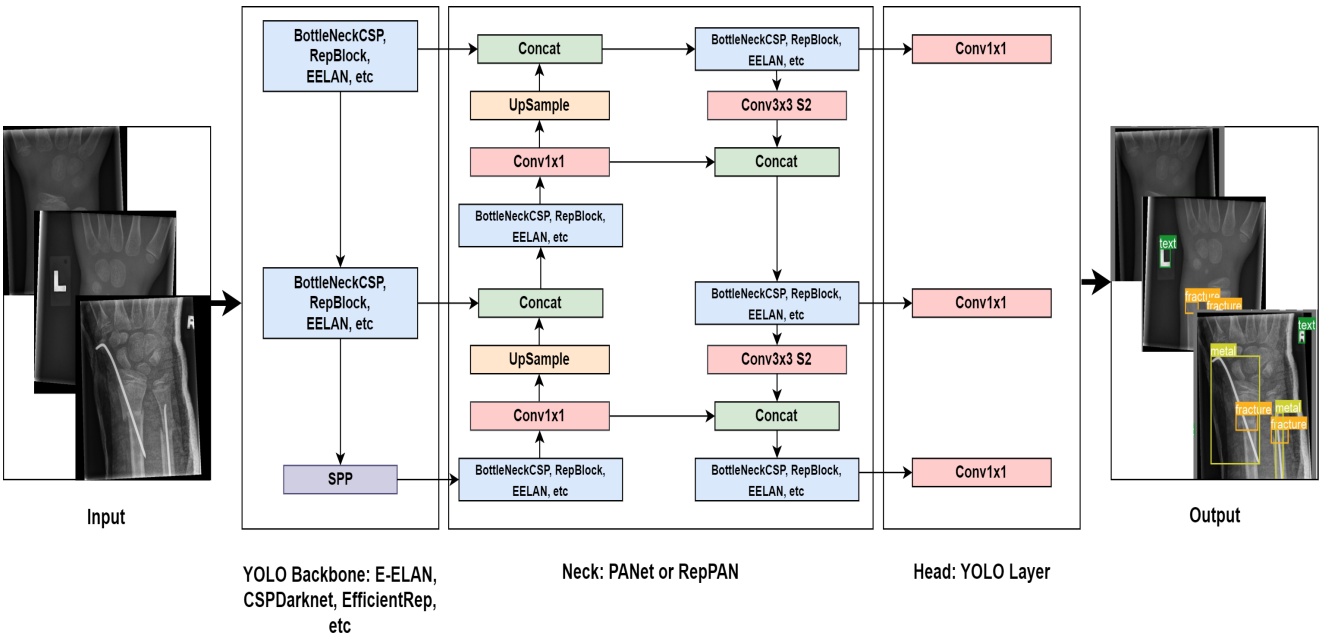


Figure 2: YOLO Architecture depicting the input, backbone, neck, head, and the output.

We conducted initial training on various variants of each YOLO model. Subsequently, we selected the highest-performing variant within each YOLO model based on the results obtained and compared them to the two-stage detection model Faster R-CNN. The models that were trained for 100 epochs, were observed to converge between 90-100 epochs, indicating no additional improvement beyond the 100th epoch, thus further training was deemed unnecessary.

The YOLO (You Only Look Once) algorithm, initially introduced by Redmon, Divvala, Girshick and Farhadi (2015) in 2016, is a single-stage object detection approach that uses a single pass of a convolutional neural network (CNN) 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 v7 Redmon et al. (2015); Redmon and Farhadi (2016); Farhadi and Redmon (2018); Bochkovskiy et al. (2020); Jiang, Ergu, Liu, Cai and Ma (2022), with v5, v6, v7, and v8 being released in 2020, 2022, and 2023 offering smaller volume, higher speed, and higher precision Solawetz (2020); Nelson (2020); Jiang et al. (2022). Fig. 2 illustrates the general structure of YOLO with backbones used in this study such as CSP, VGG, and EELAN.

R-CNN proposed by Girshick, Donahue, Darrell and Malik (2013a) was one of the first algorithms to achieve state-of-the-art performance on the PASCAL VOC object detection benchmark. R-CNN is a two-stage algorithm that takes an entire image as input, generates regions likely containing objects, extracts features using a CNN, and classifies objects within these regions. Faster R-CNN is a widely adopted and well-established model within the R-CNN family, known for its efficiency and accuracy in object detection. It has been widely utilized in the medical field, specifically in the detection of bone fractures. The Faster R-CNN model, first introduced by Ren, He, Girshick and Sun (2015), has continued to be a significant and influential contribution to the field of computer vision, remaining one of the most highly cited papers in the field to this day.

3.4.1. The YOLOv5 Model

The YOLO framework consists of three main components: the backbone, the neck, and the head. First, the input terminal performs various data processing tasks, including adaptive image filling and mosaic data augmentation Solawetz (2020). In our research, we have utilized the same data augmentation and pre-processing methods. Additionally, the YOLOv5 model uses adaptive anchor frame calculation to optimize its performance on different datasets by adjusting its anchor frame size when the dataset changes.

