ORIGINAL ARTICLE

Improving pediatric trauma care: an automated system for wrist trauma detection using GELAN

Promit Basak[1] · Adam Mushtak[2] · Mohamed Ouda[3] · Sadia Farhana Nobi[4] · Anwarul Hasan[5] · Muhammad E. H. Chowdhury[4]

Received: 28 August 2024 / Accepted: 23 July 2025 / Published online: 12 September 2025

Abstract

Trauma is a major cause of disability among children, requiring swift and accurate diagnosis for effective treatment. This paper introduces an automated method that uses deep learning to detect and categorize fractures in children using X-ray images. The system makes use of the GRAZPEDWRI-DX dataset, which consists of 20,327 annotated X-ray images of pediatric wrist fractures. Our architecture, which is built upon the generalized efficient layer aggregation network (GELAN), effectively tackles the issues of class imbalance and image resolution. As a result, it achieves state-of-the-art performance in both trauma and severity detection. Our proposed framework surpassed the most advanced techniques, showcasing exceptional precision and effectiveness, achieving a mean average precision (mAP50) score of 74.1%, 95%, and 85.5% for Task A (trauma detection), Task B (fracture detection), and Task C (fracture severity detection), respectively. The results of our study highlight the capacity of deep learning to improve the diagnosis of pediatric trauma, decrease the burden on radiologists, and boost patient outcomes.

Keywords Pediatric trauma · Fracture detection · Deep learning · X-ray · Medical imaging · Generalized efficient layer aggregation network (GELAN)

1 Introduction

Trauma continues to be one of the main causes of death and disability among children, and their demands for specialized trauma care are similar to those of adults, according to population-based statistics [1–4]. In addition to mortality and morbidity, it is also important to consider the long-term effects, such as physiological and economic impacts [5–7]. Mostly, young children aged between 7 and 14 fall victim to different types of injuries predominantly in the forearm and wrist area [8, 9]. In most cases of accidental trauma, a timely as well as accurate diagnosis is crucial for effective treatment and recovery. Nevertheless, the unavailability of appropriate equipment and expert doctors hinders the way to quick recovery and survival. This study addresses this issue to expedite the diagnosis process with automatic and accurate detection of pediatric trauma from X-ray images.

Due to anatomical and physiological differences between children and adults, there are specific traumatic effects that are unique to each age group [10]. Children's skeletal structure is more susceptible to fractures due to its weaker nature, but less prone to the spreading of fractures. This is attributed to the higher proportion of cartilage and collagen in children's bones, which makes them more elastic [11, 12]. Hence, in order to achieve the best possible interpretation of imaging, it is imperative to possess a comprehensive comprehension of the pertinent anatomy and the age-appropriate visual characteristics of the wrist and distal forearm. For example,

https://doi.org/10.1007/s00521-025-11539-1

incomplete fractures, such as bowing, torus, and greenstick are almost unique to children as shown in Fig. 1 [13]. Thus, clinical considerations for pediatric trauma include the need for specialized knowledge for accurate diagnosis and treatment and the importance of minimizing radiation exposure due to the sensitivity of developing tissues. This difference between adult and child bone structure makes it impossible to apply the adult fracture knowledge to the cases of children and requires the development of standalone algorithms and approaches.

X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound are commonly utilized in pediatric trauma diagnosis. MRI offers detailed images of soft tissues and bones and is considered the most reliable, specific, and optimal modality to detect complex trauma and fractures [10, 14–16]. However, MRI is limited by its availability, higher cost, longer scan times, and inappropriate in the presence of metal implants or pacemakers [17]. CT (Computed Tomography) provides comprehensive cross-sectional images, superior in detecting complex fractures, but involves higher radiation exposure and is slow as well as less accessible compared to X-rays. Ultrasound is useful for soft tissue evaluation and certain fracture types but is operator-dependent and less effective for visualizing bones [18]. Although other imaging modalities have their benefits, X-ray is still the most frequently utilized first-line imaging modality for trauma diagnosis. This is due to its ability to quickly and inexpensively assess the conditions, which is crucial for emergency diagnosis and treatment [19]. An inherent drawback of this modality is its inferior resolution and contrast compared to CT and MRI, which poses challenges in detecting minute fractures and requires the expertise of skilled and proficient radiologists. However, there is a shortage of radiologists not only in low- and middle-income countries (LMIC) but also in affluent countries like the UK [20]. Consequently, the need for sophisticated diagnostic techniques arises to assist in these circumstances.

Deep learning (DL) has been widely utilized in musculoskeletal radiology in recent years. DL applications have been used to analyze radiographs of the hip [21–23], ankle [24], humerus [25], and wrist [26–30] for automatic fracture detection in the realm of medical imaging. Regrettably, there is currently no research that has confirmed the accuracy of the methods used to diagnose complex fractures. The objective of this study is to develop a deep learning-based automated system for the detection and classification of trauma and fractures from pediatric X-ray images. Utilizing the novel GELAN architecture, we want to address the clinical gaps in pediatric trauma care including enhancing diagnostic accuracy in fracture and trauma detection, reducing the time required for intervention, assessing fracture severity early, and providing an alternative system in the absence of a specialist doctor. However, this study not only aims to improve diagnosis decision-making but also aligns with broader healthcare goals by optimizing clinical workflows and reducing the burden on radiologists. By automating the detection and classification of pediatric wrist fractures, our system can help mitigate the shortage of radiology experts in resource-limited settings and remote areas. The main contributions of this paper are stated below:

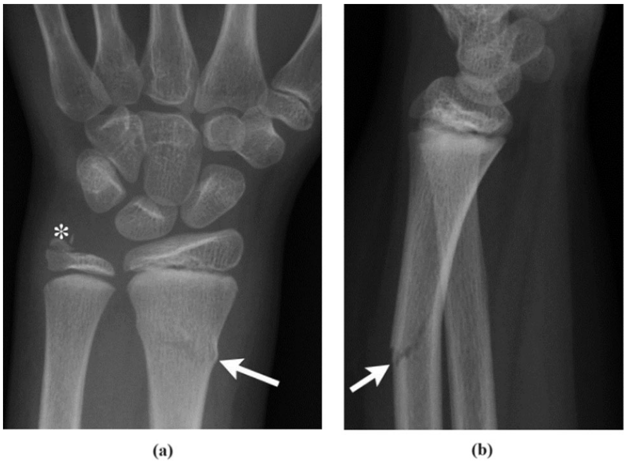


Fig. 1 X-rays of **a** buckle and **b** greenstick fractures in children [10]

- In this research, we address critical clinical gaps in pediatric trauma care by providing a robust tool for the automatic detection and localization of pediatric fractures. This tool can revolutionize pediatric trauma care and rehabilitation by ensuring fast and accurate diagnostics without needing a human expert.
- We propose a novel GELAN-E model that demonstrates superior performance compared to state-of-the-art methods, achieving high mean average precision (mAP50) scores.
- With rigorous preprocessing and augmentation process, it improves the process of automated trauma recognition (Task A), fracture detection (Task B), and severity identification (Task C).
- To the best of our knowledge, this is the first study to detect the severity of child fractures in the form of complete and incomplete fractures.

The article is segmented into five distinct sections. This section offers a concise elucidation of the study's inspiration, as well as the difficulties inherent in trauma detection from X-ray images. In Sect. 2, we will examine similar works and their respective contributions and limits. In the following section, we introduce a conceptual framework that forms the basis of our suggested technique. The findings of this investigation are succinctly outlined in Sect. 4. Section 5 encompasses the concluding remarks and aspirations for the future.

2 Related works

Fracture and trauma detection from X-ray are not a very new concept. Earlier in the previous decade, research in this domain mainly focused on signal processing and mathematical techniques to detect trauma and abnormalities from X-ray images. Liang et al. [31] presented a morphological technique for the detection of fractures in tibia bones. Prior to segmentation, the original image is dynamically partitioned into many intervals to facilitate the identification of the smallest interval containing the target. The smaller sections are subsequently thresholded using the Otsu approach in an automated manner. Smith et al. [32] proposed a multi-stage pelvic ring fracture detection system based on discrete wavelet transform (DWT). The pelvic ring was separated into windows before DWT was applied for both the denoising and the texture detection. The bone boundary is highlighted by reconstructing a picture using the chosen wavelet coefficient. Next, morphological operations are performed on the binary image and the boundaries of the ring are delineated by examining the surrounding pixels in an 8-neighborhood configuration for each edge pixel. This process generates a matrix of pixel positions. The presence of fractures was then detected by the discontinuity of boundaries. In another study [33], this task was converted to a bone shape identification problem. The algorithm performs integral projection on each pre-processed bone region in the X-ray image and then combines the resulting projection curves. Subsequently, the location of the fractured area is evaluated by analyzing the differences in sub-section curves. However, these techniques are restricted by their failure to detect fractures effectively and their sensitivity to image artifacts. Their high processing cost also hinders how to apply them in real-world circumstances.

With the advent of deep learning, there has been a change of dynamics in different fields and bone fracture detection is not an exception here. For the last few years, researchers have been working on applying deep learning techniques to both open-source and clinical bone image datasets collected from different medical devices. Yadav et al. [34] used a deep neural network model to categorize bones as either healthy or fractured. Data augmentation methods were employed to mitigate overfitting resulting from the limited dataset. The model utilized a four-layer convolutional neural network with a fully connected layer as the classifier. The proposed model attained a classification accuracy of 92.44% using a fivefold cross-validation technique. Nevertheless, the result was reported only on the validation test, and additional verification on a more extensive dataset is required to validate its performance.

Jones et al.[35] proposed an ensemble approach to predict whether the X-ray image contains a fracture. They used 10 variants of dilated residual networks [36], specifically designed to identify fractures from radiographs. Each network exhibited differences in architecture and output structure, although they all utilized the dilated residual network as their foundation. Each network was individually fine-tuned using a training set to predict the likelihood of a radiograph containing a fracture accurately. The ensemble output was derived by calculating the mean of the individual network outputs. However, the training dataset comprised 16 anatomical regions, with certain regions having a higher representation making it skewed.

Raisuddin et al. [30] worked on the localization of fractures from X-ray images. They utilized three datasets: one for training and the other two for validation and testing. At first, they trained to identify three anatomical wrist landmarks using the KNEEL method [37, 38] and then cropped the Region Of Interest (ROI) from the X-rays using those landmarks. The fracture detection block consisted of SeresNet50 models [39] which were ensembled later to get the accuracy of 95% and 82% on two test sets accordingly. The difference in accuracy of the two datasets raises a question of generalization and performance on challenging fractures.

The objective of the study by Hardalaç et al. [40] was to identify regions of wrist fractures in X-ray images by employing deep learning-based object detection models with various backbone networks. After a thorough preprocessing using CLAHE [41], the study utilized 26 models into five different ensemble models using the weighted bounding box fusion method to create a distinctive detection system. Their proposed pipeline achieved an average precision of 0.864, and an average recall of 0.33 surpassing the performance of the individual models. Nevertheless, the study exclusively examined fractures, and the ultimate models were intricate. Additionally, they did not mention the results in more widely used metrics such as mAP50 making it difficult to compare with other state-of-the-art studies.

