Biomedical Signal Processing and Control 93 (2024) 106144

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

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], Sher Muhammad Daudpota [a]

[a] Department of Computer Science, Sukkur IBA University, Sukkur, 65200, Pakistan
[b] Department of Computer Science, Norwegian University of Science & Technology (NTNU), Gjøvik, 2815, Norway
[c] Department 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 fracture detection. Additionally, compound-scaled variants of each YOLO model were compared, with YOLOv8 m demonstrating a highest fracture detection sensitivity of 0.92 and mean average precision (mAP) of 0.95. On the other hand, YOLOv6 m achieved the highest sensitivity across all classes at 0.83. Meanwhile, YOLOv8x recorded the highest mAP of 0.77 for all classes 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 [1–4]. 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. While X-ray is often the first and most common imaging modality used for wrist problems, the choice of test depends on the suspected abnormality, clinical presentation, and available resources. If an X-ray does not provide a clear diagnosis, other imaging modalities like MRI, CT, or ultrasound may be recommended. The obtained radiographs 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 [5]. Studies have shown that diagnostic errors in reading emergency X-rays can reach up to 26% [6–9]. This is compounded by the shortage of radiologists even in developed countries [10–12] and limited access to specialist reporting in other parts of the world [13] 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. [14]. While imaging modalities such as MRI, CT, and ultrasound can further assist in the diagnosis of wrist abnormalities, some fractures may still be occult [15,16].

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 [17–19]. In recent years, significant progress has

[*] Corresponding author.
E-mail addresses: ammarahmed.bscsf19@iba-suk.edu.pk (A. Ahmed), ali.imran@ntnu.no (A.S. Imran), zenun.kastrati@lnu.se (Z. Kastrati), sher@iba-suk.edu.pk (S.M. Daudpota).

https://doi.org/10.1016/j.bspc.2024.106144
Received 30 March 2023; Received in revised form 29 January 2024; Accepted 19 February 2024
Available online 23 February 2024

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 [20] 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. Key implementations of this method include cascade classifiers that employ LBP (Local Binary Patterns) or Haar-like features. These classifiers are trained using positive examples of a specific object set against random negative images of the same size. Once optimized, the classifier can accurately identify the target object within a specific section of an image. To detect the object throughout the whole image, the classifier systematically examines each segment. It is essential to differentiate between LBP and Haar-like features. LBP characterizes the local texture of an image by comparing a pixel to its neighboring ones, while Haar-like features measure differences in pixel intensities within neighboring rectangular areas. There are several disadvantages of the sliding window 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 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 and computational efficiency, 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 SOTA single-stage detectors in detecting wrist abnormalities. Additionally, this study is unique in its use of a large, comprehensively annotated dataset called GRAZPEDWRI-DX presented in a recent publication [21]. The characteristics and complexity of the dataset are discussed in Section 4.

Wrist fractures represent just one of several typical wrist abnormalities, other prevalent conditions include Carpal Tunnel Syndrome (CTS), Ganglion Cysts, Osteoarthritis, Tendinitis, as well as Sprains and Strains. Within the dataset that we use, the distinct objects are categorized as fracture, periostealreaction, metal, pronatorsign, softtissue, bone-anomaly, bonelesion, and foreignbody. It is crucial to understand that our primary goal is to detect these specific objects rather than diagnose the overarching abnormalities. In our context, the presence of these objects (including fractures) in the wrist can be considered as 'abnormal'. Moreover, the presence of objects other than fractures may suggest another associated wrist abnormality. For instance, soft tissue presence might be indicative of CTS or a ganglion cyst. In CTS, swelling of the synovial tissue that lines the tendons in the carpal tunnel may be observable. Conversely, a ganglion cyst manifests as a soft tissue structure. The term 'bone lesion' denotes an anomalous area within the bone, severe sprains can involve avulsion fractures where a fragment of bone is pulled away by the ligament.

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 [22], Li et al. [23], Wang et al. [24], and ultralytics [25] respectively.

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

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 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 sensitivity 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.
- A detailed performance analysis of single-stage detection models in comparison to the widely-used two-stage detection model, 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 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. 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 et al. [26], a Faster R-CNN model utilizing VGG16 was applied to identify distal radius fractures in anteroposterior wrist X-ray images. The model achieved an mAP of 0.87 when tested on a set of 1312 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 et al. [27] developed two separate Faster R-CNN models with Inception-ResNet for frontal and lateral projections of wrist images. The models were trained on 6515 and 6537 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 AUC-ROC value, with the frontal model having 0.92 and the lateral model having 0.93. The overall per-study specificity was 0.73, the sensitivity was 0.98, and the AUC-ROC value was 0.89.

Guan et al. [28] used a two-stage R-CNN method to achieve an average precision (AP) of 0.62 on approximately 4000 X-ray images of arm fractures MURA dataset. Wang et al. [29] developed a two-stage R-CNN network called ParallelNet, with a TripleNet backbone network, for fracture detection in a dataset of 3842 thigh fracture X-ray images, achieving an AP of 0.88 at an IoU threshold of 0.5.

Qi et al. [30] 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 an mAP score of 0.69.

Raisuddin et al. [31] 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 [32] 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 et al. [33] applied a Feature Ambiguity Mitigate Operator model along with ResNeXt101 and an 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 et al. [34] 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 et al. [35] 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 et al. [36] 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 et al. [37], a YOLOv2 model was used to detect fractures in a dataset of 5134 spinal CT images, resulting in an mAP of 0.75. In another research by the same authors [38], 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. [39] compared the performance of the YOLOv4 object detection model to that of the U-Net segmentation model proposed by Lindsey et al. [40] 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 [21]. The authors presented the baseline results for the dataset using the COCO pre-trained YOLOv5 m variant of YOLOv5. The model was trained on 15,327 (of 20,327) images and tested on 1000 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 [21,39] by conducting a comprehensive comparative study between the state-of-the-art single-stage detection algorithms (YOLOv5, v6, v7, and v8) 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 6091 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 4066, and a testing set of 1016.

In the context of binary classification, the dataset was partitioned into two categories: 'fracture' and 'No-fracture'. The 'fracture' class consisted of 13,549 images, while the 'No-fracture' class contained 6777 images. To establish training, testing, and validation sets, the dataset was split using the same ratio as for the aforementioned object detection task.

3.2. Tools & instruments

Python scripts were utilized for dataset partitioning. The PyTorch framework was employed for training object detection models. Whereas, the TensorFlow framework was employed for classification. The training process for all variants, excluding the P6 variants of YOLOv7 and the Binary Classifier, was conducted on a Windows PC equipped with an NVIDIA GeForce RTX 2080 SUPER graphics card with 8192 MB of VRAM, an Intel(R) Xeon(R) W-2223 CPU@3.60 GHz processor, and 64 GB of RAM. On the other hand, the training of the P6 variants and the binary classifiers took place on COLAB, utilizing an NVIDIA Tesla T4 GPU with 15,360 MB of VRAM and 12 GB of RAM. The Python version used was 3.9.13.