The backbone is responsible for extracting image features. It aggregates and forms features at different granularities. The specific CNN architecture used in YOLOv5 is CSPDarknet since this is the best-performing one so far Wang, Liao, Wu, Chen, Hsieh and Yeh (2020). Hence, in our work, we have utilized the same CNN architecture. The CSPDarknet architecture consists of convolutional, pooling, and residual connections represented mathematically as:

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

where F_i is the feature maps at the i-th layer, F_{i-1} is the feature maps at the $(i-1)$-th layer, W_i are the weights and biases at the i-th layer, and $f(\cdot)$ is the function applying convolution and pooling operations. The SPP structure is then

applied to the feature maps produced by the CSPDarknet to extract features at multiple scales. This can be represented mathematically as:

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

where F_{SPP} represents the multi-scale feature maps produced by the SPP structure, and $g(\cdot)$ represents the function that applies the SPP operation to the input feature maps F_i.

The neck of YOLOv5 uses Path Aggregation Network (PANet) to aggregate features from the backbone and produce higher-level features for the output layers. The same architecture is used in our study. The head constructs output vectors with class probabilities, objectness scores, and the bounding box (coordinates for the box: center, height, and width) representing objects in the image. The utilized head is the same as the original implementation.

The YOLOv5 model includes five different model variants, namely, YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. All are compound-scaled variants of the same architecture. Each of these offers different detection accuracy and performance. The differences between these variants are the number of channels i.e. (the depth of the network), and the number of layers.

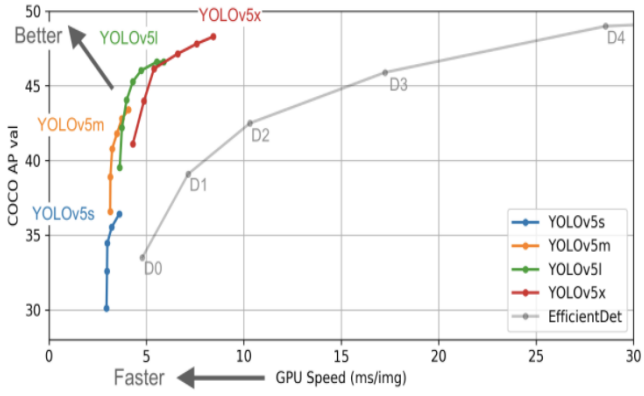


Figure 3: Performance comparison of YOLOv5 variants on COCO ultralytics (2022).

In the experimentation of YOLOv5 variants, standard hyperparameters were utilized. The input resolution was fixed at 640 pixels, and the batch size was set to 16. The optimization algorithm employed was Stochastic Gradient Descent (SGD) with an initial learning rate $\alpha = 1 \times 10^{-2}$, final learning rate $\alpha_f = 1 \times 10^{-2}$, momentum = 0.937, weight decay = 5×10^{-4}, warmup momentum = 0.8, and warmup bias lr = 0.1. Each variant underwent 100 epochs of training from scratch. During the evaluation phase, each variant was tested on 1016 randomly selected samples, using an Intersection over Union (IoU) threshold of 0.5 for inference.

3.4.2. 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 also includes five different model variants, namely, YOLOv6n, YOLOv6s, YOLOv6m, YOLOv6l, and YOLOv6l6. Fig. 4 shows the performance comparison of these variants on the COCO dataset.

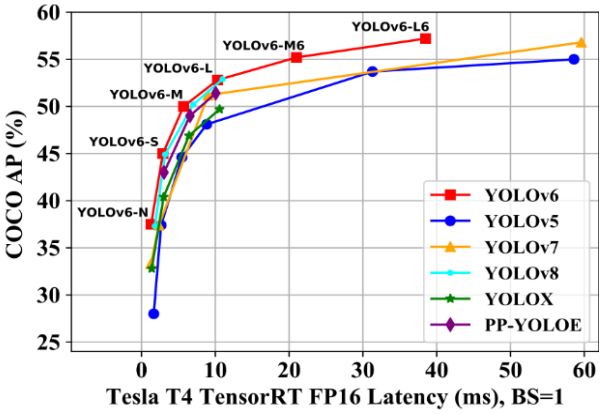


Figure 4: Performance comparison of YOLOv6 variants on COCO meituan (2023).

All 5 variants of YOLOv6 were trained for 100 epochs from scratch. Standard hyperparameters were used. The input was set to 640 pixels with a batch size of 16 samples. The optimization algorithm employed was Stochastic Gradient Descent (SGD) with the same parameters as used for YOLOv5, including initial and final learning rates, momentum, warmup momentum, weight decay, and warmup bias lr. As before, each variant was tested on 1016 randomly selected samples, using an Intersection over Union (IoU) threshold of 0.5 for inference.

3.4.3. The YOLOv7 Model

YOLOv7 also has three main components discussed before with several changes. The E-ELAN is a component in YOLOv7 that uses expand, shuffle, and merge cardinality to continuously improve network learning without disrupting the gradient path. Wang et al. (2022). Other notable changes include Model Scaling techniques, Reparameterization planning, and Auxiliary Head Coarse-to-Fine. Model scaling is a technique used to adapt key characteristics of a model to align with specific application requirements. This includes adjusting the width (number of channels), depth (number of stages), and resolution (input image size) of the model. The scaling of object detection models requires knowledge of the network depth, width, and resolution on which it is trained.