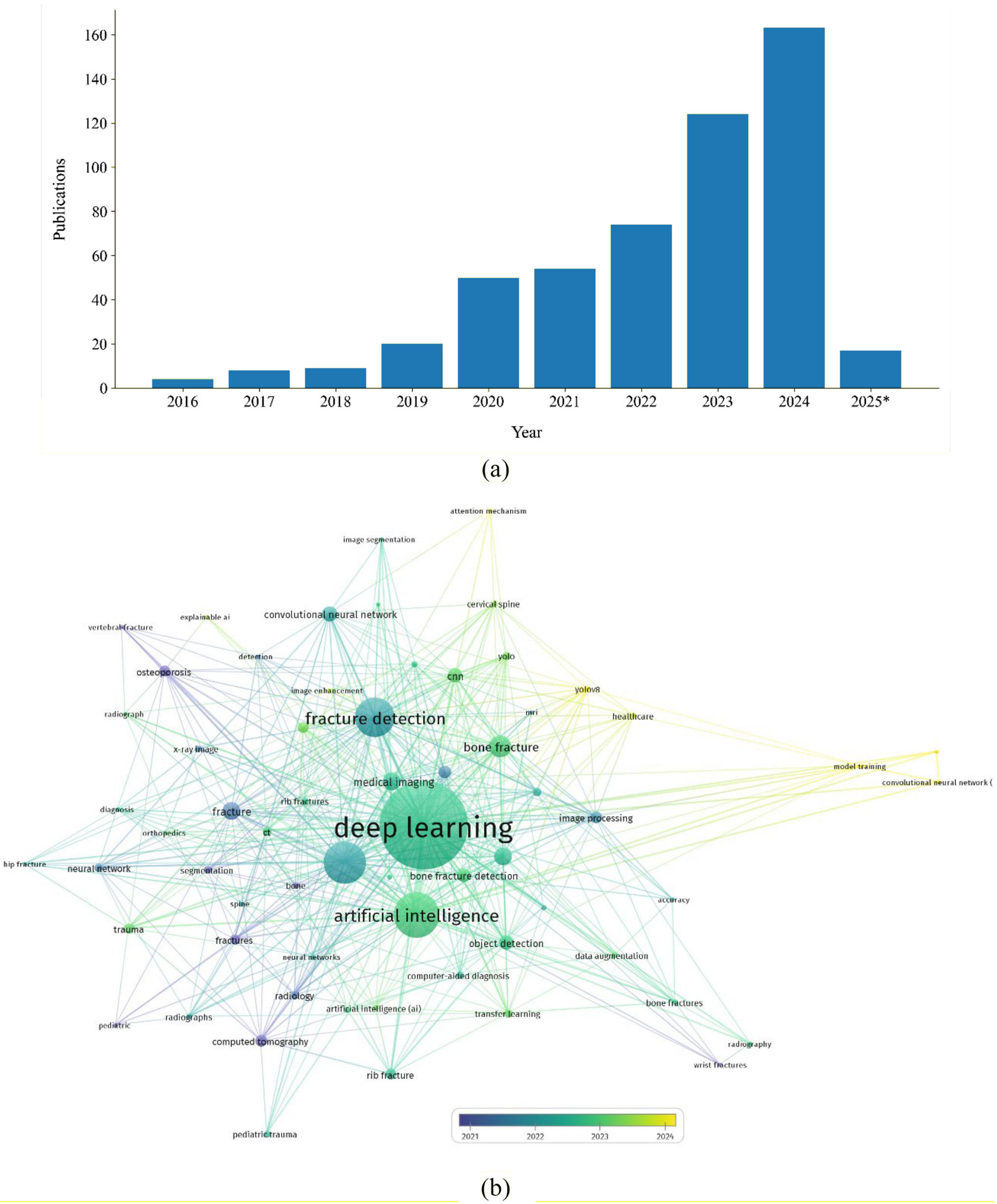
Ju et al. [42] proposed a rather straightforward approach to address this problem. Contrast and brightness adjustments were used to preprocess the images before they were applied to the YOLO v8 medium (YOLOv8m) network to detect fractures. They achieved mAP50 and mAP50-95 scores of 0.621 and 0.403, respectively, for the detection task. There was plenty of room to improve the performance and tune the architecture.

DeepLOC was proposed by Dibo et al. [43] for the localization and classification of wrist X-ray images. The main contribution of this paper was the introduction of GAM attention blocks between each detection head of YOLOv7 [44]. They used this DeepLOC model to detect four types of traumas including fracture, periosteal reaction, foreign body, and bone lesion, and achieved mAP50 of 0.654 and mAP50-95 of 0.398.

There are several other studies for classification [24, 25, 29, 45, 46], detection [47–51], and localization [26, 52, 53] of fractures and traumas from X-ray images and some of them achieved impressive performances through the novel methods. To understand the research landscape in this direction, we conducted a bibliometric analysis of related publications over the last ten years. This analysis helps identify trends, key areas of focus, and existing gaps, allowing us to position our study within the broader scientific context.

We used VOSviewer and Python to analyze publications retrieved from Scopus using keywords such as “pediatric trauma,” “fracture detection”, “deep learning,” and “medical imaging.” We found a total of 523 such publications. The results show a significant increase in research activity in this field, particularly in the last 5 years. This surge reflects the growing interest in applying machine learning to medical imaging for automated trauma diagnosis. We also analyzed the keywords of the publications in the last 5 years and identified a few major gaps in this research area. One of the most notable gaps is the lack of studies focusing on fracture severity analysis which is crucial for treatment planning. While many studies address fracture detection, few attempt to classify the severity of fractures. Additionally, there is limited research addressing class imbalance issues in pediatric trauma datasets, which can impact model performance in real-world clinical settings. The bibliometric analysis is shown graphically in Fig. 2.

The prior works in fracture detection have largely focused on binary classification (fracture vs. no fracture) or detection tasks using conventional architectures like ResNet, YOLOv4, and Faster R-CNN. While these approaches have demonstrated effectiveness, they lack scalability in handling multiple trauma indicators and do not address fracture severity. Some recent studies have attempted multi-class trauma detection but suffered from class imbalance and suboptimal localization performance.



(a)

(b)

Fig. 2 Bibliometric analysis of the publications in the last 10 years, **a** Publication counts on the relevant topic by year (*up to January 2025), **b** Co-occurrence analysis of the author keywords in the last 5 years

Unlike prior studies, our approach introduces a comprehensive framework that not only detects fractures but also classifies their severity—a critical yet underexplored aspect of pediatric trauma assessment. Our approach distinguishes itself by:

- A more advanced and flexible deep learning model that optimizes gradient propagation paths, ensuring improved accuracy in trauma detection and classification.
- Implementing a novel weighted augmentation strategy to address class imbalance and enhance the detection of rare trauma types.
- Unlike prior works, our study goes beyond simple fracture detection by distinguishing between complete and incomplete fractures.
- The chosen dataset, GRAZPEDWRI-DX, is one of the most comprehensive pediatric wrist trauma datasets that allows robust model training and validation across diverse trauma cases.

3 Methodology

In this section, we will begin with the dataset description, show a bird’s eye view of our proposed system, and finally, explain the intricate details of it.

3.1 Dataset

For this study, we have used the GRAZPEDWRI-DX dataset [54]. This is one of the few publicly available datasets that contain annotated X-ray images of the children’s wrists. It comprises 20,327 pediatric wrist radiographs from 6091 patients taken from the Department for Pediatric Surgery of the University Hospital Graz, Austria between 2008 and 2018. Out of 6091 patients, 2688 are female, 3402 are male and one is unknown with an age range of 0.2 and 19 years. The X-ray images were initially retrieved as digital imaging and communications in medicine (DICOM) images [55] and later converted to a more common 16-bit portable network graphics (PNG) image for easier access. The images are annotated in two ways: image-level tags containing image attributes and object-level annotations identifying trauma and other indications. A detailed view of different image tags is given in Table 1 and image annotations in Table 2.

One of the most useful things in this dataset is the metadata provided with each image. For example, AO classification codes may be used to further categorize fractures into different subtypes. However, the imbalance between the majority and minority classes in the annotations is quite significant. Particularly, foreign body, bone

Table 1 Image tags in the GRAZPREDWRI-DX dataset

Image tag	Count	Percentage (%)
AO classification code for fractures	14,158	69.65
Presence of cast/plaster	5,776	28.42
Whether the fracture labels or tags uncertain	537	2.64
Initial trauma presentation	10,861	53.43
Presence of metal implants	708	3.48
Signs of osteopenia	2,473	12.17
Anteroposterior projection view	10,086	49.62
Lateral projection view	10,148	49.92
Oblique projection view	93	0.46
Left side	11,135	54.78
Right side	9,192	45.22
Total	20,327	

Table 2 Image annotations in the GRAZPREDWRI-DX dataset

Image annotation	Annotation type	Object count	Present in percent of images (%)
Fracture	Box	18,090	66.66
Axis of the bone	Line	20,327	100
Bone anomaly	Box	276	0.94
Bone lesion	Polygon	45	0.21
Foreign body	Box	8	0.04
Metal	Box	819	3.48
Periosteal reaction	Polygon	3453	11
Pronator sign	Box	567	2.78
Soft tissue	Box	464	2.16
Text	Box	23,722	99.74
Total		67,771	

lesions, and bone anomalies are present in less than 1% of the images. The other trauma classes are not also comparable with the majority ‘fracture’ class considering the object quantity. We have mentioned how this huge class imbalance issue is addressed in Sect. 3.3. Sample images including the annotations are shown in Fig. 3.

3.2 Overview

A simplified overview of our proposed framework is given in Fig. 4.

As stated in Sect. 3.1, the dataset exhibits a significant imbalance, which necessitated the exclusion of certain classes from this analysis. Therefore, it was necessary for us to analyze the raw data during the ‘Feature and label selection’ phase. After consulting with radiologists, we eliminated the classes ‘bone anomaly’, ‘bone lesion’, and ‘foreign body’ due to their limited instances. Similarly, we excluded ‘metal’, ‘axis’, and ‘text’ since they did not have sufficient diagnostic significance. In this stage, we additionally transformed the bounding boxes to meet the requirements of the algorithm. X-ray images typically exhibit reduced resolution and contrast, demanding a considerable amount of care and effort to address during the ‘Image Preprocessing’ stage. We have introduced a mathematical approach to address the problem of class imbalance in detection data referred to as ‘Weighted augmentation’. The ‘Data adaptor’ block is responsible for modifying the data to suit various tasks and splitting the data for training purposes. Finally, we employed the GELAN model [56] to effectively identify various forms of trauma from the processed X-ray images.

3.3 Preprocessing and augmentation

X-ray radiographs usually have a lower resolution and contrast compared to other modern imaging modalities which may even further deteriorate due to other factors such as focal spot, presence of motion, and physical characteristics of the system [57]. We tried to enhance the image qualities with the help of several image processing techniques. At first, we used contrast limited adaptive histogram equalization (CLAHE) [41] to equalize the histogram of the image for better contrast resolution. Later, we applied adaptive gamma correction for brightness improvement using the adaptive gamma correction with weighting distribution (AGCWD) method proposed by Huang et al. [58]

One of the major challenges in medical imaging datasets is the imbalance between common and rare conditions, which can lead to biased models. We also faced this issue in our dataset. Even after excluding some minority classes, the class imbalance was very high. As each image may contain more than one object, augmenting a minority class uniformly will also increase the majority class failing the whole point of augmentation. To address this issue, we introduce a weighted augmentation strategy that dynamically adjusts augmentation rates based on class frequency. Unlike standard oversampling techniques, which may introduce redundancy, our