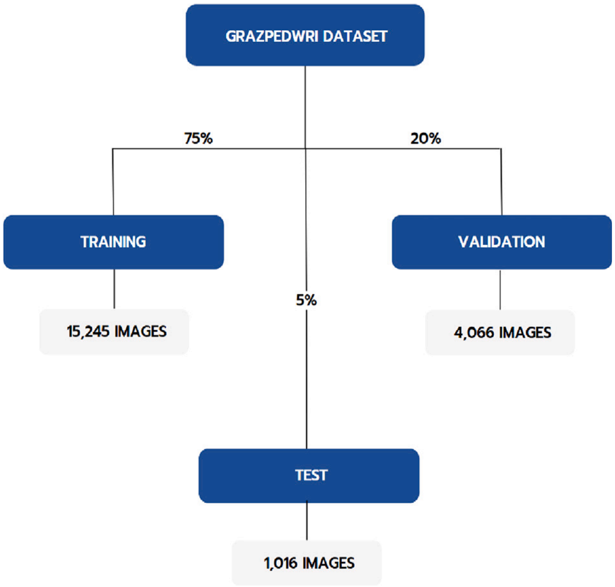


Fig. 1. Dataset split into training, validation, and test sets.

3.3. *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 variants of each YOLO model, ranging from 5 to 7 variants. This resulted in a total of 23 wrist abnormality detection procedures.

The YOLO (You Only Look Once) algorithm, initially introduced by Redmon et al. [41] 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 [42]. 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 [43]. We adopted the same architecture in our research. CSPDarknet involves convolution, pooling, and residual connections represented as:

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

(Where F_i and F_{i-1} are feature maps at ith 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:

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

(Where F_{SPP} denotes multi-scale feature maps, and $g(\cdot)$ performs 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 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. *The YOLOv8 model*

YOLOv8 is reported to provide significant advancements in object detection 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 still under development. As of now, it has 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".

3.3.5. *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. *Binary classifiers*

For our binary classification task (fracture/no-fracture), we employed a standard CNN architecture that consists of multiple convolutional layers, followed by pooling layers, fully connected layers, and a final layer with sigmoid activation. The detailed architecture of this conventional ConvNet is presented in s Fig. 4. we also explored the capabilities of DenseNets, which have been observed to outperform traditional CNNs in various tasks. It is important to note that for

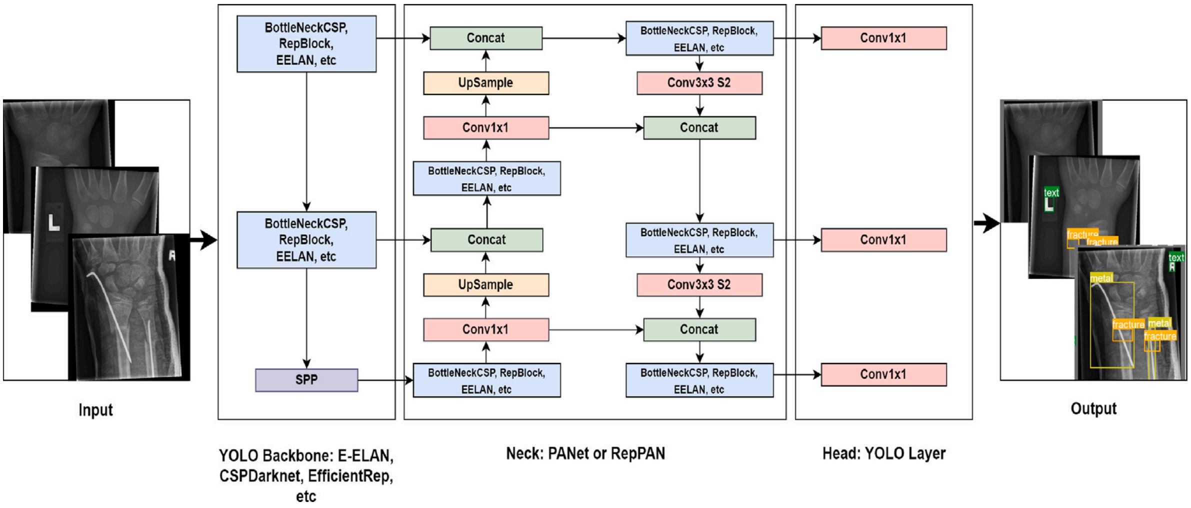

Input
BottleNeckCSP, RepBlock, EELAN, etc
BottleNeckCSP, RepBlock, EELAN, etc
SPP
Concat
UpSample
Conv1x1
BottleNeckCSP, RepBlock, EELAN, etc
Concat
UpSample
Conv1x1
BottleNeckCSP, RepBlock, EELAN, etc
BottleNeckCSP, RepBlock, EELAN, etc
Conv3x3 S2
Concat
BottleNeckCSP, RepBlock, EELAN, etc
Conv3x3 S2
Concat
BottleNeckCSP, RepBlock, EELAN, etc
Conv1x1
Conv1x1
Conv1x1
Output
text
fracture
metal
YOLO Backbone: E-ELAN, CSPDarknet, EfficientRep, etc
Neck: PANet or RepPAN
Head: YOLO Layer

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

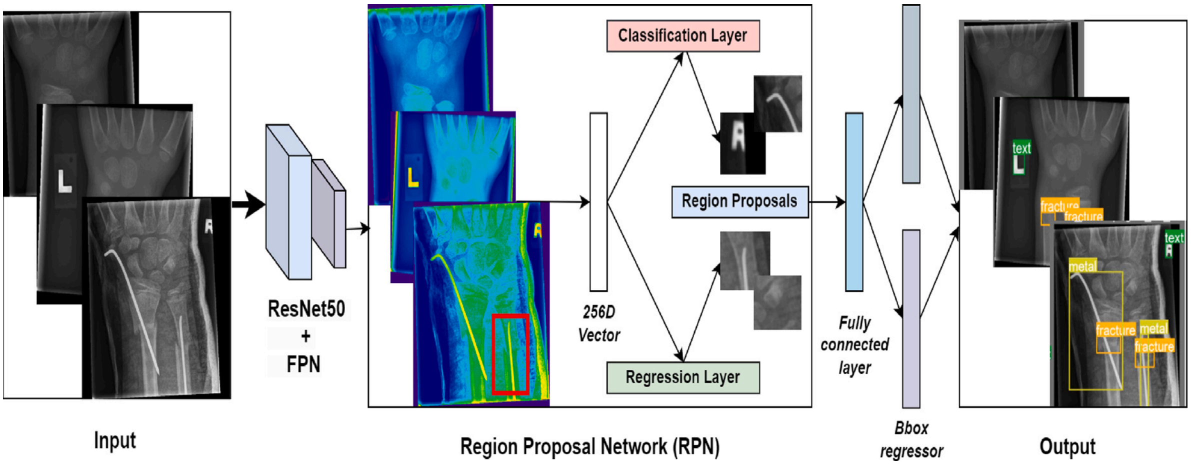


Fig. 3. Faster R-CNN Pipeline.

the conventional CNN, the selection of the number of layers and other hyperparameters was not intentional; rather, they represent commonly used configurations in the literature. We trained four variants of the DenseNet models: DenseNet121, DenseNet161, DenseNet169, and DenseNet201. Each model increases in complexity and depth from the previous one. The models are pre-trained on ImageNet data. Implementations for these variants are readily available in PyTorch. Finally, since YOLOv8 also comes with classification capabilities, initially, the YOLOv8 m model was pre-trained using chest data sourced from Kaggle [44]. Subsequently, we fine-tuned this pre-trained model on our primary dataset pertaining to wrist data. The aim was to compare the classification performance of YOLOv8 m with DenseNets and conventional CNNs in terms of sensitivity and accuracy. Our choice of using chest X-ray pretrained weights over ImageNet or COCO was influenced by the domain and content similarities between chest and wrist X-rays. Both share specific features inherent to medical imaging, such as X-ray beam artifacts and consistent grayscale patterns. In contrast, ImageNet or COCO comprises diverse, colored images from various non-medical contexts, which do not capture the nuances of medical imaging.