YOLOv7 utilizes a compound scaling technique, which simultaneously scales the depth and width of the network by concatenating layers. This method has been shown through ablation studies to maintain optimal model architecture

while scaling for different sizes. Without this technique, an increase in depth alone may cause a decrease in hardware efficiency due to a change in the ratio between input and output channels of a transition layer. YOLOv7's compound scaling technique prevents such negative effects on performance. Re-parameterization techniques aim to create a more robust model by averaging a set of weights. Recent research has focused on module-level re-parameterization, where specific parts of the network are targeted. YOLOv7 uses gradient flow propagation to determine which modules should be re-parameterized. The YOLO network head generates the final predictions, however, an auxiliary head located in the middle of the network can be beneficial during training. The auxiliary head is supervised along with the final head. However, it does not train as efficiently because it is closer to the prediction, thus the YOLOv7 authors experimented with different levels of supervision for the auxiliary head, using a coarse-to-fine approach where supervision is passed back from the final head at various granularities.

The YOLOv7 model comprises of seven different variants, which include "P5" models (v7, v7x, and v7-tiny) and "P6" models (d6, e6, w6, and e6e). These variants are compound-scaled versions of the same architecture, each of which offers a different level of detection accuracy and performance when trained on the standard COCO dataset. This variation in performance is illustrated in Fig. 5.

In the experimentation of these 7 variants of the YOLOv7 model, all variants underwent a training phase with a duration of 100 epochs from scratch with 16 samples as batch size. The standard hyperparameters were applied and the optimization algorithm employed was Stochastic Gradient Descent (SGD) with an initial learning rate $\alpha = 1 \times 10^{-2}$, final learning rate $\alpha_f = 1 \times 10^{-1}$, momentum = 0.937, weight decay = 5×10^{-4}, warmup momentum = 0.8, and warmup bias lr = 0.1. The "P5" models within YOLOv7, namely, v7, v7x, and v7-tiny were trained with an input resolution of 640 pixels, while the "P6" models, namely, d6, e6, w6, and e6e were trained with an input resolution of 448 pixels due to computational constraints, and although it may have a negative effect on the model's performance, this issue is compensated for by utilizing mosaic augmentation within YOLOv7. Mosaic augmentation is a technique used to increase the diversity of training data by combining multiple small images into a larger "mosaic" image. This can help to improve the robustness of object detection models by exposing them to a wider variety of object scales, orientations, and backgrounds. During the evaluation phase, all variants were tested on a test set of 1016 samples, using an Intersection over Union (IoU) threshold of 0.5.

3.4.4. The YOLOv8 Model

YOLOv8 is reported to provide significant advancements in object detection as well as image segmentation when compared to previous YOLO models, particularly in compact versions that are implemented on less powerful hardware. At the time of writing this paper, the architecture of YOLOv8 is not fully disclosed and some of its features are

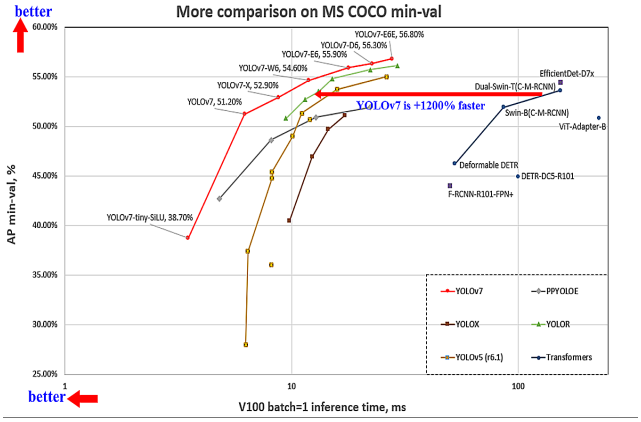


Figure 5: Performance comparison of YOLOv7 variants on COCO Wang et al. (2022).

still under development. As of now, it's been confirmed that the system has a new backbone, uses an anchor-free design, has a revamped detection head, and has a newly implemented loss function. We have included the performance of this model on the GRAZPEDWRI-DX dataset as a benchmark for future studies, as further improvements to YOLOv8 may surpass the results obtained in this study.

YOLOv8 comes in five versions at the time of release (January 10, 2023), namely, "n", "s", "m", "l", and "x". The performance of these variants on a COCO dataset is shown in Fig. 6. Just as before, all variants underwent 100 epochs of training from scratch, with standard hyperparameters. The image resolution was set at 640 pixels, with a batch size of 16 samples. The optimization algorithm employed was Stochastic Gradient Descent (SGD) as before with a starting learning rate of 1×10^{-2}, final learning rate of 1×10^{-2}, the momentum of 0.937, weight decay of 5×10^{-4}, warmup momentum of 0.8, and warmup bias lr of 0.1. As before, all variants were tested on a test set of 1016 samples, using an IoU threshold of 0.5.

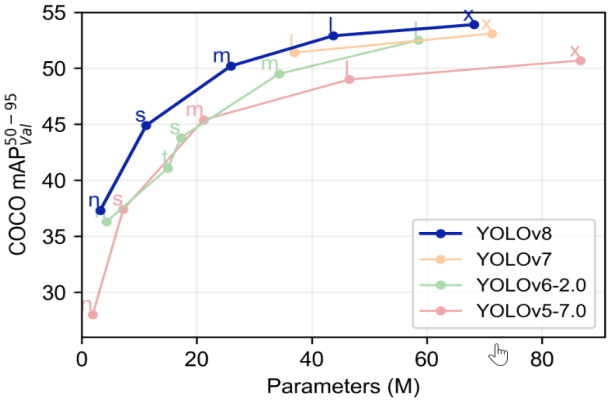


Figure 6: Performance comparison of YOLOv8 variants on COCO ultralytics (2023).