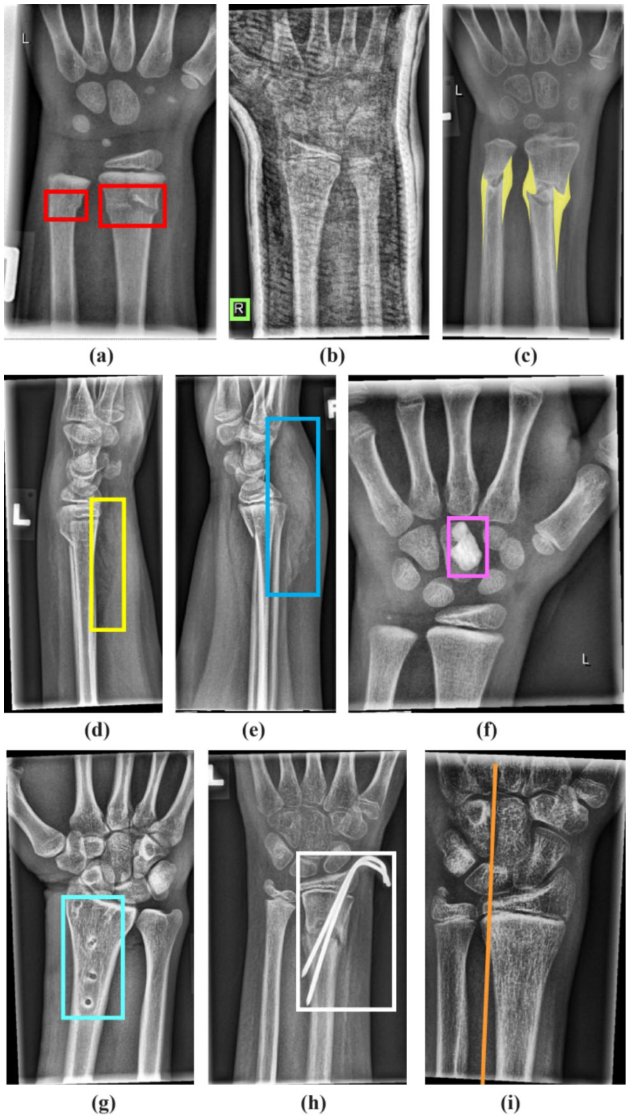


Fig. 3 Sample X-ray images from the dataset. Annotations are **a** fracture, **b** text, **c** periosteal reaction, **d** pronator quadratus sign, **e** soft tissue swelling, **f** foreign body, **g** bone anomaly, **h** metal, and **i** axis line of the bone

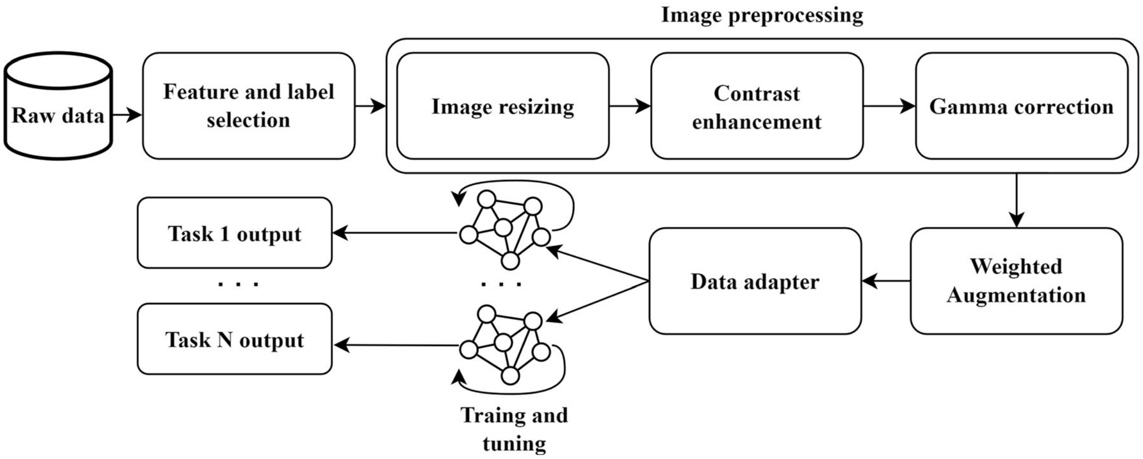


Fig. 4 An overview of the proposed framework

approach ensures that minority trauma classes such as periosteal reaction and pronator sign are adequately represented during training and testing without excessively inflating the dominant fracture class. This improves the model's ability to generalize across all trauma types, leading to more reliable detection and classification. According to this strategy, the number of augmentations depends on the number of objects of each class present in the image and the predefined weights for each class. If we have M classes with the corresponding number of objects in a particular image $N_1, N_2, N_3, \ldots, N_M$ and predefined weights $W_1, W_2, W_3, \ldots, W_M$. The number of augmentations for that corresponding image will be:

$$N_{\text{aug}} = \frac{\sum_{i=1}^{M} N_i \times W_i}{\sum_{i=1}^{M} N_i} \tag{1}$$

For the successful application of this strategy, a weight inversely proportional to the class frequency is required. In this work, we used weights for fracture, periosteal reaction, pronator sign, and soft tissue to be 1, 4, 10, and 10 accordingly.

3.4 Efficient Layer aggregation network (ELAN)

The design of deep neural networks (DNNs) has become a focal point in the quest for efficient, accurate, and low-cost solutions for various tasks. Most research adheres to the conventional understanding of deep networks, which involves extracting low-level features from shallow layers and high-level features from deep layers. Based on these concepts, one may utilize them to create neural network structures that efficiently integrate various levels of information in the data route (feed-forward path). ELAN [59] proposes a different way to think about this approach. In fact, what features a layer can learn is controlled by the objective or loss function of the layer. According to Fig. 5a, the bottom-up path has feature representations from low-level to high-level, but the top-down path represents all high-level features when guided by a loss function. Similarly, in Fig. 5b, all the layers represent high-level features. Hence, the features that the weight learns primarily depend on the type of information we provide for training it, rather than the combination of input layers.

ELAN is designed using gradient path analysis to optimize gradient propagation paths, enhancing learning efficiency and model performance. ELAN incorporates a unique layer aggregation mechanism that balances the shortest and longest gradient paths through each layer. This mechanism ensures efficient gradient flow by employing a stacking strategy within computational blocks, avoiding the excessive use of transition layers that can elongate gradient paths and hinder training. The design integrates principles from VoVNet [61] and Cross Stage Partial Network (CSPNet) [62] to further optimize gradient propagation. In CSPNET, the input feature map is segmented into two distinct parts. The first part passes into a computational module such as ResNet. The other one passes the full stage without any operation and is subsequently combined with the processed initial part. This

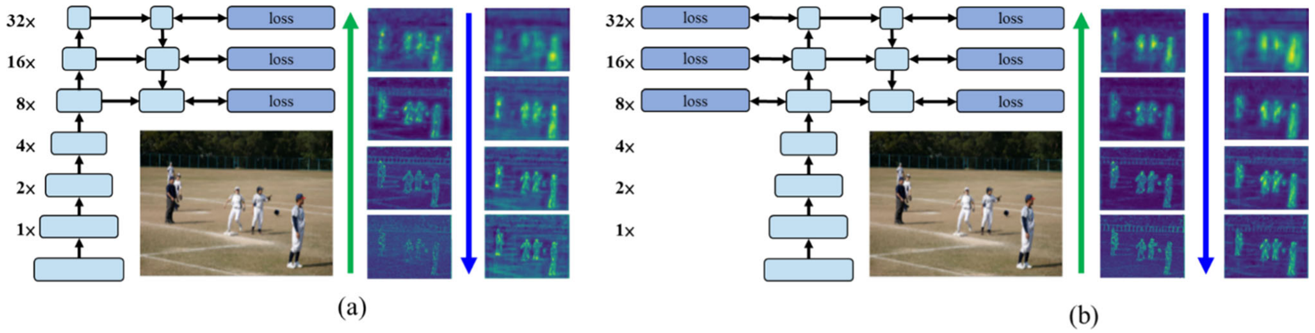
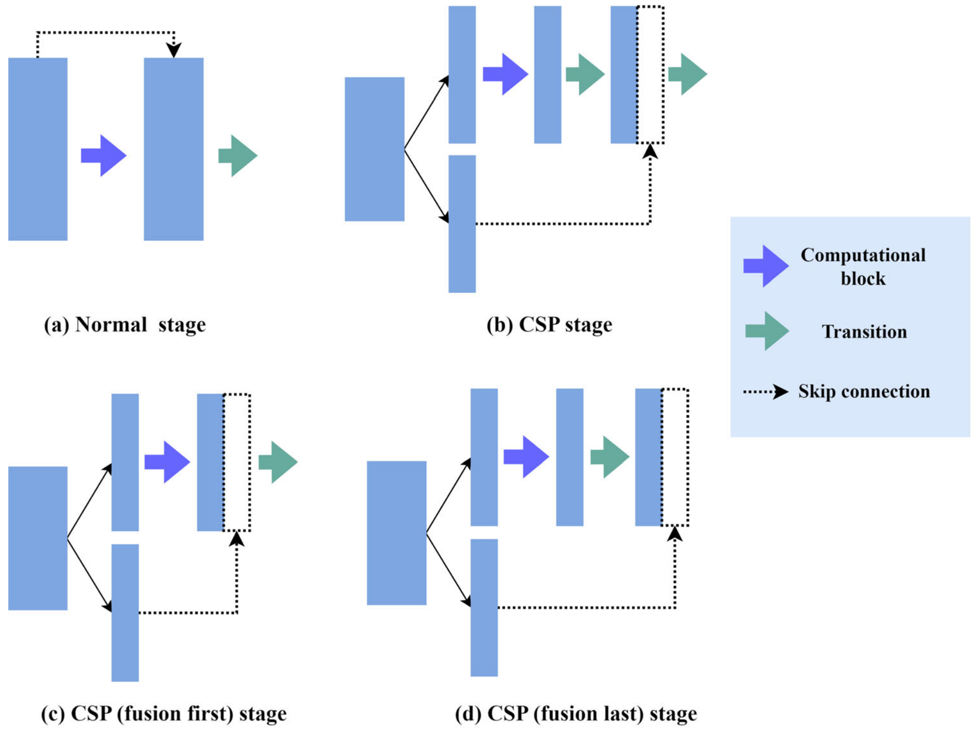


Fig. 5 Regardless of the model, whether it is shallow or deep, the deeper layers in a deep network have the ability to extract both low-level and high-level properties[59, 60]



(a) Normal stage

(b) CSP stage

(c) CSP (fusion first) stage

(d) CSP (fusion last) stage

Fig. 6 Architecture of CSPNet compared to traditional models: **a** Normal (Resnet) stage, **b** CSP stage, **c** CSP stage with fusion first, **d** CSP stage with fusion last

type of architecture can efficiently minimize the number of parameters, operations, memory traffic, and memory peak, resulting in faster inference performance for the model. CSPNet can be constructed in multiple ways which is summarized in Fig. 6.

The other component of ELAN is VoVNet [61]. VoVNet is a convolutional neural network that aims to enhance the efficiency of DenseNet [63] by concatenating all features just once in the final feature map known as one-shot aggregation (OSA). This approach ensures a constant input size and allows for the expansion of new output channels. A comparison of DenseNet and VoVNet is shown in Fig. 7.

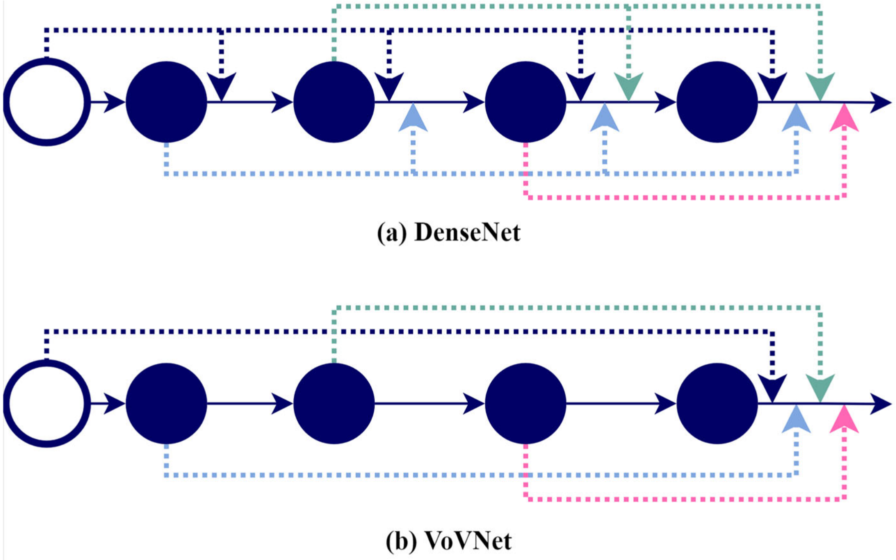
Similar to most of the popular convolutional networks, VoVNet suffers from scaling issues, that is with an increase in the number of layers after a certain point, the accuracy starts decreasing [64]. In fact, the accuracy of VoVNet degrades faster than ResNet [65] due to the utilization of the OSA module in the stacking of VoVNet. As each OSA module has a transition layer, if we add an OSA module to the stack, the shortest gradient path of all layers in the network rises by one as opposed to ResNet. In order to address this problem, ELAN prevents the rapid elongation of the shortest gradient path by reducing the number of transition layers. Consequently, ELAN chooses to aggregate features after many stages, as depicted in Fig. 8, resulting in a decreased number of transition layers.

