Comparative analysis of YOLO variants and EfficientNet for detecting bone fractures in X-ray images

Shatabdi Sarker[1], Avizit Roy[1], Shaila Sharmin[1], Shakila Rahman[1], Jia Uddin[2]
[1]Department of Computer Science and Engineering, Faculty of Science and Technology, American International University-Bangladesh, Dhaka, Bangladesh
[2]Department of Artificial Intelligence and Big Data, College of Endicott, Woosong University, Daejeon, Republic of Korea

Article Info

Article history:

Received Jan 29, 2025
Revised Nov 21, 2025
Accepted Jan 1, 2026

Keywords:

Bone fracture
Computer vision
EfficientNet
X-ray image
YOLOv11

ABSTRACT

A bone fracture is a serious medical problem, and accurate and prompt diagnosis is crucial for optimal treatment. This study highlights the progress of automatic bone fracture detection using deep learning (DL) models. A dataset containing 17 different fracture classes was used to train and evaluate the models. The dataset had class imbalance and minor fracture detection challenges. Extensive preprocessing, including data augmentation and resizing, has been applied to solve these problems, which has helped to increase the robustness of the model. Seven state-of-the-art models—you only look once (YOLO)v8, YOLOv9, YOLOv10, YOLOv11, EfficientNetB0, DenseNet169 and ResNet50—are trained and evaluated. Precision, recall, F1-score, and mean average precision (mAP) were used to evaluate the performance of the models. Among all models, YOLOv11 leads the others by achieving the highest precision, mAP, and precision-recall balance. YOLOv11 adds architectural improvements such as a deep backbone network and hybrid feature fusion, which make the model more reliable in different types of fracture detection. It is capable of reducing false detections and maintaining stable memory usage consistency even under different imaging conditions. Overall, YOLOv11 showed promising results and highlighted the potential of AI-powered diagnostic tools to improve clinical processes and patient care. As future work, the application field of the model can be extended to larger medical imaging tasks, and it can be further refined for effective use in resource-limited environments.

Corresponding Author:

Jia Uddin
Department of Artificial Intelligence and Big Data, College of Endicott, Woosong University
Daejeon, 34606, Republic of Korea
Email: jia.uddin@wsu.ac.kr

1. INTRODUCTION

Every day, countless people suffer from bone fractures, which are a common but serious health problem. Accurate fracture diagnosis at the right time is critical, as incorrect or late diagnosis can lead to long-term mobility problems or poor recovery. Currently, X-ray is the most common method of fracture detection. However, it has several limitations. Fracture appearance can vary from patient to patient; the diagnostic value depends entirely on the skill of the radiologist, and many times it may fail to detect small or subtle hairline fractures. This is where artificial intelligence (AI) plays an important role. The you only look once (YOLO) family of deep learning (DL)-based object detection methods is particularly popular, as it is able to detect fractures with high accuracy in real-time [1]. Channel-wise fusion and spatial-wise group attention (CFSG) U-Net-like models have further improved the accuracy of rib fracture detection in computed tomography (CT) scans by adding channel-based and spatial attention processes [2]. Convolutional neural network (CNN)-based methods have shown outstanding performance in medical image analysis in terms of classification, segmentation, and abnormality detection [3]. Advanced computer-aided diagnosis (CAD) systems have made fracture detection more efficient by reducing the workload of radiologists and reducing human error [4]. However, despite significant progress, several challenges remain- especially ensuring consistently reliable performance of AI models across patient heterogeneity and changing imaging modalities is a major problem [5].

Although the use of AI in medical imaging has grown rapidly, several challenges remain in using CAD systems for bone fracture detection. One of these is the uneven quality of the dataset. Variations in X-ray image clarity, annotation accuracy–these factors affect model performance [6]. In particular, poor quality X-ray images prevent effective feature extraction, resulting in loss of information, and subtle fractures are often misclassified [7]. Another major problem is class imbalance in fracture datasets. In this, the models perform well in detecting common or frequent fractures, but show weakness in detecting rare or complex fractures [8]. Limitations in generalization ability are also major challenges, as most AI models are trained on limited, similar datasets, reducing their effectiveness in real-life, diverse medical images [9]. Many AI-based fracture detection models lack transparency, making it difficult for medical professionals to interpret and trust the model output [10]. To solve this problem, researchers are trying to interpret how AI decisions are being made using attention-based visualization methods such as Grad-CAM [11]. Yet AI models that can properly adapt to different types of fractures, different imaging modalities, and diverse patient groups are still a major challenge [12]. In addition, computational cost, hardware limitations, and implementation issues in resource-constrained environments are hindering the widespread adoption of AI technologies [13].

This study presented several important innovations to address the identified challenges:

- First, a curated X-ray dataset containing 17 fracture types is used [14]. The dataset contains well-labeled images in JPG, PNG, and WEBP formats, correctly classified according to fracture type. It is created using publicly available medical images collected from various open sources.
- Second, the study applied advanced pre-processing methods–such as data augmentation and standardization–to improve dataset quality, increase model robustness, and reduce X-ray image noise.
- Third, a comparison between cutting-edge DL models, including DenseNet169, ResNet50, EfficientNetB0, YOLOv8, YOLOv9, YOLOv10, and YOLOv11, was conducted.
- Finally, a comparison of various popular models showed that YOLOv11 gave the best results. Hence, it is chosen as the main model of this study. YOLOv11 proved to be more effective than other models in terms of detection accuracy, fast working ability, and overall stability.