3.5. 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 $\alpha = 1 \times 10^{-2}$, 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.

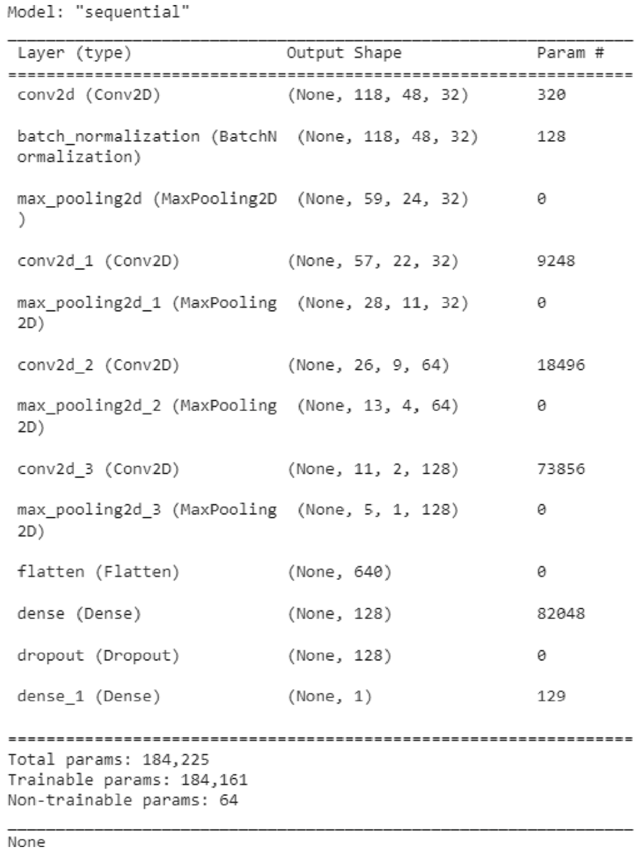
With Faster R-CNN, the only difference was the learning rate of $\alpha = 1 \times 10^{-3}$, momentum of 0.9 and weight decay of 5×10^{-4}. All other parameters were the same as YOLO variants. As with YOLO models, the selection of these parameters is not deliberate, they are the default settings.

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.

3.6. Evaluation metrics

3.6.1. mAP

To evaluate object detection, *Intersection over Union (IoU)* is commonly used to determine the accuracy of predicted object location. It is the ratio of the intersection of the predicted and ground truth bounding boxes to the union of these boxes. For a given image, let A be the set



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 118, 48, 32)       320

 batch_normalization (BatchN  (None, 118, 48, 32)      128
 ormalization)

 max_pooling2d (MaxPooling2D  (None, 59, 24, 32)       0
 )

 conv2d_1 (Conv2D)           (None, 57, 22, 32)        9248

 max_pooling2d_1 (MaxPooling  (None, 28, 11, 32)       0
 2D)

 conv2d_2 (Conv2D)           (None, 26, 9, 64)         18496

 max_pooling2d_2 (MaxPooling  (None, 13, 4, 64)        0
 2D)

 conv2d_3 (Conv2D)           (None, 11, 2, 128)        73856

 max_pooling2d_3 (MaxPooling  (None, 5, 1, 128)        0
 2D)