The architecture of ELAN ensures that each layer can be trained effectively by keeping the length of gradient paths manageable. To summarize, this final architecture has three primary advantages: (1) achieve better learning ability, (2) reduce the number of parameters and hardware resources, and (3) improve the inference speed.

3.5 Generalized efficient layer aggregation network (GELAN)

GELAN is an object detector based on YOLOv7 and modified ELAN that achieves superior performances compared to state-of-the-art models[56]. One of the limitations of ELAN is it only works with convolutional blocks. GELAN provides more flexibility than ELAN which is capable of accommodating different computational blocks. This flexibility enables it to adjust to various tasks and computational settings, resulting in wider applicability. A model architecture of the proposed GELAN model is shown in Fig. 9.

Fig. 7 Comparison between DenseNet (top) and VoVNet (bottom)



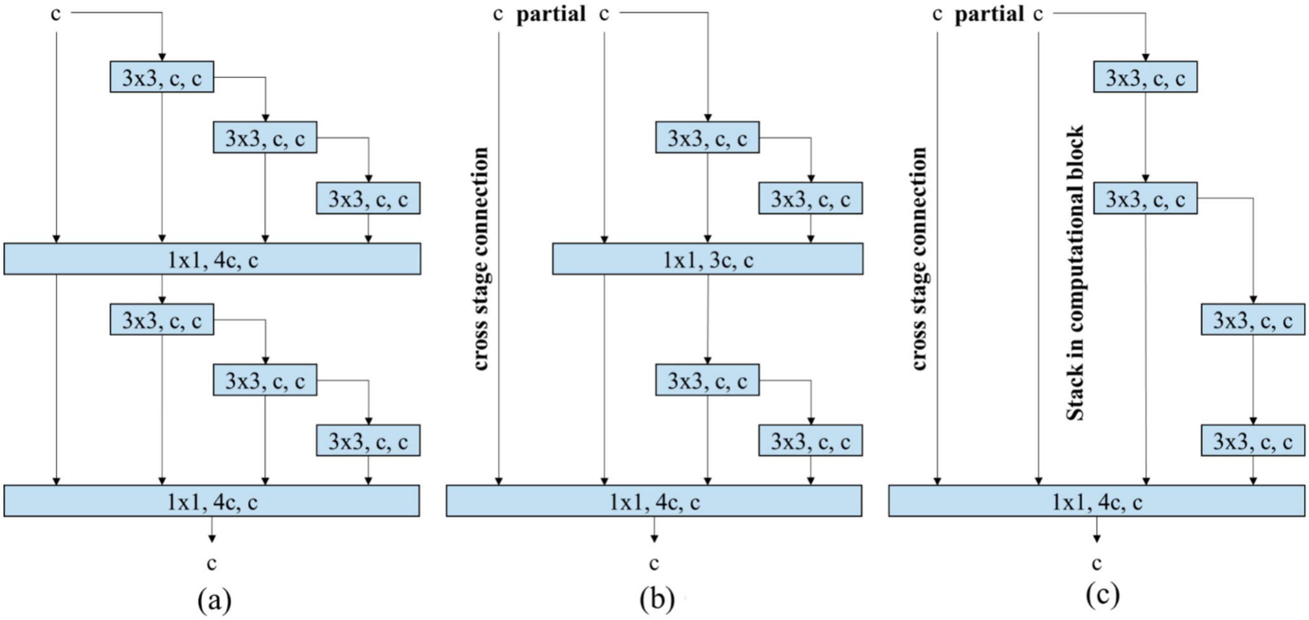
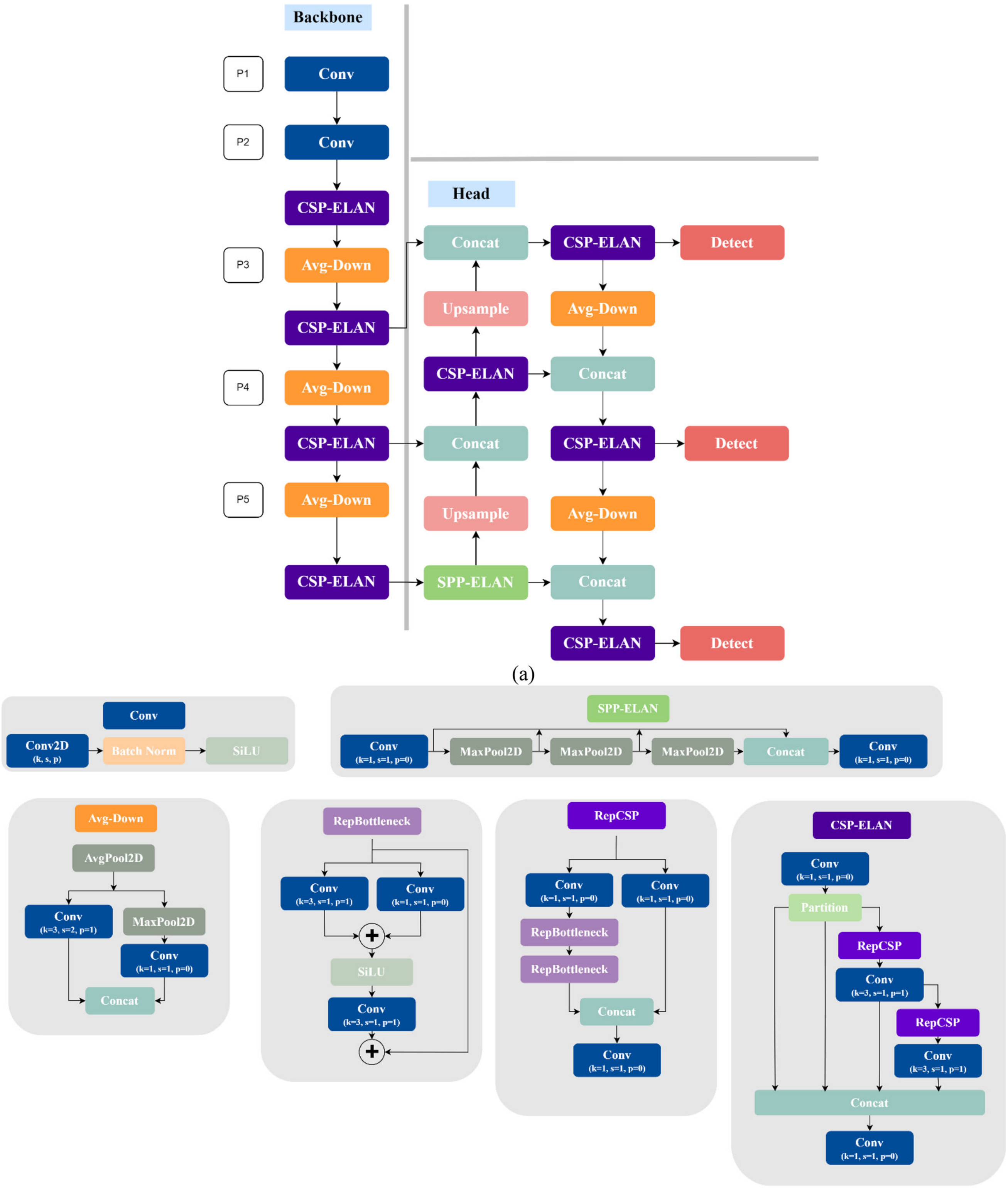


Fig. 8 Gradual modifications to obtain the architecture of ELAN: **a** VoVNet, **b** CSPVoVNet [66], and **c** ELAN

The overall architecture of GELAN is divided into two different sections: backbone and head. The backbone is primarily tasked with extracting features from the input image. It begins with convolutional blocks that downsample the image, reducing its spatial dimensions while retaining critical information. These are followed by CSP-ELAN blocks, which perform intricate operations to extract hierarchical features at various scales. By combining convolutional layers with advanced feature aggregation techniques, these blocks enable the model to capture detailed information effectively. Avg-Down blocks, which perform average pooling and downsampling, further reduce the spatial dimensions of feature maps, retaining essential details and facilitating easier processing in subsequent layers.

The head of the GELAN model processes the aggregated features for final detection. It includes the SPPELAN block, incorporating spatial pyramid pooling to capture multi-scale contextual information, which is crucial for detecting objects of different sizes. CSP-ELAN blocks are responsible for efficient feature aggregation from different levels of the backbone and further enable enhanced gradient flow. The architecture also employs



(a)

(b)

Fig. 9 Model architecture of GELAN **a** the simplified architecture and **b** detailed architecture of each block

upsample and concat layers, which increase the resolution of feature maps and merge features from different scales, facilitating comprehensive information integration. Finally, the detect block then performs the critical task of detecting objects and determining their locations within the input image.

3.6 Task description

We will provide a brief description of the tasks; we are facilitating in this study:

Task A: The primary task of our study is to automatically detect and classify different kinds of trauma from X-ray images including fractures, periosteal lesions, pronator signs, and soft tissues. It is a multi-class detection task for which we removed classes other than the selected ones for training and testing purposes.

Task B: Statistically, the number of patients is much higher for fracture cases compared to other ones. In fact, the fracture class comprised more than 75% of the images in the dataset even after removing some of the minority classes making the dataset hugely imbalanced without any augmentation. Therefore, to aid this situation, we have constructed a separate single class detection task exclusively for fractures.

Task C: As discussed in the introduction section, children's fractures are different from adult fractures. Children's bones are normally more brittle and tender resulting in incomplete fractures in many cases that are almost non-existent for adults. In order to convey the severity information, we detected and classified them into "complete" and "incomplete" fractures using the image "AO classification code" tags described in Table 1. The classification code followed the AO pediatric comprehensive classification of long bone fractures (PCCF) standard from the fracture and dislocation classification compendium—2018[67] as described in Fig. 10.

According to the standard, the fracture area can be precisely indicated using the bone, segment, (optionally, the paired bone), and the sub-segment with their respective values. The morphology of the fracture is specified using pattern and severity. However, the pattern itself has no specific meaning; rather, the bone, segment, sub-segment, and pattern combined represent the specific type of fracture. Following the standard, we labeled three types of fractures (22-D/1.1, 22-D/2.1, 23-M/2.1, and their radius/ulna variants) around the wrist as incomplete fractures and the rest of them as complete fractures. Finally, we used the labeled data as the ground truth for Task C. Figure 11 shows the incomplete fractures with their corresponding classification code.

3.7 Evaluation metrics

Evaluation metrics play a vital role in measuring the effectiveness of machine learning models and facilitating meaningful comparisons between various applications. We used a variety of commonly utilized metrics to evaluate the performance of our models at different experimental levels.