2. LITERATURE REVIEW

DL and AI have made significant improvements in bone fracture detection in medical imaging in recent years. Earlier X-ray analysis depended entirely on the experience of radiologists. This would have allowed for human errors, discrepancies, and delays in diagnosis. But now it is becoming possible to detect fractures much faster and more accurately using CNN, machine learning methods, and object detection models like YOLO or EfficientNet. Early studies have shown that using machine learning methods such as support vector machine, decision tree, and naive Bayes has yielded about 64% to 92% accuracy in fracture classification [15].

These methods were helpful, but they were not scalable and mainly relied on handcrafted features. The development of deep CNNs greatly increased the accuracy of detection. Narrative reviews confirmed that CNN architectures such as ResNet, Inception V3, and Faster R-CNN consistently surpassed conventional diagnostic techniques [16]. Further, CNN-based models such as ResNet-18 have achieved 89.8% validation accuracy with transfer learning in MATLAB on X-ray image classification into fractured or non-fractured categories, although they did not perform well in multi-class fracture detection and small fracture localization problems [17]. Also, parallel DenseNet showed test accuracy up to 74% for anomaly detection of the wrist and forearm. However, the performance decreases for complex fracture patterns [18].

EfficientNet architecture has shown notable promise in fracture detection tasks. For example, EfficientNet-B4 has been applied for vertebral fracture and osteoporosis classification from lateral spine X-rays, achieving area under the receiver operating characteristic curve (AUROC) values of 0.93 and 0.85, outperform standard clinical models, and spinal cord fracture detection using EfficientNet-B4 demonstrated superior AUROC performance for both fracture and osteoporosis identification [19]. Reviews have also emphasized the role of CNNs, U-Net architecture, and transfer learning in improving diagnostic efficiency [20]. Recent work has explored attention-enhanced frameworks. An attention-based cascade R-CNN model, for instance, achieved a mean average precision (mAP) of 0.71 in sternum fracture detection, improving sensitivity for small and concealed fractures [21]. Besides, better results were obtained by adding attention technology to the customized CNN and YOLO models. For example, the YOLOv7 model with focus on the FracAtlas dataset achieved 86.2% mAP, although very fine fractures were still difficult to detect [22].

Currently, the most advanced methods for real-time fracture detection are object detection frameworks such as YOLO. A systematic review showed that the YOLO-based model achieved an accuracy of up to 99% in distal radius fracture detection [23]. In addition, using YOLO in an AI-assisted radiology system reduced report generation time by an average of 27%, demonstrating its effectiveness in clinical workflows [24]. YOLOv5 performed well in cervical spine fracture detection on the RSNA 2022 CT dataset despite its mild weakness. It achieved an overall accuracy of 94% and an AP of 0.98 in normal cases and an AP of 0.96 in fracture cases, although recall was lower in small fractures [25]. YOLOv7 gave high speed and good mAP scores in whole-body fracture detection, but performed poorly in low-contrast X-ray [26]. On the other hand, YOLOv8 provided high precision and recall in real-time detection on multimodal images. However, it requires more computational resources, and a slight decline in mAP@0.95 scores was observed [27].

In addition to radiographs, some new non-invasive methods for bone fracture detection have also been proposed. For example, a microwave imaging system is capable of detecting fine fractures down to 1 mm [28]. Similarly, increased gain and efficiency have been achieved using metamaterial-based antennas for early fracture detection [29]. Also, high-resolution microwave transceivers have been able to detect fractures with sub-millimeter accuracy [30]. These methods show that fracture detection technology is not limited to conventional imaging. The effectiveness of AI in radiology has been highlighted in several reviews. As has been shown, AI often performs better than physicians' results in detecting hip fractures [31]. According to a meta-analysis of 42 studies, the diagnostic sensitivity of AI is comparable to that of radiologists—92% in internal validity and 91% in external validity [32]. In addition, CNN-based models have shown greater than 90% accuracy in skeletal fracture detection and reduced interobserver variability [33]. Some studies have shown that interpretable AI tools such as Grad-CAM are helpful in increasing clinician confidence [34]. However, inconsistent datasets, lack of standardized assessments, and limitations in clinical validity still remain problems. This has emphasized the need for transparent and interpretable systems to make AI acceptable in radiology. Benchmark analysis showed that YOLOv11 and its predecessors demonstrated stability in fracture detection [35]. Recent research has also shown that DL, especially CNN, is capable of demonstrating human-equivalent performance in bone fracture classification and CAD [36].

AI-based methods, especially YOLO and EfficientNet-based CNN, have made fracture detection faster and more accurate. These models can help reduce the workload of radiologists and improve patient care. Nevertheless, future clinical use requires focus on large-scale clinical trials, improving the interpretability of models, and generating standardized datasets. The future of automated bone fracture detection will depend on building AI systems that are both scalable and easy to understand.

3. RESEARCH METHOD

In this section, the detailed research method is presented. There are 5 sub-sections in this methodology. Such as dataset, resizing and labelling, augmentation, data-processing, and model selection.