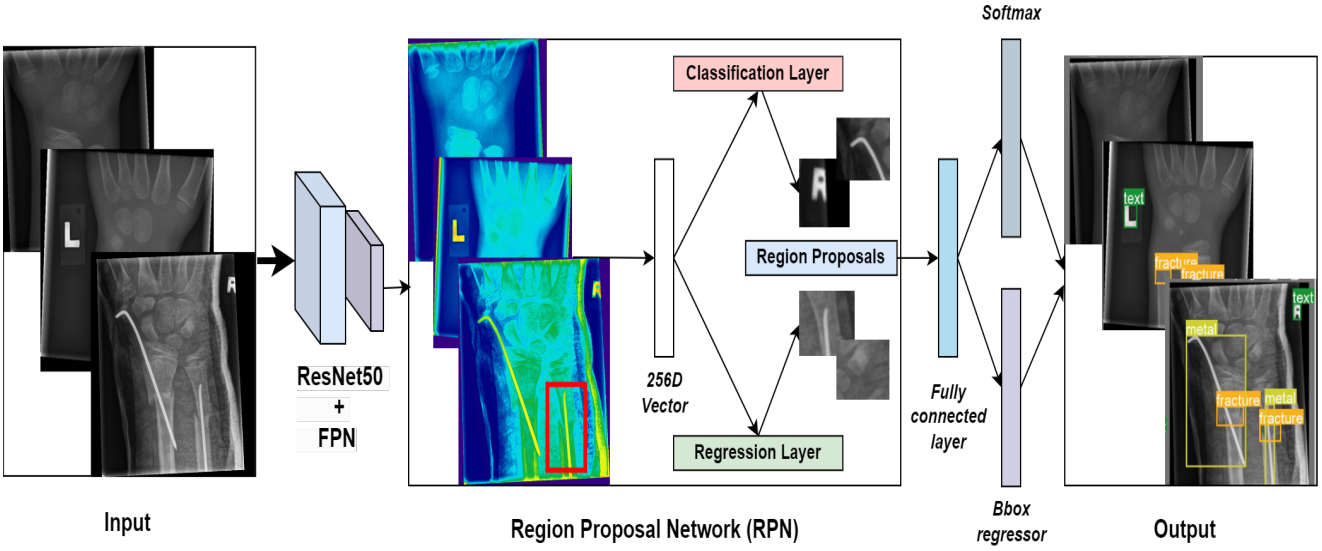


Figure 7: Faster R-CNN Pipeline.

3.4.5. Faster R-CNN

The Faster R-CNN model consists of three main components: the backbone, the region proposal network (RPN), and a detection network. In this study, a ResNet50 backbone with FPN was used for feature extraction from the input image. Anchors were generated for each feature, and a set of anchor boxes with variable sizes and aspect ratios were created for each anchor. The RPN was responsible for selecting appropriate anchor boxes and passing them onto the next layer. The classifier within RPN predicted if an anchor box contained an object, determined by an IoU threshold of 0.5. The regressor within RPN predicted offsets for the anchor boxes containing objects to fit them tightly to the ground truth labels. Lastly, the RoI pooling layer converted variable-sized proposals to a fixed size to run a classifier and regress a bounding box. Fig. 7 illustrates the architecture of Faster R-CNN.

The Faster R-CNN model underwent 100 epochs with an image size of 640 pixels and 16 samples as batch size. The default parameters were used and the optimization algorithm employed was Stochastic Gradient Descent (SGD) with a learning rate of $\alpha = 1{\times}10^{-3}$, a momentum of 0.9, and weight decay of $5{\times}10^{-4}$. It is important to note that, as with YOLO models, the selection of these parameters is not deliberate, they are the default settings. During the evaluation phase, each variant was tested on 1016 randomly selected samples, using an Intersection over Union (IoU) threshold of 0.5 for inference.

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. 8. 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 as:

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

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.

3.6. Supplementary Materials

The supplementary materials, including source code, and dataset split can be accessed through the following links:

- Source Code
- Dataset Split

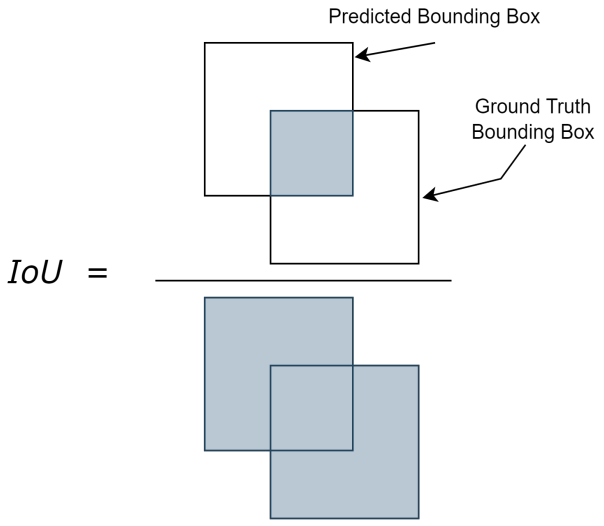


Figure 8: Visual illustration of Intersection over Union (*IoU*).