 flatten (Flatten)           (None, 640)               0

 dense (Dense)               (None, 128)               82048

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 1)                 129

=================================================================
Total params: 184,225
Trainable params: 184,161
Non-trainable params: 64
_________________________________________________________________
None
```

Fig. 4. Architecture summary of the conventional CNN binary classifier.

of predicted bounding boxes and B be the set of ground truth bounding boxes. 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 IoU > 0.5, the detection is considered true positive, else it is false positive. TP and FP can be computed using IoU to calculate 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 $IoU > 0.5$. Furthermore, it takes into account the balance between precision and recall and considers both the occurrence of false positives and false negatives.

3.6.2. 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:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6}$$

Initially, we evaluated the models using a default confidence threshold of 0.001. This default threshold will give us an initial understanding of the performance of each model in terms of fracture detection. Object detection models typically have lower confidence thresholds compared to classification models. This is because object detection involves both classifying objects and precisely localizing multiple objects within an image. Setting a lower threshold helps ensure that the model does not miss any potential objects. We then select the best-performing model and test its performance at higher thresholds (0.5, 0.7, 0.9).

3.7. Supplementary materials

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

- Source Code[1]
- Dataset Split[2]

4. Dataset

The dataset used in this study is called GRAZPEDWRI-DX *for machine learning* presented by the authors in [21] and is publicly made available to encourage computer vision research. The dataset contains pediatric wrist radiograph images in PNG format of 6091 patients (mean age 10.9 years, range 0.2 to 19 years; 2688 females, 3402 males, 1 unknown), treated at the Division of Paediatric Radiology, Department of Radiology, Medical University of Graz, Austria. It contains a total of 20,327 wrist images covering lateral and posteroanterior projections. The radiographs were acquired over a 10-year period from 2008 to 2018 and annotated between 2018 and 2020 by expert radiologists and medical students. The annotations were validated by three experienced 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. 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 has 9 objects, including periostealreaction, fracture, metal, pronatorsign, softtissue, boneanomaly, bonelesion, foreignbody, and text. The "text" object, typically present in most X-ray images, is used to identify the side of the body. However, some images might not contain any objects, including this text. Table 1 shows the number of X-ray images containing each object. Aside, from the object 'text', fracture is the most common object, followed by periostealreaction and metal. Boneanomaly, bonelesion, and foreignbody have the lowest occurrence.

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

[1] https://github.com/ammarlodhi255/pediatric-wrist-abnormality-detection

[2] https://studntnu-my.sharepoint.com/:u:/g/personal/ammaa_ntnu_no/EVUBFTvHlu9DgxcmmoTyFQoBGQRiGrKc5EzV1-b3YPnhyw?e=HNvu5j

Table 1
Class distribution.

Abnormality	Instances	Ratio
Boneanomaly	192	0.94%
Bonelesion	42	0.21%
Foreignbody	8	0.04%
Fracture	13 550	66.6%
Metal	708	3.48%
Periostealreaction	2235	11.0%
Pronatorsign	566	2.78%
Softtissue	439	2.16%
Text	20,274	99.74%

Table 2
Object occurrences.

Abnormality	Zero	One	Two	More	Total
Fracture	6777	9212	4137	201	13 550
Boneanomaly	20 135	42	24	126	192
Bonelesion	20 285	11	8	23	42
Foreignbody	20 319	0	0	8	8
Metal	19 620	347	219	141	707
Periostealreaction	18 092	1273	885	77	2235
Pronatorsign	19 761	456	71	39	566
Softtissue	19 888	221	82	136	439

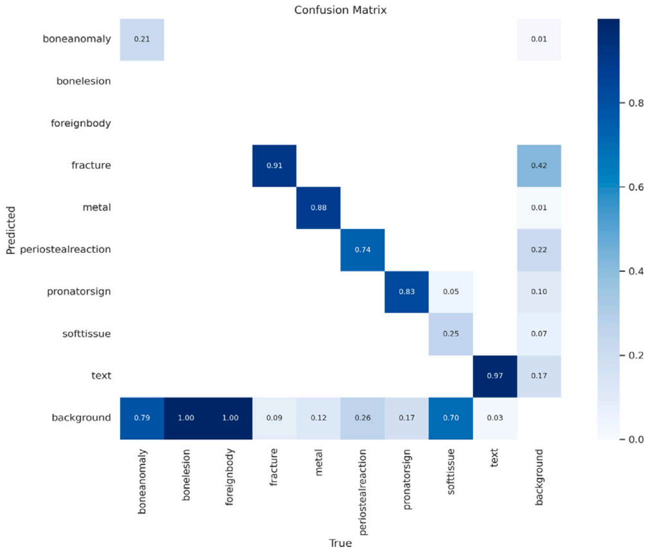


Fig. 5. Confusion Matrix (YOLOv8 m).

5. Results & discussion

This section presents a comprehensive analysis of the performance of all models utilized in this study. We start by conducting a detailed analysis of the different variants within each YOLO model. Subsequently, we identify the best-performing variant from each YOLO model based on the highest mAP score at an IoU threshold of 0.5 and recall (sensitivity) at a fixed confidence threshold of 0.001 across all classes and the fracture class. Finally, we compare the performance of the best-performing variants against each other and also against Faster R-CNN. The top-performing variant is further evaluated for sensitivity at higher confidence thresholds (0.5, 0.7, 0.9), we also look at the sensitivity vs. precision trade-off of the model.

The performance of YOLOv5 variants is presented in Table 3 illustrating their performance across all classes and also on the fracture class. The results reveal that the variants, "l", "x", and “s” achieved the highest mAP of 0.95 for fracture detection. However, the highest sensitivity of 0.91 was attained only by the variants “n” and “s” with variant “s” exhibiting superior performance in detecting fractures by excelling in both sensitivity and mAP. Notably, variant “s” also exhibits a precision of 0.89, indicating a favorable balance between precision and recall. Regarding overall performance across all classes, variant “x” obtained the highest mAP score of 0.69, while variant “s” achieved the highest sensitivity of 0.66. Increasing the complexity beyond variant “s” seems to decrease the sensitivity of the model.

Table 4 displays the performance of YOLOv6 variants. Among the variants, “n”, “s”, and “m” achieved the highest mAP of 0.94 for fracture detection, with “s” having the highest sensitivity of 0.89. However, we observed that increasing the model complexity beyond the “s” variant had a diminishing effect on its sensitivity in fracture detection. We found that the sensitivity improved by 0.03 when transitioning from the “n” variant to the “s” variant, suggesting that the “s” variant struck the optimal balance between complexity and performance for this specific task of fracture detection. When considering performance across all classes, the “m” variant exhibited the highest overall mAP of 0.64, along with a sensitivity of 0.83. The “s” variant followed closely behind with an mAP of 0.62 and a sensitivity of 0.82.