3.1. Dataset collection and annotation

This study employs the human bone fracture C17 dataset from the Mendeley data repository in Figure 1 and Table 1 [14]. The dataset consists of X-ray images with 17 types of fractures: avulsion, closed, comminuted, compression, dislocated, greenstick, hairline, impacted, intra-articular, longitudinal, oblique, open, pathological, segmental, spiral, stress, and transverse fractures. There are 2,192 data points in the collection, considering all types of fractures. The images are in JPG, PNG, and WEBP formats. Each fracture type is in a separate folder. The dataset combines samples from Kaggle, Radiopaedia, and Shutterstock; hence, it is quite diverse in its conditions of imaging. For better performance and prevention of overfitting, all images were resized and normalized, and different augmentation procedures were performed. Then, the data was divided into training, validation, and test sets. Due to its varied nature and well-organized labeling, it is ideal for the assessment of YOLO variants and EfficientNet.

We labeled the data using the RoboFlow platform, with precise bounding boxes assigned to each fracture location. If there are multiple fractures in an image, a separate bounding box is created for each. Labelers followed quality guidelines, and the job was done correctly. The dataset is then formatted to be usable with YOLO, where the class label and bounding box coordinates for each image are appended to a text file. The dataset is split to ensure sufficient data for training and robust validation: 70% training, 15% validation, and 15% testing.

3.2. Resizing and labelling

As shown in Figure 2, input data is normalized by resizing all images to 640×640 pixels to be compatible with DL models. The dataset was labeled using the RoboFlow annotation tool, which made it easy to create bounding boxes according to class IDs. Figure 2 shows how the data is organized using resized images, annotated bounding boxes, and different color codes. This structured annotation process enables accurate fracture detection in real-life applications and makes model training simple and effective.

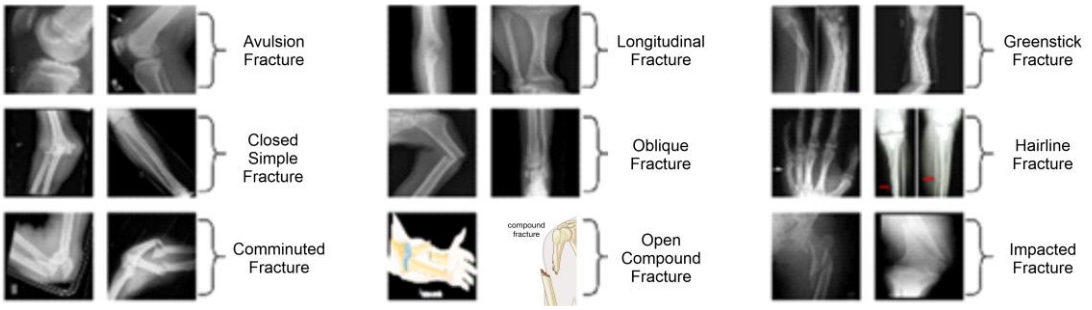


Figure 1. Sample dataset

Table 1. Dataset statistics

Types of fracture	Quantity within each type	Types of fracture	Quantity within each type
Avulsion fracture	141	Intra-articular fracture	104
Closed (simple) fracture	112	Longitudinal fracture	129
Comminuted fracture	219	Oblique fracture	124
Compression-crush fracture	150	Open(compound) fracture	86
Fracture dislocation	159	Pathological fracture	129
Greenstick fracture	136	Segmental fracture	74
Hairline fracture	139	Spiral fracture	134
Impacted fracture	161	Stress fracture	148
Transverse fracture	47		

Figure 2. Sample labeled data

3.3. Augmentation

Table 2 illustrates those various techniques of data augmentation have been used to increase the generalizability of the model and overcome the limitations of small datasets. These methods simulate a variety of imaging settings, allowing the model to perform stably under real-world conditions. Augmentation includes random rotation for slight changes in image angle (±15°), scaling for variation in fracture size (±10%), horizontal and vertical flipping to simulate orientation changes, brightness adjustment (±20%) that reflects changes in X-ray exposure and intensity, and random blur (up to 2 pixels) that simulates motion or focus blur in real images. By using these augmentations randomly during training, the dataset becomes more diverse, which significantly improves the performance and stability of the model.

Table 2. Augmentation techniques and parameters

Techniques	Parameters	Techniques	Parameters
Random rotation	±15°	Vertical flip	p 0.2
Scaling	±10%	Brightness adjustment	±20%
Horizontal flip	p =0.5	Random blur	Up to 2 pixels
Stochastic application	Random during training		

3.4. Data preprocessing

As shown in Figure 3, X-ray images are preprocessed in multiple steps to make them compatible with the selected model, without compromising the diagnostic value. The images are resized to 640×640 pixels for the YOLO model and 224×224 pixels for EfficientNetB0, to match the spatial dimensions of the input to the model. This increases performance in both training and estimation. Image contrast is enhanced to clarify bone structure and fracture lines. Local contrast is enhanced using contrast-limited adaptive histogram equalization (CLAHE), which highlights fine fracture lines within the bone, which are not easily seen on uncontrasted X-ray images.

A Gaussian blur filter is used for noise reduction, which reduces background artifacts and high-frequency noise. This stabilizes feature extraction and helps the model focus more deeply on relevant fracture patterns. Morphological operations and Otsu thresholding were used to separate bone regions from soft tissue. This made the diagnostic analysis more accurate by reducing the influence of irrelevant anatomical features.