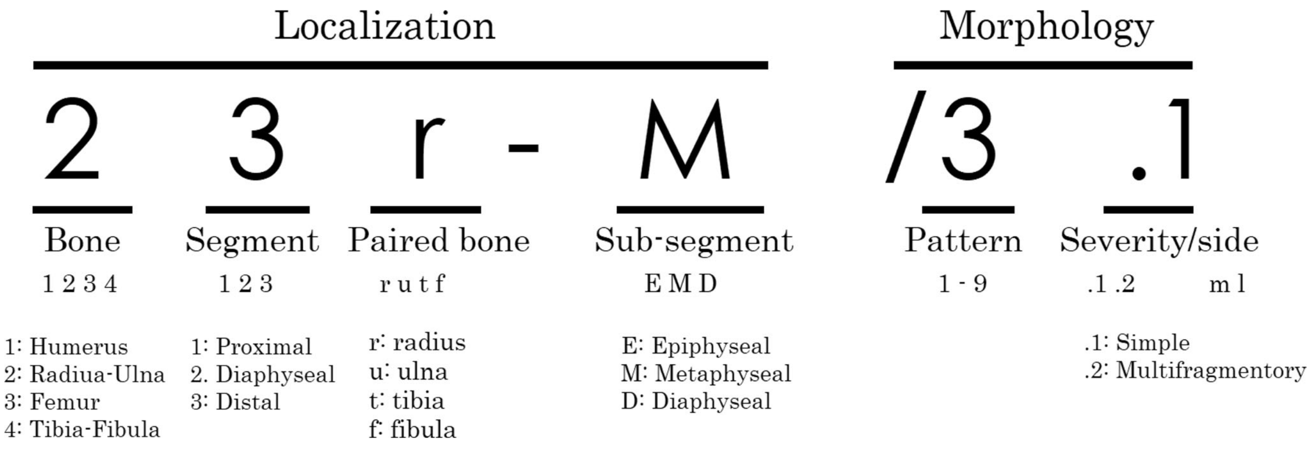


Fig. 10 The standard of PCCF fracture classification code

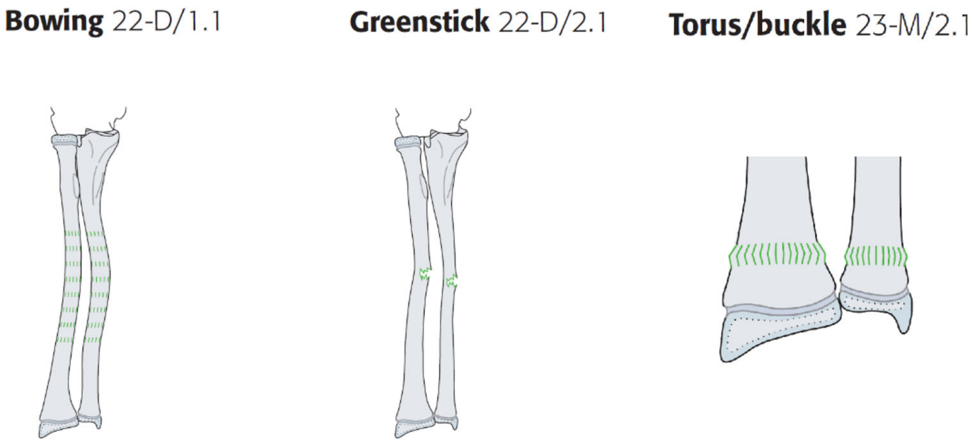


Fig. 11 Selected incomplete fractures around the wrist

Accuracy (Acc): Accuracy is a commonly used metric for assessing the performance of a machine learning model, as it quantifies the model's overall capacity to predict the correct output. This metric represents the ratio of correctly predicted observations to the total number of observations. While accuracy is a valuable measure, it may not offer a comprehensive evaluation of the model's efficacy, especially when working with datasets containing imbalanced classes. Denoting true positives, false positives, true negatives, and false negatives as TP, FP, TN, and FN respectively, accuracy can be defined as:

$$\text{Accuracy}, Acc = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2}$$

Precision (Pr): Precision is a quantitative measure used to evaluate the accuracy of positive predictions produced by a model. The metric in issue quantifies the proportion of correct positive predictions out of all positive predictions, thereby providing useful insights into the model's ability to reduce false positives. Precision is especially important in situations where false positives might have significant consequences, such as in the field of medical diagnostics.

$$\text{Precision}, Pr = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Sensitivity (Se): Sensitivity, sometimes referred to as recall or true positive rate, refers to a model's ability to completely identify all positive cases in a given dataset. The approach calculates the accuracy by dividing the number of properly detected positive predictions by the total number of positive cases in reality. The notion of sensitivity is especially pertinent in situations when the inability to detect positive occurrences might lead to substantial consequences.

$$\text{Sensitivity}, Se = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

F1 score: The F1 score is a composite metric that integrates precision and sensitivity (or recall), offering an equitable assessment of a model's efficacy. The harmonic mean underscores the significance of precision and recall in equal measure. The F1 score is a useful metric in cases where there is an imbalance in the dataset, as it takes into account both the false positives and false negatives.

$$F1\text{score} = \frac{2 \times Pr \times Se}{Pr + Se} \tag{5}$$

Intersection over Union (IoU): Intersection over Union (IoU) serves as a performance metric used to evaluate the precision of segmentation and object detection algorithms. It gauges the level of overlap between the predicted bounding box or segmented region and the ground truth bounding box or annotated region within a dataset. IoU quantifies the degree of agreement between a predicted object and its corresponding object annotation. Mathematically, IoU is computed by dividing the intersection of the predicted and ground truth areas by their union, expressed as:

$$\text{Intersectionoverunion}, IoU = \frac{\text{Areaofintersection}}{\text{Areaofunion}} \tag{6}$$

Mean average precision (mAP): Mean average precision (mAP) is a commonly employed assessment metric in object identification that offers a consolidated estimate of a model's performance across different levels of precision and recall. Different levels of IoU and confidence thresholds yield varying accuracy and recall scores in practical applications. Therefore, to comprehend the nature of this trade-off, precision-recall curves are usually plotted. Average precision (AP) for a single class is essentially the area under this precision-recall curve and mean average precision is the mean of AP across all the classes. If average precision is denoted by AP and the total number of classes is N, it can be shown mathematically,

$$\text{Meanaverageprecision}, mAP = \frac{1}{N}\sum\nolimits_{i=1}^{n} AP_i \tag{7}$$

Usually, mAP is calculated at a fixed IoU threshold and the most common practice is to mention the threshold as $mAP : T$ where T is the IoU threshold. However, for different thresholds, the calculated mAP will be different which makes it harder to compare between mAP s with different thresholds. A standard procedure is to calculate the mAP at different thresholds from 0.5 to 0.95 with an increment of 0.05 and take the average of all the calculated $mAPs$. This is known as $mAP50 - 95$ or $mAP0.5 : 0.95$.

$$mAP50 - 95 = \frac{1}{10} \times \sum\nolimits_{t=0.5}^{0.95} mAP : t \tag{8}$$

3.8 Experimental setup

The experiments in this study were conducted using a cloud-based computing environment equipped with high-performance GPUs. This choice of environment provided us with the computational resources needed to

Table 3 Experimental setup

Component	Specification
CPU	13th Gen Intel® Core^{TM} i7-13700KF @ 3.40 GHz
GPU	NVIDIA RTX 4090 (24 GB VRAM)
RAM	64 GB DDR5
Storage	2 TB SSD
Operating system	64-bit Microsoft Windows® 10
Software environment	Python 3.11, Torch 2.1.0, CUDA 11.8
Other Python libraries	Numpy, Pandas, Matplotlib, SciPy, OpenCV, Pillow, Albumentations, Torchvision,

efficiently train and evaluate our deep learning model on a sizable dataset of pediatric respiratory sounds. Technical specifications of the computing setup are given in Table 3**.**

4 Results

4.1 Performance evaluation

For all three tasks, we conducted a fivefold cross-validation strategy to validate the performance of our proposed model. To address subject-level data leaking, no single subject was present in two different splits. The performance for all three tasks using different models including our proposed framework is provided in Tables 4, 5, and 6, respectively. As the YOLOv5 and YOLOv7 models performed worst in the primary task (Task A), we did not include them in later experiments.

Detailed class-wise performance metrics are presented in Table 7 and 8 for Task A and Task C, respectively. As Task B contains a single class, its overall performance represents the class performance.

Additionally. The precision-recall curves for all the tasks are given in Fig. 12:

In all three tasks, GELAN-E performed the best with a mAP50-95 score of 44.9%, 61%, and 55.6% respectively. YOLOv9 was the closest competitor of GELAN models. Though YOLOv9, with the implementation of PGI and auxiliary branches, achieved better performance in the MS COCO dataset benchmark, GELAN performed better in our study. A few sample predictions from each task are provided in Fig. 13 for qualitative analysis:

To enhance trust and transparency in clinical applications, we integrated Eigen-Grad-CAM, an explainable AI technique that provides visual heatmaps highlighting the most relevant image regions contributing to the model's predictions. This helps radiologists and clinicians interpret AI-driven decisions effectively and provides an opportunity for error analysis in real-world clinical settings. We have overlayed the visual heatmaps from Eigen-Grad-CAM on the X-ray images in Fig. 14 to provide interpretability of the model's decision:

A comparison of our proposed model with other state-of-the-art approaches in MRI image segmentation is discussed in Table 9:

As seen from the table, our proposed methodology performed better than most other studies. [30, 40, 42, 47, 49] was similar to Task B in this study. Our approach performed much better in the GRAZ-PREDWRI-DX dataset compared to these state-of-the-art studies Except for [47], our approach superseded the performance in other datasets also. However, [47] used a very small set of data comprising only 38 images that questioned the achieved score. [43] did a similar experiment as our Task A, and our proposed approach achieved 13% better performance. We did not find any research that investigated the severity of the fracture as we did in Task C.

Table 4 Performance of different models in Task A

Model	#Params	Precision	Sensitivity	F1 Score	Mean IoU	mAP50	mAP50-95
YOLOv5l	46.5M	0.6317	0.6211	0.626	0.688	0.6143	0.3189
YOLOv5x	86.7M	0.6004	0.6441	0.621	0.701	0.6246	0.3211
YOLOv7-X	71.3M	0.6579	0.6109	0.634	0.742	0.6525	0.3411
YOLOv7-E6	97.2M	0.6685	0.6264	0.647	0.749	0.6677	0.3538
YOLOv8l	43.7M	0.64	0.655	0.647	0.72	0.659	0.352
YOLOv8x	68.2M	0.659	0.662	0.66	0.726	0.692	0.354
YOLOv9-C	25.3M	0.715	0.63	0.67	0.764	0.693	0.413
YOLOv9-E	67.3M	0.73	0.641	0.683	0.791	0.72	0.419
GELAN-C	25.3M	0.728	0.654	0.689	0.798	0.714	0.421
GELAN-E	**67.3M**	**0.805**	**0.652**	**0.720**	**0.798**	**0.741**	**0.448**

Bold values represent the best performing cases

Table 5 Performance of different models in Task B

Model	#Params	Precision	Sensitivity	F1 Score	Mean IoU	mAP50	mAP50-95
YOLOv8l	43.7M	0.892	0.861	0.876	0.761	0.89	0.56
YOLOv8x	68.2M	0.893	0.883	0.888	0.77	0.901	0.577
YOLOv9-C	25.3M	0.925	0.854	0.888	0.799	0.922	0.583
YOLOv9-E	67.3M	0.922	0.861	0.890	0.803	0.923	0.599
GELAN-C	25.3M	0.916	0.874	0.895	0.803	0.936	0.598
GELAN-E	**67.3M**	**0.943**	**0.899**	0.920	**0.805**	**0.95**	**0.611**