4. 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. 9 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%

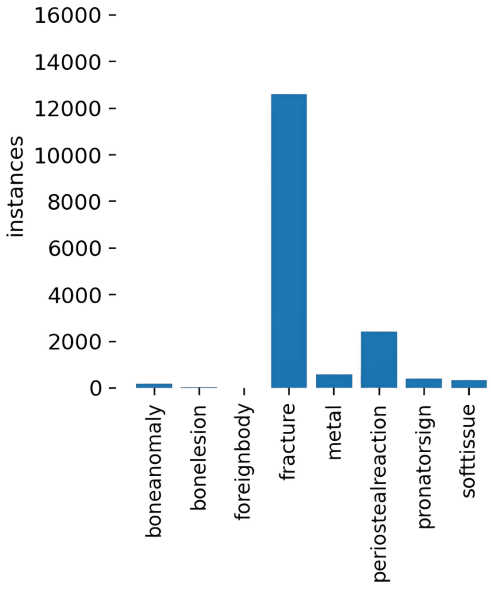
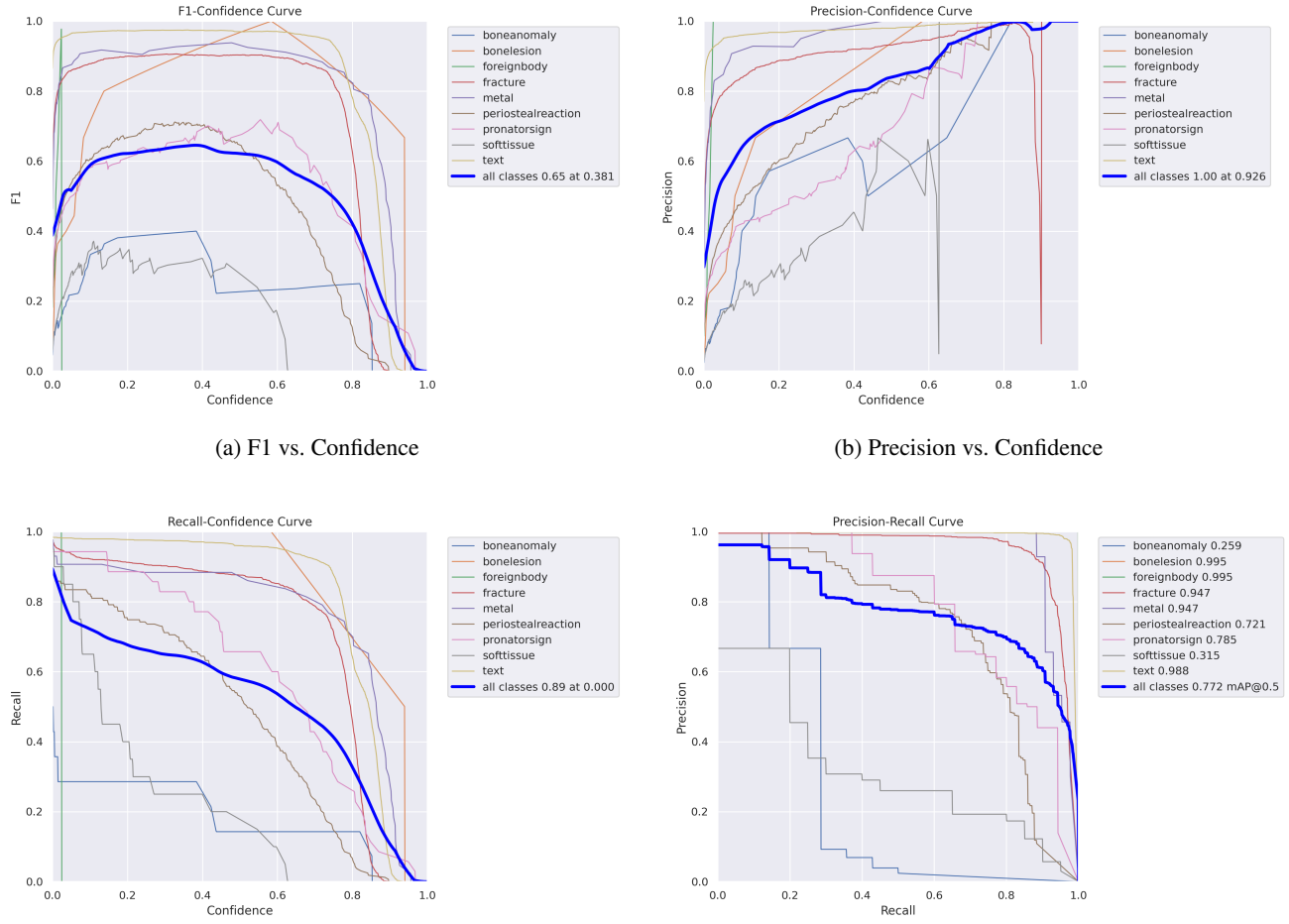


Figure 9: Histogram of Class Distribution.

In Table 2, we show the number of images in which a particular anomaly occurs only once, twice, or multiple times. The column "Total" represents the total number of images in which a particular anomaly is present.