Pixel intensity values during training are normalized to the [0, 1] range by dividing by 255, which ensures numerical stability and consistent gradient updates. Images are kept in grayscale to avoid overcomplication and keep key structural information intact. Based on various visual inspections and histogram analysis, this preprocessing pipeline increased the performance and stability of the model while retaining important diagnostic information.

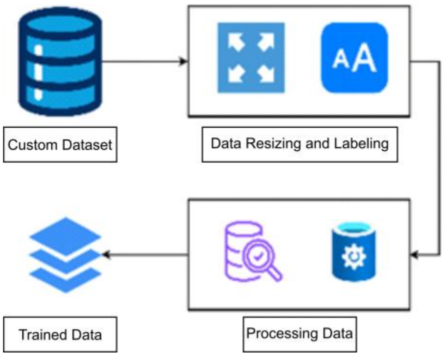


Figure 3. Data preprocessing

3.5. Model selection and architecture

This study evaluated seven advanced DL models–YOLOv8, YOLOv9, YOLOv10, YOLOv11, EfficientNetB0, DenseNet169, and ResNet50–with the objective of determining the best model for accurate and reliable fracture detection in medical X-ray images. To ensure fair and consistent comparisons, all models are trained and evaluated after preprocessing with the same care and applying data augmentation. The evaluation results showed that YOLOv11 achieved the highest accuracy and is considered the most suitable model for this application. In the field of medical diagnostics, where even a small mistake can have a big impact, high accuracy is very important. For this reason, the selection of YOLOv11 can be called logical and justifiable.

YOLOv11 is the latest version of the YOLO series, which brings many architectural improvements. Its advanced backbone network is capable of capturing high-resolution fine features, which is helpful in detecting often-missed fracture lines. Multi-scale feature integration is accomplished through the neck architecture, which combines the path aggregation network and the advanced feature pyramid network. It is very important to identify fractures of different shapes, sizes, and positions.

Moreover, YOLOv11 has an anchor-free detection head, which reduces dependence on predefined anchors. This results in increased bounding box location accuracy and confidence scores, which improves model flexibility and generalizability to different datasets. While high recall is important in medical use, it is equally important to reduce false positives in order to avoid unnecessary treatment or misdiagnosis [35]. Finally, YOLOv11 demonstrated increased accuracy, robustness, and adaptability compared to other models–YOLOv8, YOLOv9, YOLOv10, EfficientNetB0, DenseNet169, and ResNet50. It stands out as a very strong candidate for use in automated fracture detection and clinical decision-making.

4. RESULTS AND DISCUSSION

The training setup and hyperparameters for YOLOv11-based bone fracture detection are shown in Table 3. The model uses the AdamW optimizer, which has an initial learning rate of 0.000476 and a momentum of 0.9. This setup helps to update weights quickly and efficiently over 200 epochs. Pre-trained weights (pre-trained = true) from the COCO dataset are used to speed up training. Also, if validation for 100 epochs does not change the performance, training is stopped by applying early stopping (patience =100).

Automatic mixed precision (AMP) is used to increase computational efficiency. It combines 16-bit and 32-bit floating-point, resulting in less memory and faster training. To maintain learned features, a single convolutional layer (model.23.dfl.conv.weight) is frozen. For regularization, input images are resized to 640×640 pixels with a batch size of 16 and a weight decay of 0.0005. The robustness of datasets is improved by data augmentation methods such as erasing (erasing =0.4), horizontal flipping (fliplr =0.5), and randaugment. Outputs are saved in TorchScript format for deployment, with a maximum detection limit of 300 items per image and overlapping masks managed by a mask ratio of four. Object tracking is configured with Botsort.yaml to ensure precise detection.

Table 3. YOLOv11-based bone fracture detection model hyperparameters

Parameters	Value	Parameters	Value
Batch size	16	AMP	True
Number of epochs	200	Max_det	300
Optimizer	Auto (AdamW selected)	Format	Torch script
Pre-trained weights	True	Tracker	Botsort.yaml
Learning rate(lr0)	0.000476	Auto augment	Randaugment
Momentum	0.9	Overlap mask	Enabled (mask ratio =4)
Patience	100	Erasing	0.4
Image size	640×640 pixels	Flipping(fliplr)	0.5
Weight decay	0.0005	Freezing layer	‘model.23.dfl.conv.weight’

4.1. Model evaluation

Table 4 summarizes an advanced YOLOv11-based DL model which is developed to accurately detect and classify cracks or fractures in bone X-ray images. This model has a total of 319 layers. It contains a total of 9,434,371 parameters, of which 9,434,355 are actually trainable–that is, they can be changed during training to make the model more accurate. These parameters are fine-tuned during training so that the model can understand subtle and complex fracture patterns within X-ray images and can tell very accurately where the fracture is.

The computational cost of the model is only 21.6 giga floating point operations (GFLOPs), meaning it can work quickly without much processing power. It can therefore be used in real-time systems, such as detecting fractures immediately after an X-ray is taken in a hospital. The structure of the model is defined in the ‘YOLO11s.yaml’ file. Here's how each part of the network is laid out, what layers there are, how data flows–all the information. On the other hand, the 'data.yaml' file contains important details of the training and validation datasets–such as which classes are there (eg, fracture, no fracture), where the images are stored, and where the label files are. These two YAML files–one for the architecture and the other for the dataset configuration–together create a well-organized and systematic system for training and evaluating the entire model. That is, the entire process is done in a clean framework, which helps improve model accuracy and performance.