Bold values represent the best performing cases

Table 6 Performance of different models in Task C

Model	#Params	Precision	Sensitivity	F1 Score	Mean IoU	mAP50	mAP50-95
YOLOv8l	43.7M	0.776	0.761	0.768	0.766	0.799	0.511
YOLOv8x	68.2M	0.781	0.766	0.773	0.771	0.813	0.52
YOLOv9-C	25.3M	0.806	0.798	0.802	0.795	0.828	0.53
YOLOv9-E	67.3M	0.818	0.819	0.818	0.801	0.856	0.544
GELAN-C	25.3M	0.817	0.808	0.812	0.805	0.853	0.544
GELAN-E	**67.3M**	**0.824**	**0.832**	**0.828**	**0.809**	**0.855**	**0.555**

Bold values represent the best performing cases

Table 7 Detailed class-wise performance evaluation of the GELAN-E model for Task A

Class	Precision	Sensitivity	F1 score	Mean IoU	mAP50	mAP50-95
Fracture	0.933	0.912	0.922	0.801	0.949	0.604
Periosteal reaction	0.742	0.68	0.71	0.779	0.715	0.409
Pronator sign	0.783	0.696	0.737	0.781	0.794	0.474
Soft Tissue	0.682	0.375	0.484	0.774	0.51	0.282
Overall	**0.805**	**0.652**	**0.72**	**0.798**	**0.741**	**0.448**

Bold values represent the best performing cases

Table 8 Detailed class-wise performance evaluation of the GELAN-E model for Task C

Class	Precision	Sensitivity	F1 score	Mean IoU	mAP50	mAP50-95
Complete fracture	0.875	0.792	0.831	0.8	0.869	0.537
Incomplete fracture	0.788	0.849	0.817	0.821	0.859	0.576
Overall	**0.824**	**0.832**	**0.828**	**0.809**	**0.855**	**0.555**

Bold values represent the best performing cases

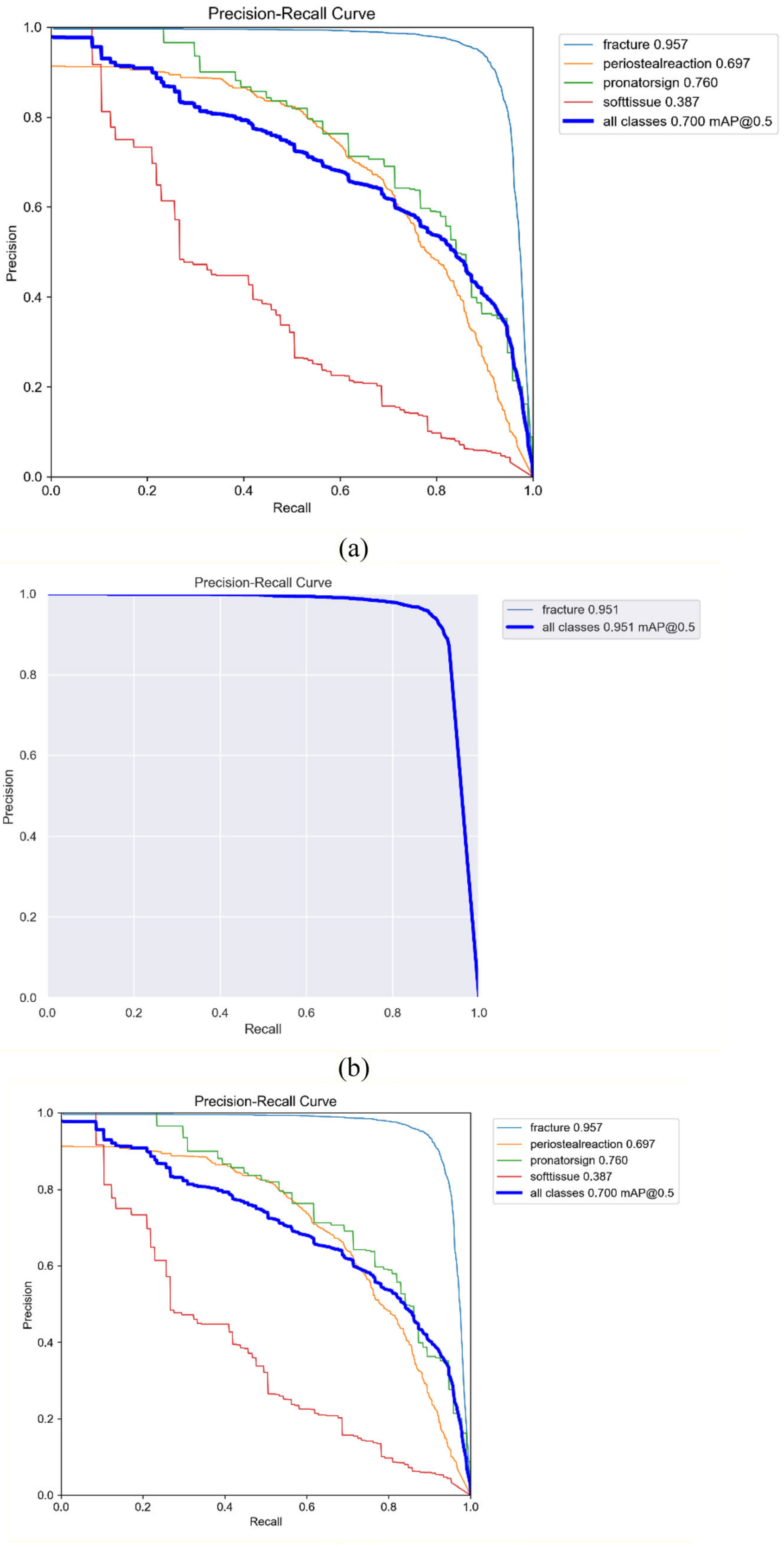
To assess the computational feasibility of our model, we evaluated its inference speed and resource consumption on an NVIDIA RTX 4090 GPU (the complete experimental setup is shown in Table 3). We have summarized the inference time and GPU usage in Table 10. While evaluating the computation, the batch size was set to 1 and the image size was 960 × 960.

The table shows that, even though our proposed model shows superior performance compared to the state-of-the-art models, it is computationally feasible for real-life usage.

4.2 Statistical significance

To rigorously validate the performance improvements of our proposed GELAN model, we conducted statistical significance tests and computed confidence intervals (CIs) for key evaluation metrics, including mean average

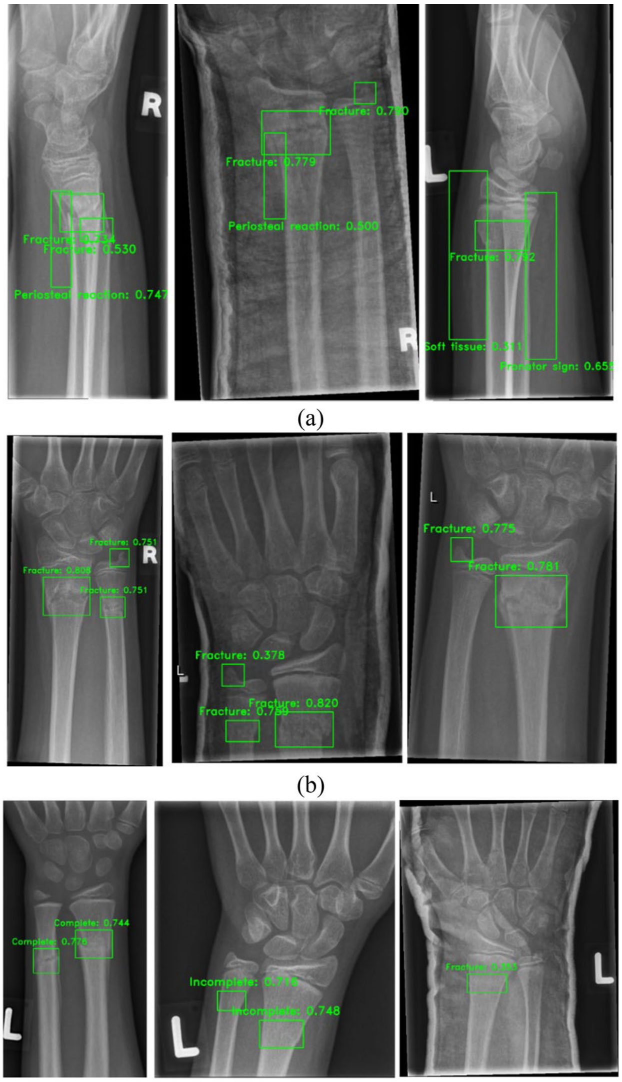
Fig. 12 Precision-recall curves for **a** Task A, **b** Task B, and **c** Task C



(a)

(b)

(c)



(a)

(b)

(c)

Fig. 13 Sample predictions made by the proposed model: **a** Task A, **b** Task B, and **c** Task C

precision (mAP), precision, and recall. These analyses ensure that our model's superior performance is not due to chance but represents a meaningful improvement over existing methods.

For each performance metric, we computed 95% confidence intervals (CIs) using the t-distribution method:

$$\mathrm{CI} = \overline{\mathrm{x}} \pm \mathrm{t}_{\frac{\alpha}{2},\mathrm{df}} \times \frac{\mathrm{s}}{\sqrt{\mathrm{n}}} \tag{9}$$

where $\bar{x}$ is the mean performance, s is the standard deviation, s is the number of test runs (fivefold cross-validation), and $t_{\alpha/2,df}$ is the critical t-value for α confidence level. Table 11 presents the computed 95% confidence intervals for mAP50, mAP50-95, precision, and recall for the GELAN-E model.

To assess whether the performance improvements of our proposed model are statistically significant, we conducted a paired statistical analysis comparing our model with the two other models: YOLOv8x and YOLOv9-E. These models were chosen as they show performances closest to our proposed model. Since the evaluation was performed using k-fold cross-validation, we treated the results from each fold as paired observations.

First, we applied a paired t-test, which assumes a normal distribution of differences, to compare our model's mean performance against the baselines. Additionally, we performed the Wilcoxon signed-rank test to validate