(a) F1 vs. Confidence

(b) Precision vs. Confidence

(c) Recall vs. Confidence

(d) Precision vs. Recall

Figure 10: Performance analysis curves (YOLOv8x)

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

5. 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 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

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

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 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 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.

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

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

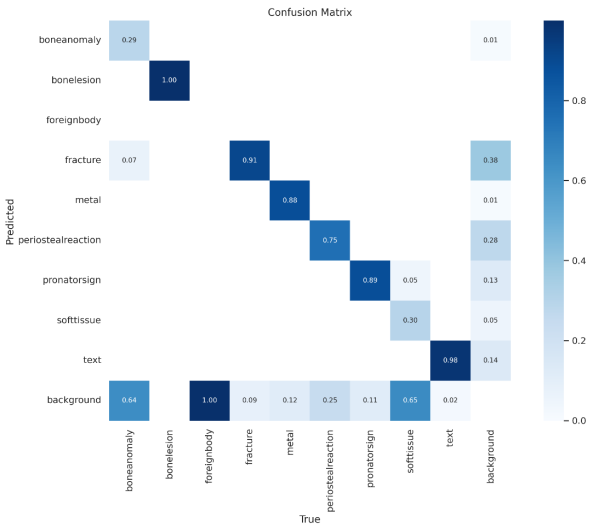


Figure 11: Confusion Matrix (YOLOv8x).

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 12 and 13 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 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. 14.

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. 10a, 10b, 10c, and 10d 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

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

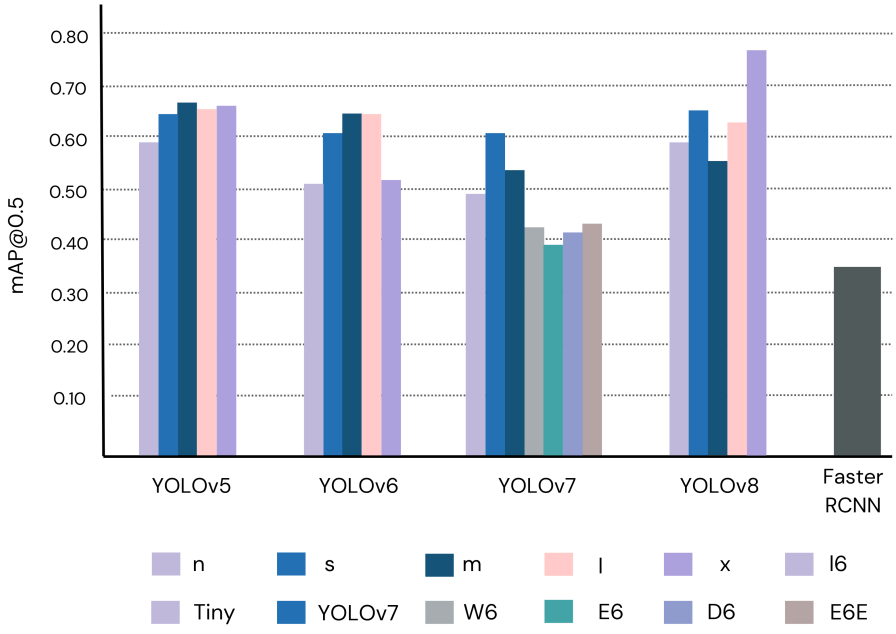


Figure 12: mAP Scores (Across All Classes).

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

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 11 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.

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

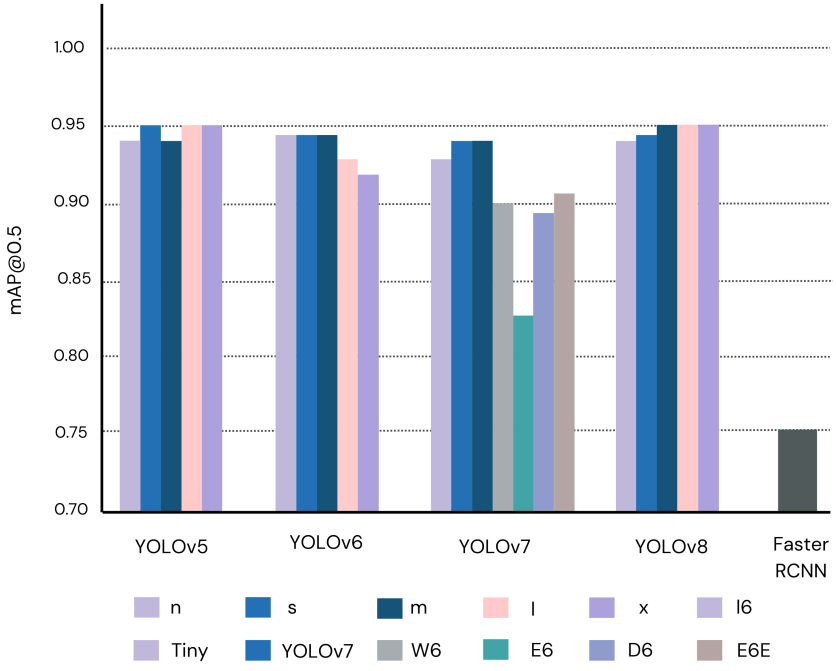


Figure 13: mAP Scores (Fracture Class).