Table 4. Model parameters

Parameters	Value	Parameters	Value
Model layers	319	Gradients	9,434,355
Model parameters	9,434,371	GFLOPs	21.6
Data configuration	Data.yaml	Model configuration	YOLO11s.yaml

4.2. Hardware description

Table 5 illustrates that all tests were performed on a computer with a 10th Gen Intel Core i7-10700 processor, NVIDIA RTX 3050 graphics card with 16 GB DDR4 RAM, and 8 GB memory. The system was running on the Ubuntu 24.04 operating system. This hardware configuration gave us enough power to comfortably train, test, and evaluate the DL models we used.

Table 5. System specification

Component	Specification	Component	Specification
CPU	10th Gen Intel Core i7 10700	RAM	16 GB DDR4
OS	Ubuntu 24.04	GPU	NVIDIA RTX 3050 8 GB

4.3. Analysis of results

To understand the improvement of YOLOv11 in bone fracture detection, it is compared with YOLOv10, YOLOv8, YOLOv9, EfficientNetB0, DenseNet169, and ResNet50 models. The most important comparison was with YOLOv10, as both are newer versions of the same architecture. From the results, which were given in Figure 4, YOLOv11 (Figure 4(a)) showed better behavior than YOLOv10 (Figure 4(b)) in all types of significant loss or loss means during training. For example, box regression loss–which tells how well the model learns fracture locations–dropped from 0.65 to 0.30 for YOLOv10, while YOLOv11 dropped even better from 0.75 to 0.22. This suggests that over time, YOLOv11 has learned the location of bone ruptures more accurately. While classification loss - which tells how well the model can tell if an image has fractures–dropped relatively quickly for YOLOv10 (from 6.0 to 0.7), YOLOv11 dropped relatively slowly but more steadily from 4.0 to 1.0. This means YOLOv11 learns more stably and reliably to recognize subtle and complex fractures. Similar results were seen in distribution focal loss (DFL). YOLOv11 managed to reduce this loss from 1.38 to 0.95, which is similar to YOLOv10's reduction from 1.33 to 0.98 - but slightly better. Overall, the data show that YOLOv11 is not only better than previous models but also learns more accurately, consistently, and reliably in detecting bone fractures.

The results of the validation phase proved the power of YOLOv11 more clearly. YOLOv11 showed that its validation box loss and DFL loss decreased slowly and smoothly–meaning the model learned steadily. But for YOLOv10, these graphs fluctuated, showing that YOLOv10 did not learn as stably and had trouble adapting well to new data. The same thing can be seen for classification loss–YOLOv11 consistently reduced loss, but YOLOv10's graph was much more volatile. This implies that YOLOv11 can more consistently and reliably understand whether there are fractures in the image. YOLOv11 was ahead in detection performance as well. The model showed an improvement in precision from about 0.10 to 0.48, indicating that it reduced the tendency to falsely detect "fractures" (false positives). While YOLOv10 eventually reached the same place in accuracy, YOLOv11 maintained a good balance between correct detection and false alarms throughout training. YOLOv11 was also more stable in terms of recall - how many true fractures the model was able to detect. At first, YOLOv10 showed artificially high recall (because it was saying "fracture" too many places), but later it decreased. In contrast, YOLOv11 maintains a constant recall of around 0.45 throughout, which is practically more meaningful and reliable.

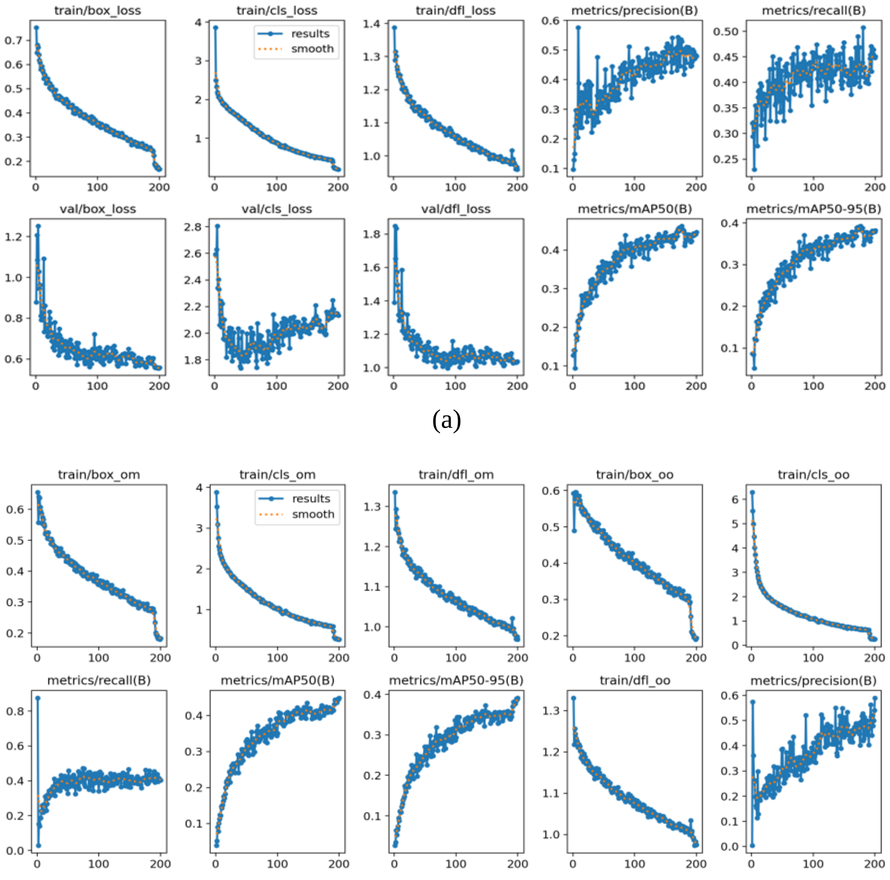
The true power of YOLOv11 is most clearly seen in the mAP results. Here, YOLOv11's mAP@50 value increased from 0.15 to 0.52, while YOLOv10 only managed to rise from 0.15 to 0.45. That is, under the same conditions, YOLOv11 learns to detect fractures much more accurately. Similar results were obtained by looking at the more stringent evaluation metric mAP@50-95. While both models improved, YOLOv11's value increased from 0.08 to 0.38–marginally surpassing YOLOv10's increase from 0.07 to 0.38. This means that at different intersection over union (IoU) thresholds (i.e., different stiffness criteria), YOLOv11 is able to detect fractures slightly more accurately. Overall, the results show that YOLOv11 can adapt to new data better than previous versions (better generalization). Its bounding box estimation (location of fracture) is more accurate, and its overall detection capability is also more powerful. Therefore, YOLOv11 is a more reliable and effective choice for automatic bone fracture detection in real hospitals or medical imaging systems.