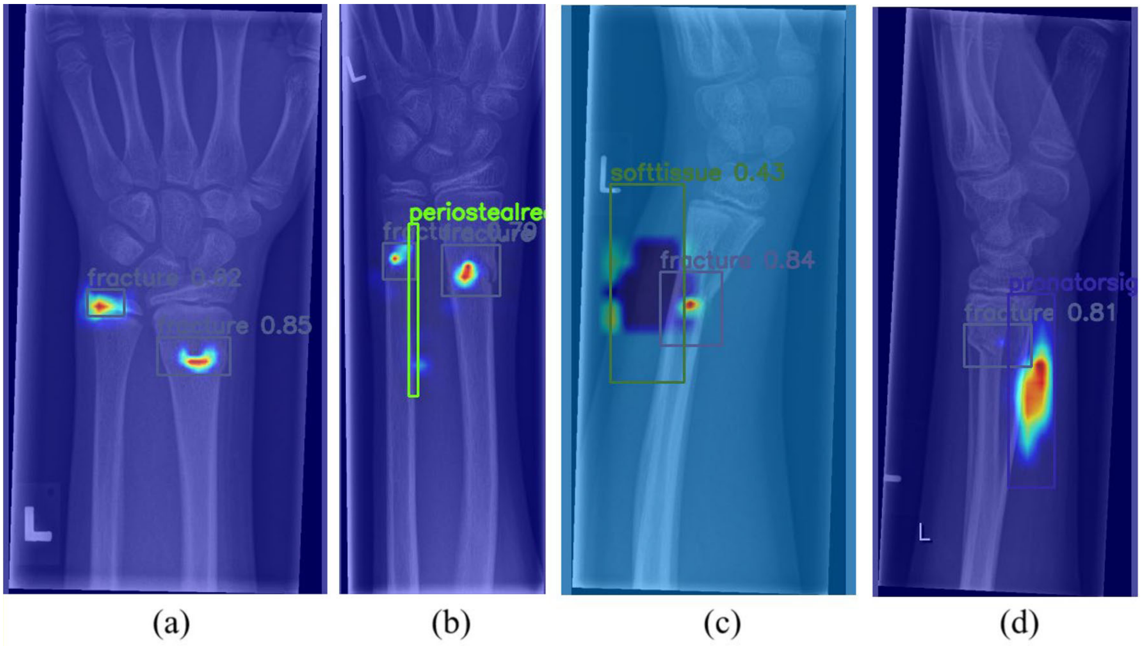


Fig. 14 Visual heatmap overlay using Eigen-Grad-CAM

Table 9 Performance comparison with the state-of-the-art methods

Year	Ref	Dataset	Classes	Method	Performance
2019	[47]	Private (38 images)	1: Fracture	Faster R-CNN	mAP: 0.866
2021	[30]	Private (1946 images)	1: Fracture	KNEEL, SeResNet	Test #1: Acc: 0.96 Test #2: Acc: 0.82
2022	[40]	Private (542 images)	1: Fracture	Weighted bounding box fusion	AP: 0.8639 AR: 0.33
2022	[49]	GRAZPEDWRI-DX	1: Fracture	YOLOv4	AUC-ROC: 0.965
2023	[43]	GRAZPEDWRI-DX	4: Fracture, periosteal reaction, foreign body, and bone anomaly	YOLOv7 with GAM attention	mAP50: 0.654 mAP50-95: 0.398
2023	[50]	GRAZPEDWRI-DX	8: Fracture, bone anomaly, bone lesion, foreign body, metal, periosteal reaction, pronator sign, and soft tissue	YOLOv8	mAP50: 0.591 mAP50-95: 0.372
2023	[42]	GRAZPEDWRI-DX	1: Fracture	YOLOv8m	mAP50: 0.621 mAP50-95: 0.403
	Proposed	GRAZPEDWRI-DX	Task A: 4: Fracture, periosteal reaction, pronator sign, and soft tissue	GELAN-E	mAP50: 0.741 mAP50-95: 0.448
			Task B: 1: Fracture		mAP50: 0.95 mAP50-95: 0.611
			Task C: 2: Complete, incomplete fracture		mAP50: 0.855 mAP50-95: 0.555

the significance of our findings without assuming normality. The results show that our proposed model significantly outperforms both baselines across multiple evaluation metrics. Specifically, the Wilcoxon signed-rank tests yielded a p-value of 0.0312 and 0.0312 when comparing our model to YOLOv8-x and YOLOv9-E, respectively. In case of paired t-test, the p-values were 0.000, and 0.001 both of which are below the standard significance threshold ($p < 0.05$) confirming statistical significance.

Table 10 Inference speed and resource consumption

Stage	Time (ms)	VRAM usage (GB)
Preprocessing	0.1	0
Inference	6.5	2.1
Non-max suppression	1	2.1
Post-processing	1.5	0

Table 11 95% Confidence intervals for different evaluation metrics for the proposed model

Task	Precision CI	Sensitivity CI	mAP50 CI	mAP50-95 CI
Task A	0.783–0.827	0.624–0.680	0.733–0.749	0.443–0.454
Task B	0.938–0.948	0.892–0.906	0.947–0.953	0.608–0.614
Task C	0.809–0.839	0.813–0.851	0.838–0.871	0.547–0.562

To further quantify the improvement, we computed Cohen's d effect size, which measures the magnitude of the difference between models. The effect sizes were 13.58 and 4.23 when comparing our model to YOLOv8-x and YOLOv9-E, respectively, indicating a large effect size ($d > 0.8$), demonstrating the practical significance of our improvements. These statistical results confirm that our model's superior performance is not due to random chance but represents a significant improvement over the baseline models.

4.3 Ablation study

During the process of designing our model, we came up with various ideas and concepts to explore and test. Although some of those excelled, not all of them achieved the same level of performance. It is vital to assess the significance of the characteristics in each system and eliminate any that are not needed. An ablation study refers to the systematic process of deleting or replacing components inside a machine learning framework to assess the impact of these changes on performance. In this study, we conducted a series of ablation tests to eliminate superfluous components from the system. We utilized a randomized search technique to tune the hyperparameters.

Image size: An ablation study was conducted to assess the impact of image size on the GELAN-E model's performance. We tested input sizes of 640 × 640, 960 × 960, and 1280 × 1280 pixels. The raw images had different resolutions, but a significant portion of them were close to 1000 pixels. As a result, 960 pixels and 1080 pixels images gave almost the same performances. To maintain computational efficiency, we selected 960-pixel images for this experiment.

Confidence threshold: A confidence threshold is a predefined value that determines the minimum confidence score an object detection model must-have for a predicted bounding box to be considered a valid detection. We performed an ablation study to examine the impact of different confidence thresholds on the performance of the proposed model. Various confidence thresholds, in the range of 0.1–0.5, were tested to evaluate their effect on detection accuracy and false positive rates. Lower thresholds, like 0.1, increased sensitivity but resulted in a higher number of false positives. Conversely, higher thresholds, such as 0.5, reduced false positives but also missed more true positives. The optimal balance was achieved at a confidence threshold of 0.3, which provided a favorable trade-off between sensitivity and precision, enhancing the model's overall robustness in detecting and classifying pediatric fractures.

Learning rate: The learning rate is a crucial hyperparameter in model training. In this study, we examined several learning rates to determine the optimal value for achieving higher performance and faster training. By doing the ablation investigation, we found that the learning rates of 0.03 and 0.1 were too high, which hindered the convergence of the model. On the other hand, 0.003 was too low for the learning rate. A learning rate of 0.01 was identified as the optimal selection, as emphasized in Table 12. All the performances were evaluated in Task A.

Table 12 Result of the ablation studies

Hyperparameter	Value	mAP50	mAP50-95
Image Size	640	0.738	0.44
	960	**0.743**	**0.449**
	1280	0.743	0.448
Confidence threshold	0.1	0.726	0.436
	0.2	0.741	0.442
	0.3	**0.743**	**0.449**
	0.4	0.74	0.445
	0.5	0.738	0.442
Learning rate	0.003	0.734	0.436
	0.01	**0.743**	**0.449**
	0.03	0.733	0.446
	0.1	0.741	0.442

5 Discussion

Traditional pediatric trauma diagnosis often relies on the manual interpretation of X-ray images by radiologists. Despite being effective sometimes, it is subject to several limitations, including high inter-observer variability, diagnostic delays, and limited availability of specialists. Automated deep learning models have emerged as potential solutions, but existing approaches primarily focus on adults which is anatomically different from pediatrics based on bone structure, fracture complexity, and the presence of growth plates. Deep learning-based object detection methods have demonstrated promising results in fracture detection; yet, they often struggle with class imbalance and fail to incorporate fracture severity classification. Our study focuses on these gaps and improves upon these methods by leveraging GELAN, a more efficient architecture for feature aggregation and gradient optimization, resulting in superior detection accuracy and classification performance.

According to Tables 4**,** 5, and 6, our proposed approach showed significantly better performance compared to recent state-of-the-art models such as YOLOv8 and YOLOv9. In comparison, GELAN-E achieved 26.8%, 5.72%, and 6.92% better mAP50-95 scores in Task A, B, and C, respectively. Randomly selected sample predictions from Fig. 13 also justify this result qualitatively. Furthermore, compared to the state-of-the-art research on this problem area, our work beats all of them in both binary and multi-class detection tasks proposing a new severity detection approach never addressed by previous researchers.

Based on the significant improvement made by our proposed approach, it can assist in multiple clinical scenarios:

- Rapid preliminary assessments to triage patients and prioritize urgent cases.
- Decision support for identifying fractures and assessing severity (complete vs. incomplete fractures), reducing workload and interobserver variability.
- Enabling fracture detection in under-resourced hospitals, assisting non-specialist clinicians.
- Providing a consistent AI-assisted interpretation to mitigate diagnostic variability.

While not a replacement for radiologists, our proposed system can serve as a valuable adjunct in high-volume hospitals and resource-limited settings, improving clinical efficiency and patient outcomes. Future validation in real-world environments will further establish its clinical utility.

Despite the promising results of our study, several limitations need to be addressed. First, the GRAZPEDWRI-DX dataset, while comprehensive, may not fully represent the diversity of pediatric wrist fractures encountered in clinical practice. The dataset is limited to annotated X-ray images, which might not capture all the variations seen in different populations or under different clinical conditions. A multimodal dataset containing CT scan and MRI modalities could have been more helpful in this scenario. Additionally, the performance of our model may vary when applied to data from different imaging equipment or settings. Another limitation is the inherent challenge of class imbalance within the dataset. Although our GELAN architecture addresses this issue to some extent, further refinement is needed to ensure consistent performance across all classes, particularly for rare fracture types. Furthermore, while our model demonstrated high accuracy in trauma and severity detection, it has not yet been validated in a clinical setting. Real-world implementation requires extensive testing and validation to ensure the model's reliability and robustness in everyday medical practice.

In terms of future directions, expanding the dataset to include a more diverse range of images and fracture types would enhance the model's generalizability. Incorporating images from different anatomical regions and other imaging modalities like CT and MRI could also improve the model's applicability. Moreover, integrating our deep learning framework into clinical workflows would require the development of user-friendly interfaces and collaboration with healthcare professionals to refine the system based on practical feedback. Future research should also focus on improving the interpretability of the model's predictions, potentially through advanced visualization techniques and explainable AI methods. Finally, large-scale clinical trials are necessary to validate the effectiveness of the proposed system in real-world settings. By addressing these limitations and pursuing these future directions, we can further enhance the utility and impact of automated pediatric trauma detection and classification systems.

6 Conclusion

In this study, we used deep learning techniques to develop an automated system for detecting and classifying pediatric trauma and fractures in X-ray images. Our proposed framework based on the generalized efficient layer aggregation network (GELAN) addresses the critical need for timely and accurate diagnosis in pediatric trauma care, particularly given the scarcity of radiologists and specialized equipment. Our framework performed particularly well across a wide range of tasks, including multi-class detection of various trauma types and fracture severity detection into complete and incomplete categories. According to our findings, the GELAN-E model outperformed the state-of-the-art method. The model's ability to detect and classify pediatric fractures shows promise for improving clinical workflows and patient outcomes. The use of advanced deep learning techniques in pediatric radiology has the potential to reduce radiologists' workload, shorten diagnostic delays, and ensure that children receive the care they require as soon as possible. Future work will concentrate on improving the model to handle a wider range of trauma types and investigating its applicability to other anatomical regions. In addition, we plan to investigate the integration of this automated system into clinical settings, assessing its real-world impact on diagnostic accuracy and efficiency. By continuing to improve deep learning capabilities in medical imaging, we hope to contribute to ongoing efforts to improve pediatric healthcare and outcomes.

Acknowledgements Open access publication of this article is supported by the Qatar National Library (QNL).

Author contribution Promit Basak: Conceptualization, Methodology, Software, Writing—Original draft preparation, Writing—Reviewing and Editing. Adam Mushtak: Conceptualization, Validation, Writing- Original draft preparation, Writing- Original draft preparation. Mohamed Ouda: Validation, Supervision, Writing—Original draft preparation, Writing—Reviewing and Editing, Funding Acquisition. Sadia Farhana Nobi: Methodology, Software, Validation, Writing- Original draft preparation. Anwarul Hasan: Conceptualization, Methodology, Software, Writing—Original draft preparation, Writing—Reviewing and Editing. Muhammad E. H. Chowdhury: Conceptualization, Methodology, Supervision, Funding Acquisition, Writing—Original draft preparation, Writing- Reviewing and Editing.

Funding Open Access funding provided by the Qatar National Library.