Table 5 presents the performance evaluation of YOLOv7 variants. The results demonstrate that the second variant of the YOLOv7 model achieved the highest mAP of 0.94 and a sensitivity of 0.91 for fracture detection. This variant also exhibits superior performance across all classes, attaining the highest mAP of 0.61 and a sensitivity of 0.54. Consistent with previous observations in YOLO5 and v6, increasing the model’s complexity beyond the second variant negatively impacts its sensitivity.

Table 6 illustrates the performance of YOLOv8 model variants for all classes, including the fracture class. The “m” variant achieved a mAP of 0.95 and the highest sensitivity of 0.92 in detecting fractures. Beyond this variant, increasing complexity appears to decrease fracture detection sensitivity while maintaining the same mAP score. For all classes, the variant “x” demonstrated the highest mAP of 0.77 and sensitivity of 0.64. Interestingly, here increasing the model complexity generally enhanced performance in terms of both mAP and especially sensitivity across all classes.

In the majority of YOLO models, excluding YOLOv8, a rise in model complexity corresponded to reduced sensitivity for fracture detection and across all classes. Several factors might account for this observation. The “s” variant appears to optimally balance model complexity and performance. Simpler variants like “n” may not capture intricate patterns adequately, while highly complex variants might overfit to training data. Additionally, the architecture and parameters of the “s” variant may be intrinsically better aligned for discerning features pertinent to wrist abnormalities in the dataset. However, in the latest version of YOLO, YOLOv8, the “m” variant appears to excel specifically in fracture detection. Adding further complexity beyond this variant might introduce redundant parameters or layers without significantly enhancing fracture detection capabilities. It is worth noting that increased complexity led to enhanced performance across all classes, the “x” variant’s superior performance suggests a more robust generalization capability. This could be due to its ability to effectively extract diverse and generalized features across various abnormality classes.

The experimental evaluation results using the Faster R-CNN model are presented in Table 7. The obtained sensitivity values for fracture detection and overall performance across all classes are 0.75 and 0.36, respectively. The results clearly demonstrate that all variants of the YOLO model significantly outperform Faster R-CNN. This conclusion is further supported by the higher mean mAP scores observed for every YOLO variant compared to Faster R-CNN, both in terms of fracture detection and overall performance across all classes. Moreover, the performance of YOLO across all classes suggests that YOLO is also capable of handling high-class imbalance compared to Faster R-CNN. These findings strongly suggest that the YOLO model, with its single-stage detection algorithm, is a more effective choice for this particular task.

Table 3
YOLOv5 results.

Model variant	Precision (All)	Sensitivity (All)	mAP@0.5 (All)	Precision (Fracture)	Sensitivity (Fracture)	mAP@0.5 (Fracture)
YOLOv5n	0.77	0.52	0.59	0.87	0.91	0.94
YOLOv5s	0.75	0.66	0.65	0.89	0.91	0.95
YOLOv5 m	0.80	0.62	0.69	0.91	0.90	0.94
YOLOv5l	0.76	0.61	0.68	0.92	0.90	0.95
YOLOv5x	0.73	0.64	0.69	0.91	0.90	0.95

Table 4
YOLOv6 results.

Model variant	Precision (All)	Sensitivity (All)	mAP@0.5 (All)	Precision (Fracture)	Sensitivity (Fracture)	mAP@0.5 (Fracture)
YOLOv6n	0.50	0.73	0.51	0.94	0.86	0.94
YOLOv6s	0.51	0.82	0.62	0.92	0.89	0.94
YOLOv6 m	0.59	0.83	0.64	0.94	0.87	0.94
YOLOv6l	0.60	0.80	0.64	0.94	0.87	0.93
YOLOv6l6	0.49	0.77	0.52	0.91	0.86	0.92

Table 5
YOLOv7 results.

Model variant	Precision (All)	Sensitivity (All)	mAP@0.5 (All)	Precision (Fracture)	Sensitivity (Fracture)	mAP@0.5 (Fracture)
YOLOv7-Tiny	0.59	0.52	0.50	0.79	0.91	0.93
YOLOv7	0.79	0.54	0.61	0.86	0.91	0.94
YOLOv7x	0.68	0.49	0.53	0.85	0.90	0.94
YOLOv7-W6	0.84	0.44	0.47	0.86	0.88	0.92
YOLOv7-E6	0.81	0.46	0.48	0.86	0.88	0.92
YOLOv7-D6	0.74	0.48	0.49	0.84	0.88	0.92
YOLOv7-E6E	0.69	0.50	0.47	0.85	0.87	0.90

Table 6
YOLOv8 results.

Model variant	Precision (All)	Sensitivity (All)	mAP@0.5 (All)	Precision (Fracture)	Sensitivity (Fracture)	mAP@0.5 (Fracture)
YOLOv8n	0.73	0.58	0.59	0.87	0.88	0.93
YOLOv8s	0.72	0.63	0.65	0.87	0.91	0.94
YOLOv8 m	0.60	0.60	0.56	0.84	0.92	0.95
YOLOv8l	0.74	0.60	0.62	0.92	0.90	0.95
YOLOv8x	0.79	0.64	0.77	0.91	0.89	0.95

Table 7
Faster R-CNN results.

Abnormality	Sensitivity	mAP@0.5
Fracture	0.64	0.75
All	0.36	0.36

Table 8
Classification report of binary classifiers.

Model	Acc	Class	P	R	F1
Conventional CNN	0.80	0	0.74	0.64	0.68
		1	0.83	0.89	0.86
DenseNet121	0.61	0	0.40	0.35	0.38
		1	0.70	0.74	0.72
DenseNet161	0.63	0	0.44	0.39	0.41
		1	0.71	0.76	0.73
DenseNet169	0.62	0	0.40	0.28	0.33
		1	0.69	0.79	0.74
DenseNet201	0.61	0	0.40	0.35	0.37
		1	0.70	0.74	0.72
Pretrained YOLOv8m	**0.92**	0	0.86	0.90	0.88
		1	**0.95**	**0.93**	**0.94**

We now move towards binary classification results. Table 8 displays the summary of the classification report for the binary classifiers. For clarification, class 0, represents no fracture. The table reveals that the pretrained YOLOv8 m achieved the highest fracture recall (sensitivity) of 0.93 and an overall accuracy of 0.92. Followed by conventional CNN, which achieved a fracture sensitivity of 0.89 and an overall accuracy of 0.80. The diminished accuracy observed in the DenseNet models may be attributed to the significant class imbalance present in the binary dataset. While DenseNets exhibit better performance in identifying the presence of a fracture, they falter in detecting its absence, likely due to the limited instances of this class compared to fractures. DenseNets, with their dense inter-layer connections, are designed to retain and reuse features throughout the network. In situations of dataset imbalance, the features predominantly reflect the majority class, potentially undermining the representation of the minority class. Furthermore, the inherent depth and intricacy of DenseNets elevate the risk of overfitting, particularly when dealing with underrepresented classes.