4.4. Evaluation metrics

Various important metrics, including precision, recall, F1-score, and precision-recall (PR) curve, are used to understand how well the YOLOv11 model performs in bone fracture detection. The model showed a precision of 0.99 at the 1.0 confidence level, which means that with very high confidence, the model can accurately detect fractures with almost no error. However, lowering the confidence level increases false positives, thereby decreasing precision. On the other hand, the recall, i.e., how many real fractures the model can detect, reaches a maximum of 0.70 when the confidence level is 0.0, and the recall gradually decreases as the confidence increases. The F1-score graph shows that the best balance between precision and recall is obtained at a confidence level of 0.690, while the highest F1-score was 0.46. The performance of the model is not equal for different types of erosion. For example, the F1-score in the "closed simple fracture" class was very good, but in some classes the results were poor, which suggests that there were too many features or that there was relatively little data. Average accuracy (AP) was measured by PR curve, where "closed-simple-fracture" and "compression-crush-fracture" classes obtained 0.855 and 0.819 AP, respectively, showing very good detection ability. In contrast, the AP in the "longitudinal-fracture" and "transverse-fracture" classes were only 0.095 and 0.100, indicating that such fractures were quite difficult to detect. Finally, the overall mAP of the model at the 0.5 IoU threshold was 0.462. This shows that performance is fairly good, but there is still room for improvement, especially for difficult and underrepresented fracture classes.

The comparative results in Tables 6 and 7 show that there is some trade-off between precision and recall among the different tested models. The proposed YOLOv11 model achieves a very high accuracy, i.e., 0.99, while EfficientNetB0 has an accuracy of 0.94, which shows that these models are less likely to falsely identify non-broken bones as fractures. This is very important from a clinical point of view, as additional misdiagnoses can lead to unnecessary patient treatment, additional suffering, and increased diagnostic costs. Hence, the high accuracy of YOLOv11 ensures more reliable and effective detection in medical systems.

In fracture detection, 'recall' refers to how efficiently the model can detect actual fractures. The recall of ResNet50 (0.43) and YOLOv10 (0.58) is relatively low, which means that these models can often miss fractures. False negative results are particularly dangerous because if a fracture is present, if it goes undetected, the patient's treatment may be delayed, the bone may not rotate properly, or long-term functional loss may occur.



(a)

(b)

Figure 4. Training graph with 200 epochs based on (a) YOLO11 and (b) YOLOv10

Table 6. Testing performance of YOLOv11 with YOLOv10, YOLOv9, YOLOv8 and EfficientNetB0

Model	Epoch	Class	Trainable parameters	Precision	Recall	F1-score	mAP@0. 5
YOLOv10	100	All	2.71M	0.80	0.58	0.37	0.347
	200			0.92	0.78	0.45	0.447
YOLOv9	100	All	25.45M	0.86	0.70	0.36	0.352
	200			0.94	0.80	0.44	0.452
YOLOv8	100	All	25.86M	0.85	0.60	0.44	0.472
	200			0.94	0.70	0.50	0.488
EfficientNetB0	100	All	4.02M	0.65	0.63	0.63	0.700
	200			0.94	0.93	0.93	0.979
Proposed YOLOv11	100	All	9.43M	0.90	0.60	0.38	0.352
	200			0.99	0.70	0.46	0.462