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

6. 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.

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

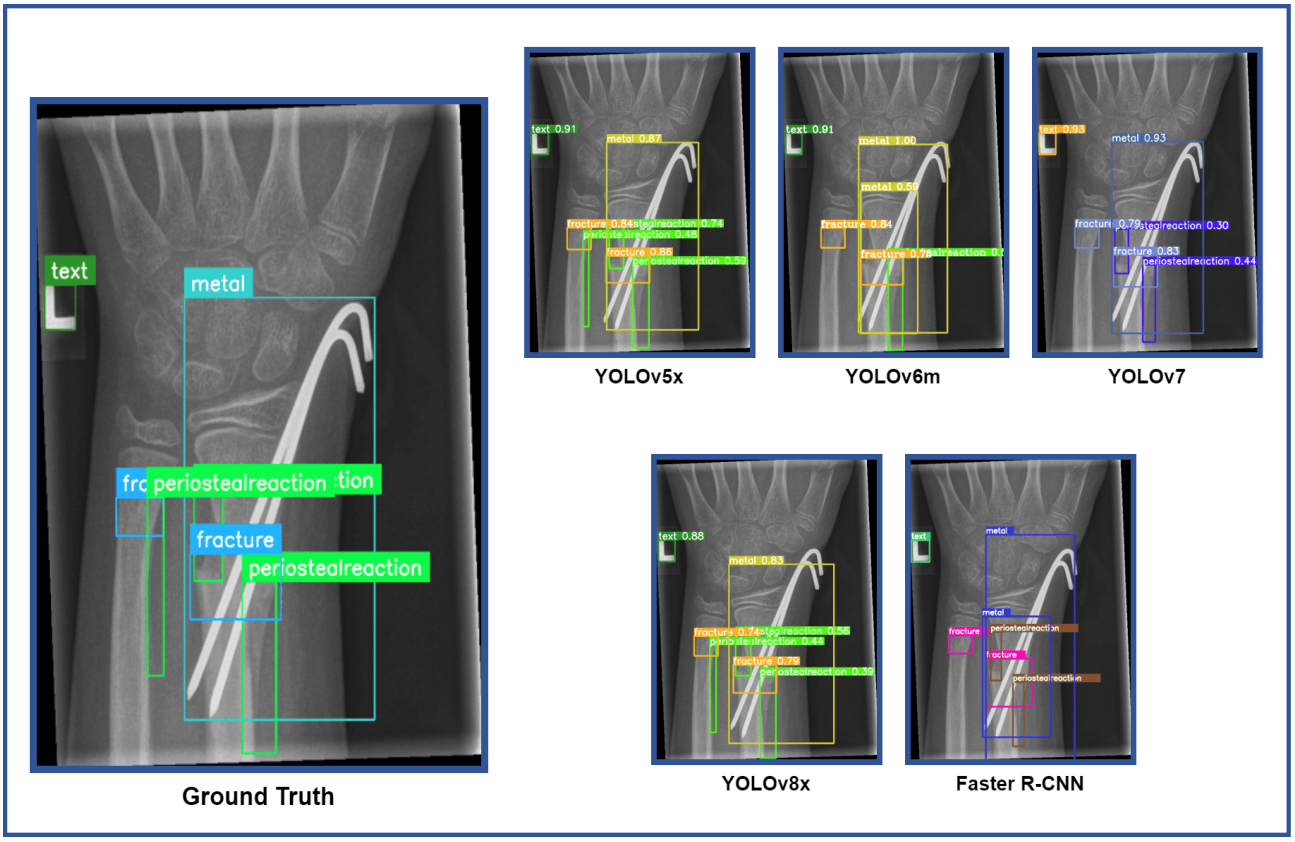


Figure 14: Bounding box estimated by YOLO variants and Faster R-CNN.

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

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 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 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.

7. Acknowledgement

This work was supported in part by the Department of Computer Science (IDI), Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway; and in part by the Curricula Development and Capacity Building in Applied Computer Science for Pakistani Higher Education Institutions (CONNECT) Project NORPART-2021/10502, funded by DIKU.

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.

Chung, S.W., et al., 2018. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 89, 468–473.

Courane, S., Conway, R., Creagh, D., et al., 2016. Radiology imaging delays as independent predictors of length of hospital stays for emergency medical admissions. Clin Radiol 71, 912–8.

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

Farhadi, A., Redmon, J., 2018. Yolov3: An incremental improvement. arXiv URL: https://arxiv.org/abs/1804.02767.

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.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2013a. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524 .

Girshick, R.B., Donahue, J., Darrell, T., Malik, J., 2013b. Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR abs/1311.2524. URL: http://arxiv.org/abs/1311.2524, arXiv:1311.2524.

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.

Jiang, P., Ergu, D., Liu, F., Cai, Y., Ma, B., 2022. A review of yolo algorithm developments. Procedia Computer Science 199, 1066–1073. doi:10.1016/j.procs.2022.01.135.

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, arXiv:https://www.pnas.org/content/115/47/11591.full.pdf.

Ma, Y., Luo, Y., 2021. Bone fracture detection through the two-stage system of cracksensitive convolutional neural network. Informatics in Medicine 236, 24–40.