Fig. 6 provides an overview of the sensitivity scores obtained for fracture class as well as across all classes by all best-performing YOLO variants and Faster R-CNN. Moreover, it is clear from Table 9 that the variant “YOLOv8m” is the best-performing variant out of all the variants employed in this study when it comes to fracture detection. Interestingly, YOLOv6s outperformed all other variants when it comes to sensitivity across all classes. For binary classification, YOLOv8 m outperformed other classifiers achieving an impressive fracture sensitivity of 0.93. It should be noted that YOLOv8 m was pretrained on chest X-rays as mentioned earlier. The results presented in this study using the variant “YOLOv8m” represent a significant improvement upon the ones originally presented in [21] for the fracture class. In that paper, the model variant “YOLOv5m” trained on COCO weights achieved a mAP score of 0.93 and a sensitivity of 0.89 for fracture detection. In contrast, the results obtained in this study demonstrate a higher mAP score of 0.95 for fracture detection and a sensitivity of 0.92.

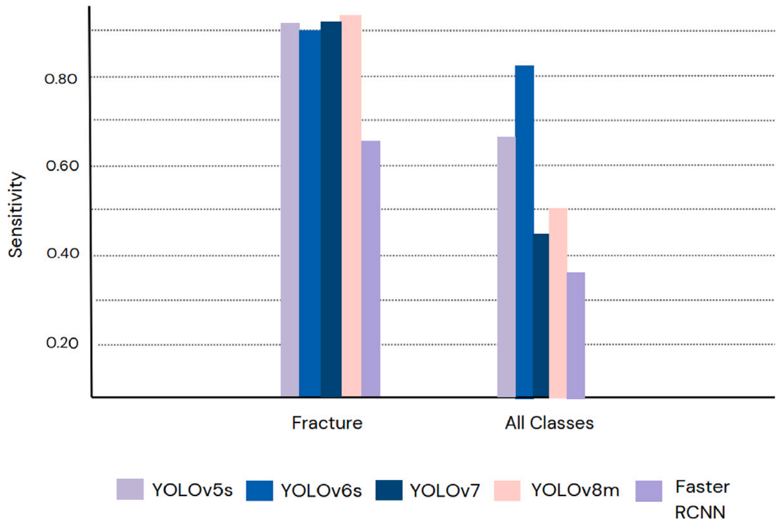


Fig. 6. Overview of sensitivity scores of best performing YOLO variants and Faster R-CNN.

Table 9
Sensitivity scores summary of detection models as well as binary classifiers.

Task	Model	Fracture	All
Detection	YOLOv5s	0.91	0.52
	YOLOv6s	0.89	**0.82**
	YOLOv7	0.91	0.54
	YOLOv8 m	**0.92**	0.60
	Faster R-CNN	0.64	0.36
Binary Classification	Conventional CNN	0.89	–
	DenseNet169	0.79	–
	Pretrained YOLOv8 m	**0.93**	–

Table 10
Threshold analysis for fracture detection (YOLOv8 m).

Confidence threshold	Sensitivity
0.001 (Default)	0.92
0.5	0.83
0.7	0.67
0.9	0.00

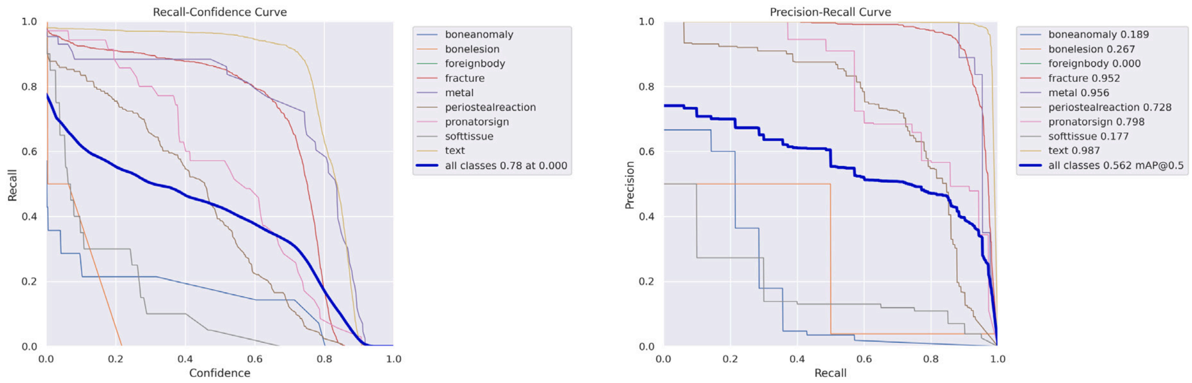
In applications where the cost of false positives is substantial, a model exhibiting high precision is generally preferred, whereas in situations where the cost of missed detections is significant then the model with high recall (sensitivity) is preferred. However, in the context of fracture detection, our research prioritizes high sensitivity due to the potentially severe consequences of false negatives when fractures are overlooked by the AI-based screening system. Achieving high sensitivity is crucial to ensure fractures are not missed and to facilitate prompt treatment for patients. Consequently, the selection of the optimal model is primarily based on sensitivity, while also considering a favorable balance of precision. A high precision score helps minimize the occurrence of unnecessary additional tests or procedures, ultimately reducing patient anxiety and healthcare costs. While maximizing sensitivity may lead to an increase in false positives, emphasizing precision may result in missed fractures. Thus, achieving the right equilibrium is essential to avoid unwarranted investigations while ensuring accurate fracture detection.

Ensuring patient safety by detecting all fractures without missing any is crucial for an AI algorithm. Therefore, a sensitivity close to 100% is desirable. We now conduct a threshold analysis for the "YOLOv8m" variant to examine the impact of increasing the confidence threshold (0.5, 0.7, 0.9) on the fracture detection sensitivity score, which is shown in Table 10. It is observed that as the confidence threshold is increased, the sensitivity of the algorithm decreases. This decrease is expected because object detection tasks, unlike classification tasks, perform both detection and classification hence requiring a low threshold to avoid missing any fractures. However, "YOLOv8m" performs reasonably well at a confidence threshold of 0.5, exhibiting a sensitivity of 0.83. Fig. 8 displays the bounding box estimations for fractures in an X-ray image from the best-performing models, including Faster R-CNN. Among these models, "YOLOv8m" accurately predicted the bounding boxes of fractures while avoiding false positive detections. We also evaluated the performance of the variant "YOLOv8m" on a challenging image containing multiple objects of interest overlapping each other, including 2 fractures, 3 periosteal reactions, 1 metal, and 1 text. The bounding box estimates for these objects are illustrated in Fig. 9. It can be seen that the model demonstrates an excellent job in detecting fractures even in the presence of other abnormalities that may conceal the fractures. However, it is important to acknowledge that the variant does not fully cover the bottom edges of the "metal" object and missed one "periosteal reaction" object. These instances highlight the ongoing necessity for continuous improvement and fine-tuning of algorithms to effectively handle complex cases like these. Additionally, it should be noted that the limited number of instances for classes other than fractures might have contributed to these observations.

Figs. 7(a), and 7(b) present Recall (sensitivity) versus Confidence and Precision versus Recall curves, respectively, for the variant "YOLOv8m" across all classes. These curves provide a visual representation of the model's performance and allow for a more thorough evaluation of its capabilities. The Recall versus Confidence curve illustrates the model's ability to correctly identify abnormalities at different confidence thresholds, while 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 5 is shown for the variant "YOLOv8m".

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, we evaluated a conventional binary convolutional CNN and DenseNets and compared their sensitivity to that of YOLO. Moreover, 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



(a) Recall vs. Confidence

(b) Precision vs. Recall

Fig. 7. YOLOv8 m Curves for Recall vs. Confidence and Precision vs. Recall.

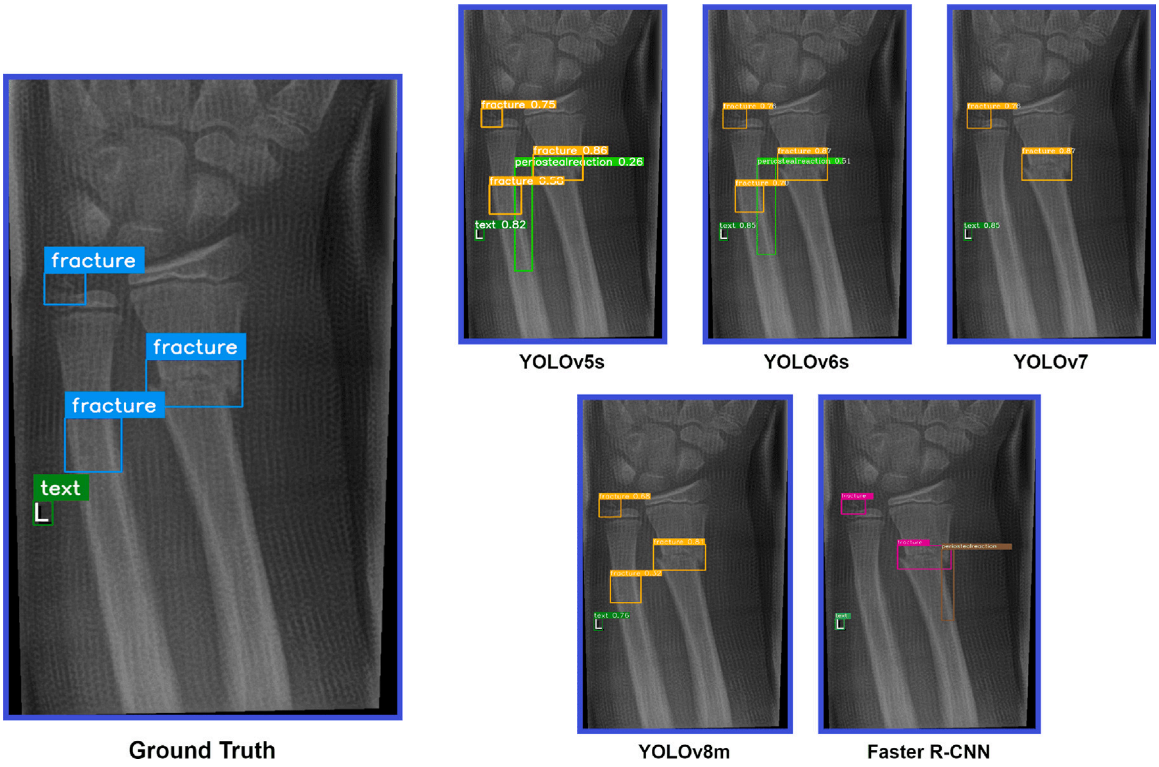


Fig. 8. Bounding box estimates for fractures in an X-ray image.

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, hlit also outperforms the DenseNets and conventional CNN in terms of sensitivity when it comes to fracture recognition.

Furthermore, an analysis of YOLO models revealed that the YOLOv8 variant "YOLOv8m" achieved the highest sensitivity score in fracture detection and the mAP score. On the other hand, "YOLOv6m" achieved the highest sensitivity across all classes. Meanwhile, "YOLOv8x" achieved the highest mAP across all classes. 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 dataset revealed that the performance of YOLO variants did not consistently improve with increasing complexity, with the exception of YOLOv8.