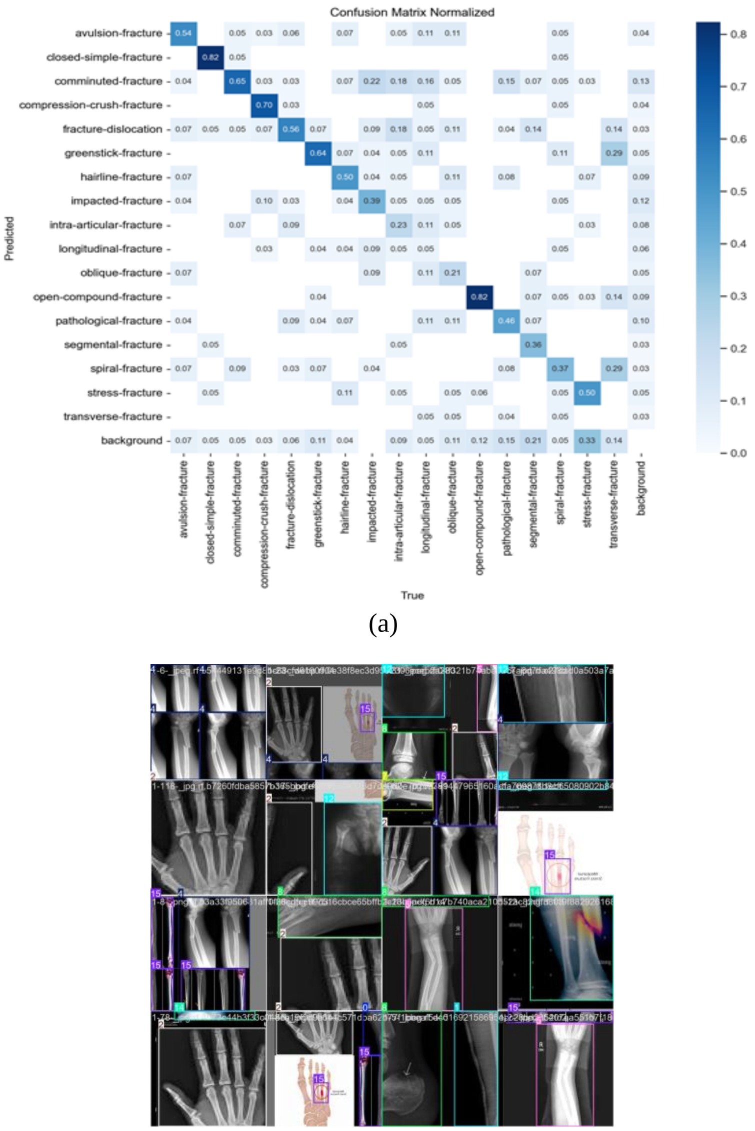
Table 7. Comparative analysis of YOLOv11 with DenseNet169 and ResNet50

Model	Epoch	Trainable parameters	Precision	Recall	F1-score	mAP@0. 5
Proposed(YOLOv11)	200	0.43M	0.99	0.70	0.46	0.462
DenseNet169	200	12.51M	0.56	0.56	0.55	0.575
ResNet50	200	23.57M	0.43	0.43	0.42	0.402

On the other hand, the YOLOv11 model shows high precision (0.99) and moderate recall (0.70), indicating a good balance. This means the model is able to detect most fractures, and the rate of false positives is low. Although EfficientNetB0 shows a recall of 0.93, its low precision increases the possibility of false positives, i.e., sometimes the model may falsely report a fracture when there is no fracture. In summary, reducing false negatives is paramount in fracture detection, as the clinical risk of missing a fracture, even if present, is high. In this respect, YOLOv11 may prove to be a safe and reliable system for clinical use due to its high precision and good recall.

4.5. Visualization

The multi-class fracture detection performance of the YOLOv11-based model is shown in Figure 5. Figure 5(a) shows a confusion matrix, showing the correct and incorrect predictions of the model for 17 fracture types. Entries in originally diagonal lines are correct predictions of the model, and dashed lines indicate misclassifications. Some classes, such as 'spiral fracture' and 'segmental fracture', are more challenging to identify, but the model showed good accuracy in classes such as 'compression-crush fracture' (0.70) and 'open compound fracture' (0.62). Figure 5(b) shows the fracture localization of the model in the X-ray image. Here, the bounding boxes and confidence scores indicate the correct identification of different fracture types. These visualizations show that the model is not only able to detect fractures, but also pinpoint exactly where they are. On the one hand, it shows the localization capability; on the other hand, it highlights the power of the model in classification and the opportunity to reduce misidentification.