Dataset availability The dataset used in this study was made publicly available by Nagy et al. [54].

Declarations

Conflict of interest The authors have no conflicts of interest to disclose for this study.

References

1. Aoki M, Abe T, Saitoh D, Oshima K (2019) Epidemiology, patterns of treatment, and mortality of pediatric trauma patients in Japan. Sci Rep 9(1):917
2. Oliver J, Avraham J, Frangos S, Tomita S, DiMaggio C (2018) The epidemiology of inpatient pediatric trauma in United States hospitals 2000 to 2011. J Pediatr Surg 53(4):758–764
3. Kundal VK, Debnath PR, Sen A (2017) Epidemiology of pediatric trauma and its pattern in urban India: a tertiary care hospital-based experience. J Indian Assoc Pediatr Surg 22(1):33–37. https://doi.org/10.4103/0971-9261.194618
4. Chaudhari PP et al (2022) Epidemiology of pediatric trauma during the coronavirus disease-2019 pandemic. J Pediatr Surg 57(2):284–290
5. Malek M, Chang B-H, Gallagher SS, Guyer B (1991) The cost of medical care for injuries to children. Ann Emerg Med 20(9):997–1005
6. Owens PL, Zodet MW, Berdahl T, Dougherty D, McCormick MC, Simpson LA (2008) Annual report on health care for children and youth in the United States: focus on injury-related emergency department utilization and expenditures. Ambulatory Pediatr 8(4):219–240
7. Mokdad AH et al (2016) Global burden of diseases, injuries, and risk factors for young people's health during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 387(10036):2383–2401
8. Cintean R, Eickhoff A, Zieger J, Gebhard F, Schütze K (2023) Epidemiology, patterns, and mechanisms of pediatric trauma: a review of 12,508 patients. Eur J Trauma Emerg Surg 49(1):451–459
9. Randsborg P-H et al (2013) Fractures in children: epidemiology and activity-specific fracture rates. J Bone Joint Surg 95(7):e42
10. Little JT, Klionsky NB, Chaturvedi A, Soral A, Chaturvedi A (2014) Pediatric distal forearm and wrist injury: an imaging review. Radiographics 34(2):472–490
11. Frost HM, Schönau E (2000) The" muscle-bone unit" in children and adolescents: a 2000 overview. J Pediatr Endocrinol Metab 13(6):571–590
12. Davis KW (2010) Imaging pediatric sports injuries: upper extremity. Radiol Clin North Am 48(6):1199–1211
13. Randsborg P-H, Sivertsen EA (2009) Distal radius fractures in children: substantial difference in stability between buckle and greenstick fractures. Acta Orthop 80(5):585–589
14. Breitenseher MJ, Gaebler C (1997) Trauma of the wrist. Eur J Radiol 25(2):129–139
15. Fotiadou A, Patel A, Morgan T, Karantanas AH (2011) Wrist injuries in young adults: the diagnostic impact of CT and MRI. Eur J Radiol 77(2):235–239
16. Wright AA, Hegedus EJ, Lenchik L, Kuhn KJ, Santiago L, Smoliga JM (2016) Diagnostic accuracy of various imaging modalities for suspected lower extremity stress fractures: a systematic review with evidence-based recommendations for clinical practice. Am J Sports Med 44(1):255–263
17. Kraus R and Dresing K (2023) Rational usage of fracture imaging in children and adolescents," *Diagnostics,* 13(3): 538. [Online]. Available: https://www.mdpi.com/2075-4418/13/3/538.
18. Mobasseri A, Noorifard P (2022) Ultrasound in the diagnosis of pediatric distal radius fractures: does it really change the treatment policy? An orthopedic view. J Ultrasonogr 22(90):e179–e182. https://doi.org/10.15557/jou.2022.0029
19. Mujoomdar M et al. (2014) Optimizing health system use of medical isotopes and other imaging modalities.
20. Burki TK (2018) Shortfall of consultant clinical radiologists in the UK. Lancet Oncol 19(10):e518
21. Adams M, Chen W, Holcdorf D, McCusker MW, Howe PD, Gaillard F (2019) Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. J Med Imaging Radiat Oncol 63(1):27–32
22. Krogue JD et al (2020) Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell 2(2):e190023
23. Badgeley MA et al (2019) Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med 2(1):31
24. Kitamura G, Chung CY, Moore BE (2019) Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. J Digit Imaging 32:672–677
25. Chung SW et al (2018) Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 89(4):468–473

26. Lindsey R et al (2018) Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 115(45):11591–11596
27. Blüthgen C, Becker AS, de Martini IV, Meier A, Martini K, Frauenfelder T (2020) Detection and localization of distal radius fractures: deep learning system versus radiologists. Eur J Radiol 126:108925
28. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT (2019) Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 1(1):e180001
29. Kim D, MacKinnon T (2018) Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 73(5):439–445
30. Raisuddin AM et al (2021) Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep 11(1):6006
31. Liang J, Pan B-C, Huang Y-H and Fan X-Y (2010) Fracture identification of X-ray image. In: 2010 international conference on wavelet analysis and pattern recognition, IEEE, pp 67–73.
32. Smith R, Ward K, Cockrell C, Ha J and Najarian K (2010) Detection of fracture and quantitative assessment of displacement measures in pelvic X-RAY images. In: 2010 IEEE international conference on acoustics, speech and signal processing, 2010: IEEE, pp 682–685
33. Wei Z and Liming Z (2010) Study on recognition of the fracture injure site based on X-ray images. In: 2010 3rd international congress on image and signal processing, vol 4: IEEE, pp 1947–1950
34. Yadav D and Rathor S (2020) Bone fracture detection and classification using deep learning approach. In: 2020 international conference on power electronics & IoT applications in renewable energy and its control (PARC), IEEE, pp 282–285
35. Jones RM et al (2020) Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digit Med 3(1):144
36. Yu F, Koltun V, and Funkhouser T (2017) Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 472–480
37. Tiulpin A, Melekhov I, and Saarakkala S (2019) KNEEL: knee anatomical landmark localization using hourglass networks. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 0–0.
38. Newell A, Yang K, and Deng J (2016) Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, 2016: Springer, pp 483-499
39. Hu, J. Shen L and Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141.
40. Hardalaç F et al (2022) Fracture detection in wrist X-ray images using deep learning-based object detection models. Sensors 22(3):1285
41. Pizer SM, Johnston RE, Ericksen JP, Yankaskas BC and Muller KE (1990) Contrast-limited adaptive histogram equalization: speed and effectiveness. In: [1990] Proceedings of the first conference on visualization in biomedical computing, 22–25 May 1990, pp 337–345, https://doi.org/10.1109/VBC.1990.109340.
42. Ju R-Y, Cai W (2023) Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. Sci Rep 13(1):20077
43. Dibo R, Galichin A, Astashev P, Dylov DV, and Rogov OY (2024) DeepLOC: deep learning-based bone pathology localization and classification in wrist X-ray images. Cham: Springer Nature Switzerland, in analysis of images, Social networks and texts, pp 199–211
44. Wang C-Y, Bochkovskiy A and Liao H-YM (2023) YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7464–7475
45. Olczak J et al (2017) Artificial intelligence for analyzing orthopedic trauma radiographs: deep learning algorithms—are they on par with humans for diagnosing fractures? Acta Orthop 88(6):581–586
46. Heimer J, Thali MJ, Ebert L (2018) Classification based on the presence of skull fractures on curved maximum intensity skull projections by means of deep learning. J Forensic Radiol Imaging 14:16–20
47. Yahalomi E, Chernofsky M, and Werman M (2019) Detection of distal radius fractures trained by a small set of X-ray images and Faster R-CNN. In: Intelligent computing: proceedings of the 2019 computing conference, vol 1: Springer, pp 971–981
48. Abbas W, Adnan SM, Javid MA, Majeed F, Ahsan T, and Hassan SS (2020) Lower leg bone fracture detection and classification using faster RCNN for X-rays images. In: 2020 IEEE 23rd international multitopic conference (INMIC), IEEE, pp 1–6
49. Hržić F, Tschauner S, Sorantin E, Štajduhar I (2022) Fracture recognition in paediatric wrist radiographs: an object detection approach. Mathematics 10(16):2939
50. Erzen EM, BÜtÜn E, Al-Antari MA, Saleh RA, and Addo D (2023) Artificial intelligence computer-aided diagnosis to automatically predict the pediatric wrist trauma using medical X-ray images. In: 2023 7th international symposium on innovative approaches in smart technologies (ISAS), IEEE, pp 1–7
51. Medaramatla SC, Samhitha CV, Pande SD, Vinta SR (2024) Detection of hand bone fractures in X-ray images using hybrid YOLO nas. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3379760

52. Gan K et al (2019) Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthop 90(4):394–400
53. Cheng C-T et al (2019) Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 29(10):5469–5477
54. Nagy E, Janisch M, Hržić F, Sorantin E, Tschauner S (2022) A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning. Sci Data 9(1):222
55. Graham RN, Perriss RW, Scarsbrook AF (2005) Dicom demystified: a review of digital file formats and their use in radiological practice. Clin Radiol 60(11):1133–1140
56. Wang C-Y, Yeh I-H and Liao H-YM (2024) YOLOv9: learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616
57. Huda W, Abrahams RB (2015) X-ray-based medical imaging and resolution. AJR Am J Roentgenol 204(4):W393–W397
58. Huang S-C, Cheng F-C, Chiu Y-S (2012) Efficient contrast enhancement using adaptive gamma correction with weighting distribution. IEEE Trans Image Process 22(3):1032–1041
59. Wang C-Y, Liao H-YM and Yeh I-H (2022) Designing network design strategies through gradient path analysis. arXiv preprint arXiv:2211.04800
60. Yang F, Lu C, Guo Y, Latecki LJ and Ling H (2019) Dually supervised feature pyramid for object detection and segmentation. arXiv preprint arXiv:1912.03730, 2019.
61. Lee Y, Hwang J-W, Lee S, Bae Y and Park J (2019)”An energy and GPU-computation efficient backbone network for real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 0–0
62. Wang C-Y, Liao H-YM, Wu Y-H, Chen P-Y, Hsieh J-W, and Yeh I-H (202) 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, pp 390–391
63. Huang G, Liu Z, Van Der Maaten L and Weinberger KQ (2019) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
64. Lee Y and Park J (2020) Centermask: real-time anchor-free instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13906–13915
65. He K, Zhang X, Ren S, and Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
66. Wang C-Y, Bochkovskiy A, and Liao H-YM (2021) Scaled-yolov4: Scaling cross stage partial network. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp 13029–13038
67. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF (2018) Fracture and dislocation classification compendium—2018. J Orthop Trauma 32:S1–S10

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Promit Basak[1] · Adam Mushtak[2] · Mohamed Ouda[3] · Sadia Farhana Nobi[4] · Anwarul Hasan[5] · Muhammad E. H. Chowdhury[4]

✉ Anwarul Hasan
hasan.anwarul.mit@gmail.com

✉ Muhammad E. H. Chowdhury
mchowdhury@qu.edu.qa

Promit Basak
promit-2017614892@eee.du.ac.bd

Adam Mushtak
amushtak@hamad.qa

Mohamed Ouda
mohamed.ouda@udst.edu.qa

Sadia Farhana Nobi
dr.sadianobi@gmail.com

[1] Department of Electrical and Electronic Engineering, University of Dhaka, Dhaka 1000, Bangladesh

[2] Department of Radiology, Hamad Medical Corporation, Doha, Qatar

[3] Telecommunication and Network Engineering, College of Engineering and Technology, University of Doha for Science and Technology, Doha, Qatar

[4] Department of Electrical Engineering, Qatar University, 2713 Doha, Qatar

[5] Department of Bioengineering, King Fahd University of Petroleum and Minerals, 31261 Dhahran, Saudi Arabia