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, the YOLO models for detecting wrist abnormalities in clinical settings. These algorithms are faster than their two-stage counterparts, making them more practical for emergency situations 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.

Physicians can utilize single-stage detection systems as valuable screening tools to assist in the initial assessment of patients. These systems can offer additional information and insights to support physicians in their clinical decision-making. They can effectively highlight suspicious areas on medical images, draw attention to potential fractures, and provide precise measurements or annotations for accurate assessment. By leveraging the capabilities of fracture detection systems,

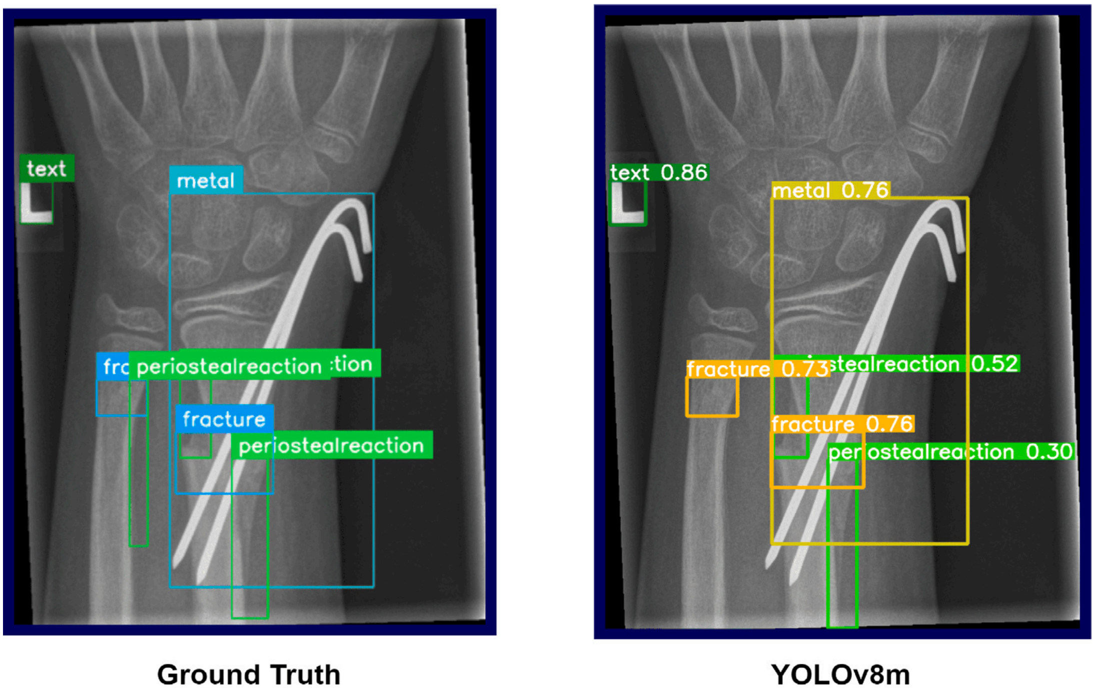


Fig. 9. Bounding box estimates by YOLOv8 m variant using a challenging image with overlapping objects.

physicians can enhance their ability to detect fractures and make informed decisions regarding patient care.

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 did not explore the entire hyperparameter space and finding the best hyperparameters for each YOLO model may improve the performance on the dataset. Computational limitations restricted the input resolution to 640 pixels, but higher resolutions could further improve performance. In future research, it would also be valuable to assess the model's robustness by evaluating its performance using cross-validation on multiple folds of the data. Moreover, the study used a dataset with class imbalance, so increasing their instances through augmentation or image generation could enhance performance. Additionally, our classification outcomes indicated that YOLO surpassed DenseNets in performance. This advantage may stem from the use of pre-trained weights based on chest data. Further investigations into the implications of transfer learning, especially in object detection, are warranted.

CRediT authorship contribution statement

Ammar Ahmed: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. **Ali Shariq Imran:** Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. **Abdul Manaf:** Data curation, Formal analysis, Methodology, Writing – original draft. **Zenun Kastrati:** Methodology, Supervision, Writing – review & editing. **Sher Muhammad Daudpota:** Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

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, Norway.

References

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

[2] P.H. Randsborg, et al., Fractures in children: epidemiology and activity-specific fracture rates, J. Bone Joint Surg. - Am. Vol. 95 (2013) e42.

[3] L.A. Landin, Epidemiology of children's fractures, J. Pediatric Orthopaedics B 6 (1997) 79–83.

[4] J. Cheng, W. Shen, Limb fracture pattern in different pediatric age groups: A study of 3350 children, J. Orthop. Trauma. 7 (1993) 15–22, http://dx.doi.org/10.1097/00005131-199302000-00004.

[5] P. Hallas, T. Ellingsen, Errors in fracture diagnoses in the emergency department: Characteristics of patients and diurnal variation, BMC Emerg. Med. 6 (4) (2006).

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

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

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

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

[10] T.K. Burki, Shortfall of consultant clinical radiologists in the UK, Lancet Oncol. 19 (e518) (2018).

[11] A. Rimmer, Radiologist shortage leaves patient care at risk, warns royal college, BMJ 359 (j4683) (2017).

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

[13] D.o. Rosman, Imaging in the land of 1000 hills: Rwanda radiology country report, 2015.

[14] R. Smith-Bindman, M. Kwan, E. Marlow, et al., Trends in use of medical imaging in US healthcare systems and in Ontario, Canada, 2000–2016, JAMA 322 (9) (2019) 843–856.

[15] A. Fotiadou, A. Patel, T. Morgan, A.H. Karantanas, Wrist injuries in young adults: The diagnostic impact of CT and MRI, Eur. J. Radiol. 77 (2011) 235–239, http://dx.doi.org/10.1016/j.ejrad.2010.06.029.

[16] J. Neubauer, et al., 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 (12) (2016) e0164859, http://dx.doi.org/10.1371/journal.pone.0164859.

[17] S.J. Adams, R.D.E. Henderson, X. Yi, P. Babyn, Artificial intelligence solutions for analysis of X-ray images, Canad. Assoc. Radiol. J. l'Association canadienne des radiologistes 846537120941671 (2020).

[18] L. Tanzi, et al., Hierarchical fracture classification of proximal femur X-Ray images using a multistage deep learning approach, Eur. J. Radiol. 133 (2020) 109373.

[19] J.W. Choi, et al., Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography, Investigat. Radiol. 55 (2020) 101–110.

[20] C. Lampert, M. Blaschko, T. Hofmann, Beyond sliding windows: Object localization by efficient subwindow search, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.

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

[22] ultralytics, YOLOv5 in PyTorch > ONNX > coreml > tflite, 2022, URL: https://github.com/ultralytics/yolov5.

[23] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, YOLOv6: A single-stage object detection framework for industrial applications, 2022, arXiv, arXiv:2209.02976.

[24] C. Wang, A. Bochkovskiy, H.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022, arXiv.org, URL: https://arxiv.org/abs/2207.02696. (Accessed: 30 December 2022).

[25] ultralytics, YOLOv8 in PyTorch > ONNX > CoreML > TFLite, 2023, GitHub repository, GitHub, https://github.com/ultralytics/ultralytics.

[26] E. Yahalomi, M. Chernofsky, M. Werman, Detection of distal radius fractures trained by a small set of X-ray images and faster R-CNN, 2018, arXiv.org, URL: https://arxiv.org/abs/1812.09025. (Accessed: 3 January 2023).

[27] Y.L. Thian, Y. Li, P. Jagmohan, D. Sia, V.E.Y. Chan, R.T. Tan, Convolutional neural networks for automated fracture detection and localization on wrist radiographs, Radiology: Artif. Intell. 1 (1) (2019) e180001.

[28] B. Guan, G. Zhang, J. Yao, X. Wang, M. Wang, Arm fracture detection in X-rays based on improved deep convolutional neural network, Comput. Electr. Eng. 81 (2020) 106530.

[29] M. Wang, J. Yao, G. Zhang, B. Guan, X. Wang, Y. Zhang, ParallelNet: Multiple backbone network for detection tasks on thigh bone fracture, Multimedia Syst. 27 (2021) 1091–1100.

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

[31] A. Raisuddin, E. Vaattovaara, M. Nevalainen, et al., Critical evaluation of deep neural networks for wrist fracture detection, Sci. Rep. 11 (2021) 6006, http://dx.doi.org/10.1038/s41598-021-85570-2.