Makary, M.S., Takacs, N., 2022. Are we prepared for a looming radiologist shortage? URL: https://www.diagnosticimaging.com/view/are-we-prepared-for-a-looming-radiologist-shortage-.

meituan, 2023. Yolov6: a single-stage object detection framework dedicated to industrial applications. https://github.com/meituan/YOLOv6.

Mounts, J., Clingenpeel, J., McGuire, E., Byers, E., Kireeva, Y., 2011. Most frequently missed fractures in the emergency department. Clin Pediatr (Phila) 50, 183–186.

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. doi:10.1038/s41597-022-01328-z.

Nelson, S., 2020. YOLOv5 is Here. URL: https://blog.roboflow.com/yolov5-is-here/.

Neubauer, J., et al., 2016. Comparison of diagnostic accuracy of radiation dose-equivalent radiography, multidetector computed tomography and cone beam computed tomography for fractures of adult cadaveric wrists. PLoS One 11, e0164859. doi:10.1371/journal.pone.0164859.

Perotte, R., Lewin, G., Tambe, U., et al., 2018. Improving emergency department flow: reducing turnaround time for emergent ct scans. AMIA Annu Symp Proc , 897–906.

Qi, Y., Zhao, J., Shi, Y., Zuo, G., Zhang, H., Long, Y., Wang, F., Wang, W., 2020. Ground truth annotated femoral x-ray image dataset and object detection based method for fracture types classification. IEEE Access 8, 189436–189444.

Raisuddin, A., Vaattovaara, E., Nevalainen, M., et al., 2021. Critical evaluation of deep neural networks for wrist fracture detection. Scientific Reports 11, 6006. doi:10.1038/s41598-021-85570-2.

Randsborg, P.H., et al., 2013. Fractures in children: epidemiology and activity-specific fracture rates. The Journal of Bone and Joint Surgery - American Volume 95, e42.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2015. You only look once: Unified, real-time object detection. arXiv URL: https://arxiv.org/abs/1506.02640.

Redmon, J., Farhadi, A., 2016. Yolo9000: Better, faster, stronger. URL: https://arxiv.org/abs/1612.08242.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 91–99.

Rimmer, A., 2017. Radiologist shortage leaves patient care at risk, warns royal college. BMJ 359.

Rosman, D.e.a., 2015. Imaging in the land of 1000 hills: Rwanda radiology country report .

Sha, G., Wu, J., Yu, B., 2020a. Detection of spinal fracture lesions based on improved yolov2, in: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 235–238. doi:10.1109/ICAICA50127.2020.9182582.

Sha, G., Yu, B., Wu, J., 2020b. Detection of spinal fracture lesions based on improved faster-rcnn, in: 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), pp. 29–32. doi:10.1109/ICAIIS49377.2020.9194863.

Smith-Bindman, R., Kwan, M., Marlow, E., et al., 2019. Trends in use of medical imaging in us healthcare systems and in ontario, canada, 2000-2016. JAMA 322, 843–856.

Solawetz, 2020. What is yolov5? a guide for beginners. https://blog.roboflow.com/yolov5-improvements-and-evaluation/.

Tanzi, L., et al., 2020. Hierarchical fracture classification of proximal femur x-ray images using a multistage deep learning approach. Eur J Radiol 133, 109373.

Thian, Y.L., Li, Y., Jagmohan, P., Sia, D., Chan, V.E.Y., Tan, R.T., 2019. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiology: Artificial Intelligence 1, e180001.

ultralytics, 2022. Yolov5 in pytorch > onnx > coreml > tflite. URL: https://github.com/ultralytics/yolov5.

ultralytics, 2023. Yolov8 pytorch > onnx > coreml > tflite. https://github.com/ultralytics/ultralytics.

Wang, C., Bochkovskiy, A., Liao, H.M., 2022. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv.org URL: https://arxiv.org/abs/2207.02696.

Wang, C., Liao, H., Wu, Y., Chen, P., Hsieh, J., Yeh, I., 2020. Cspnet: A new backbone that can enhance learning capability of cnn, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Springer International Publishing, Cham. pp. 390–391.

Wang, M., Yao, J., Zhang, G., Guan, B., Wang, X., Zhang, Y., 2021. Parallelnet: Multiple backbone network for detection tasks on thigh bone fracture. Multimedia Systems 27, 1091–1100.

Welling, R.D., et al., 2008. Mdct and radiography of wrist fractures: radiographic sensitivity and fracture patterns. AJR. American journal of roentgenology 190, 10–16. doi:10.2214/AJR.07.2502.

Wu, H.Z., Yan, L.F., Liu, X.Q., Yu, Y.Z., Geng, Z.J., Wu, W.J., Han, C.Q., Guo, Y.Q., Gao, B.L., 2021. The feature ambiguity mitigate operator model helps improve bone fracture detection on x-ray radiograph. Scientific Reports 11, 1589.

Xue, L., Yan, W., Luo, P., Zhang, X., Chaikovska, T., Liu, K., Gao, W., Yang, K., 2021. Detection and localization of hand fractures based on ga_faster r-cnn. Alexandria Engineering Journal 60, 4555–4562.

Yahalomi, E., Chernofsky, M., Werman, M., 2018. Detection of distal radius fractures trained by a small set of x-ray images and faster r-cnn. arXiv.org URL: https://arxiv.org/abs/1812.09025.