[32] Y. Ma, Y. Luo, Bone fracture detection through the two-stage system of CrackSensitive convolutional neural network, Inform. Med. 236 (2021) 24–40.

[33] H.-Z. Wu, L.-F. Yan, X.-Q. Liu, Y.-Z. Yu, Z.-J. Geng, W.-J. Wu, C.-Q. Han, Y.-Q. Guo, B.-L. Gao, The feature ambiguity mitigate operator model helps improve bone fracture detection on X-ray radiograph, Sci. Rep. 11 (2021) 1589.

[34] L. Xue, W. Yan, P. Luo, X. Zhang, T. Chaikovska, K. Liu, W. Gao, K. Yang, Detection and localization of hand fractures based on GA_Faster R-CNN, Alex. Eng. J. 60 (2021) 4555–4562.

[35] F. Hardalaç, F. Uysal, O. Peker, M. Çiçeklidaˇ g, T. Tolunay, N. Tokgöz, U. Kutbay, B. Demirciler, F. Mert, Fracture detection in wrist X-ray images using deep learning-based object detection models, Sensors 22 (3) (2022) 1285, http://dx.doi.org/10.3390/s22031285.

[36] D. Joshi, T. Singh, A. Joshi, Deep learning-based localization and segmentation of wrist fractures on X-ray radiographs, Neural. Comput. Appl. 34 (2022) 19061–19077, http://dx.doi.org/10.1007/s00521-022-07510-z.

[37] G. Sha, J. Wu, B. Yu, Detection of spinal fracture lesions based on improved Yolov2, in: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA, 2020, pp. 235–238, http://dx.doi.org/10.1109/ICAICA50127.2020.9182582.

[38] G. Sha, B. Yu, J. Wu, Detection of spinal fracture lesions based on improved faster-RCNN, in: 2020 IEEE International Conference on Artificial Intelligence and Information Systems, ICAIIS, 2020, pp. 29–32, http://dx.doi.org/10.1109/ICAIIS49377.2020.9194863.

[39] F. Hrži'c, et al., Fracture recognition in paediatric wrist radiographs: An object detection approach, Mathematics 10 (16) (2022) 2939, http://dx.doi.org/10.3390/math10162939.

[40] R. Lindsey, A. Daluiski, S. Chopra, A. Lachapelle, M. Mozer, S. Sicular, D. Hanel, M. Gardner, A. Gupta, R. Hotchkiss, et al., Deep neural network improves fracture detection by clinicians, Proc. Natl. Acad. Sci. 115 (47) (2018) 11591–11596, http://dx.doi.org/10.1073/pnas.1807792115, arXiv:https://www.pnas.org/content/115/47/11591.full.pdf, URL: https://www.pnas.org/content/115/47/11591.

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

[42] A. Bochkovskiy, C. Wang, H. Liao, YOLOv4: Optimal speed and accuracy of object detection, 2020, arXiv.org.

[43] C. Wang, H. Liao, Y. Wu, P. Chen, J. Hsieh, I. Yeh, 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, 2020, pp. 390–391.

[44] Kaggle, Chest X-Ray images (pneumonia), 2023, Kaggle, URL: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia, https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia.

Ammar Ahmed is a final-year undergraduate student pursuing a degree in computer science at Sukkur IBA University (SIBAU). His research interests include deep learning technologies and their applications, with a particular interest in the intersection of computer vision and natural language processing. He has been recognized for his academic achievements through the receipt of three scholarships based on merit, including the Institutional Scholarship and two Governmental Scholarships.

Ali Shariq Imran received a master's degree in software engineering and computing from the National University of Science and Technology (NUST), Pakistan, in 2008 and a Ph.D. in computer science from the University of Oslo (UiO), Norway, in 2013. He is associated as an Associate Professor with the Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), Norway. With over 15 years of teaching and research experience, he devised innovative ways to design effective multimedia learning objects and integrate the teaching-research nexus frameworks at the graduate level. He served as a commission member of the Ministry of Education of Macedonia in setting up Mother Theresa University in Skjope. He leads a capacity-building project called CONNECT (https://norpart-connect.com) funded by the Higher Education Commission of Norway, DIKU, under the NORPART scheme as a coordinator and three Erasmus+ KA2 projects (PhDICTKES (https://phdictkes.eu), RAPID, and TKAEDiT) as a project manager at NTNU, along with an Excited mini-project funded by NTNU. Dr. Ali is also leading a research group on Deep NLP (http://deep-nlp.net) and specializes in applied deep learning research to address various multi-modality media analysis application areas for audio–visual and text processing. He has co-authored over 100 peer-reviewed journals and conference publications and has served as an editor and reviewer for many reputed journals. He is a member of the Intelligent Systems and Analytics research group at NTNU and an IEEE/ACM Member.

Abdul Manaf is a final-year undergraduate student pursuing a degree in computer science at Sukkur IBA University (SIBAU) and currently holds a scholarship at the university. He has a keen interest in the fields of Deep Learning and Computer Vision, with a special focus on Generative Models such as GAN for image generation and for text processing. With a passion for research, He has been actively engaged in exploring the latest advancements in the field of Artificial Intelligence.

Zenun Kastrati received a master's degree in computer science through the EU TEMPUS Programme developed and implemented jointly by the University of Pristina, Kosovo, the Universite de La Rochelle, France, and the Institute of Technology Carlow, Ireland, and the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Norway, in 2018. He works as an Associate Professor at the Department of Informatics at Linnaeus University, Sweden. His research interests lie in the field of artificial intelligence with a special focus on NLP, machine/deep learning, and sentiment analysis. He is the author of more than 60 peer-reviewed journals and conferences and has served as a reviewer for many reputed journals over the years.

Sher Muhammad Daudpota received his Masters and Ph.D. degrees from Asian Institute of Technology, Thailand in the year 2008 and 2012, respectively. His research areas include deep learning, natural language processing, video and signal processing. He is author of more than 35 peer-reviewed journal and conference publications. Presently he is serving as a Professor of Computer Science at Sukkur IBA University, Pakistan. Alongside his Computer Science contribution, he is also a Quality Assurance expert in higher education. He has reviewed more than 50 universities in Pakistan for quality assurance on behalf of Higher Education Commission in the role of educational quality reviewer.