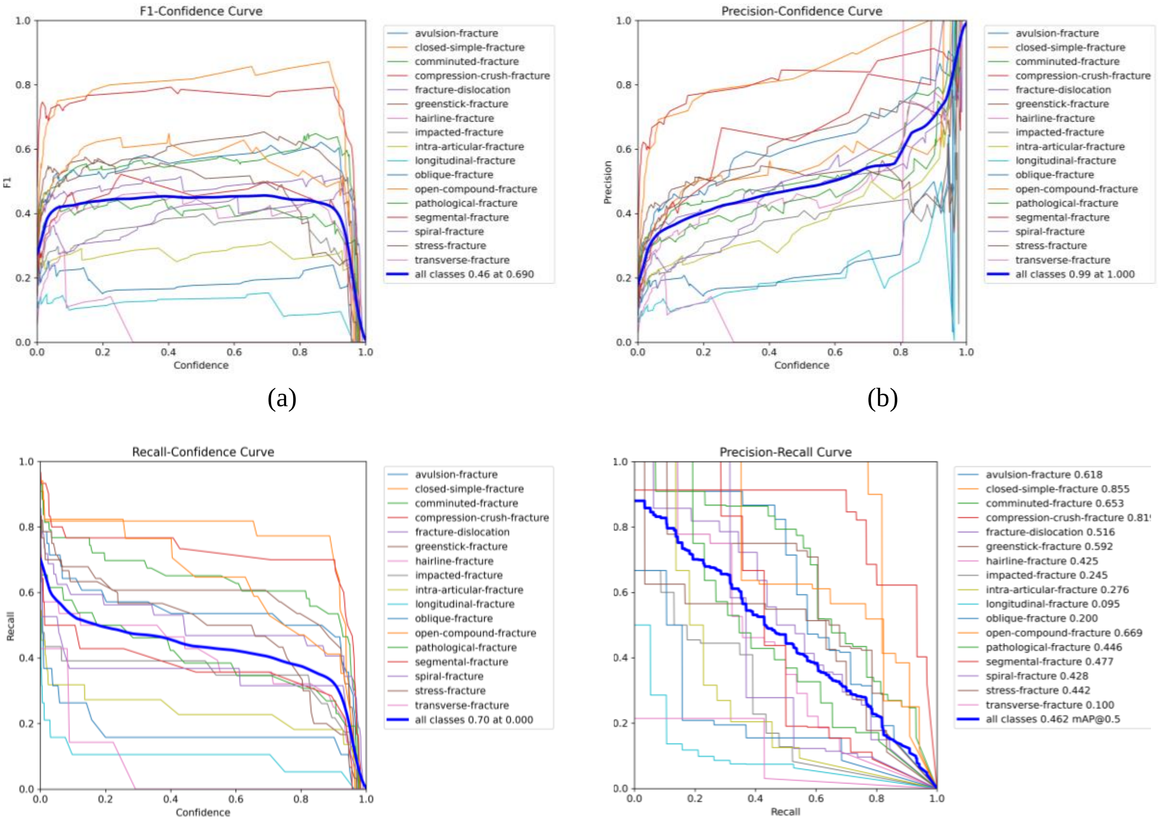
Predicted \ True	avulsion-fracture	closed-simple-fracture	comminuted-fracture	compression-crush-fracture	fracture-dislocation	greenstick-fracture	hairline-fracture	impacted-fracture	intra-articular-fracture	longitudinal-fracture	oblique-fracture	open-compound-fracture	pathological-fracture	segmental-fracture	spiral-fracture	stress-fracture	transverse-fracture	background
avulsion-fracture	0.54		0.05	0.03	0.06		0.07		0.05	0.11	0.11				0.05			0.04
closed-simple-fracture		0.82	0.05												0.05			
comminuted-fracture	0.04		0.65	0.03	0.03		0.07	0.22	0.18	0.16	0.05		0.15	0.07	0.05	0.03		0.13
compression-crush-fracture				0.70	0.03					0.05					0.05			0.04
fracture-dislocation	0.07	0.05	0.05	0.07	0.56	0.07		0.09	0.18	0.05	0.11		0.04	0.14			0.14	0.03
greenstick-fracture						0.64	0.07	0.04	0.05	0.11					0.11		0.29	0.05
hairline-fracture	0.07						0.50	0.04	0.05		0.11		0.08			0.07		0.09
impacted-fracture	0.04			0.10	0.03		0.04	0.39	0.05	0.05	0.05				0.05			0.12
intra-articular-fracture			0.07		0.09				0.23	0.11	0.05					0.03		0.08
longitudinal-fracture				0.03		0.04	0.04	0.09	0.05	0.05					0.05			0.06
oblique-fracture	0.07							0.09		0.11	0.21			0.07				0.05
open-compound-fracture						0.04						0.82		0.07	0.05	0.03	0.14	0.09
pathological-fracture	0.04				0.09	0.04	0.07			0.11	0.11		0.46	0.07				0.10
segmental-fracture		0.05							0.05					0.36				0.03
spiral-fracture	0.07		0.09		0.03	0.07		0.04					0.08		0.37		0.29	0.03
stress-fracture		0.05					0.11		0.05		0.05	0.06			0.05	0.50		0.05
transverse-fracture										0.05	0.05		0.04		0.05			0.03
background	0.07	0.05	0.05	0.03	0.06	0.11	0.04		0.09	0.05	0.11	0.12	0.15	0.21	0.05	0.33	0.14	

(a)

(b)

Figure 5. YOLOv11 of (a) confusion matrix and (b) testing performance of random data

Figure 6 shows how well the proposed YOLOv11 model performed for 17 fracture types. Figure 6(a) shows that the model produces very accurate results at high confidence, indicating low false positives. The F1-confidence curve in Figure 6(b) shows an overall F1-score of about 0.46, indicating a balanced performance of the model. The recall-confidence curve in Figure 6(c) shows that the model's recall is about 0.70, and the recall decreases with increasing confidence–meaning that some fractures may be missed while predicting more reliably. The mAP@0.5 value in Figure 6(d) is 0.462, which shows a respectable balance between precision and recall. In summary, the YOLOv11 model is reliable and stable in detecting different types of fractures, which makes it suitable for automatic fracture detection in medical X-ray images.



(a) (b)

(c) (d)

Figure 6. YOLOv11-based of (a) precision curve, (b) F1-score curve, (c) recall curve, and (d) mAP curve

Figure 7 compares the prediction performance of five models–YOLOv11, YOLOv10, YOLOv9, YOLOv8, and EfficientNetB0–on a 17-class fracture detection dataset. Each row shows a fracture class, and each column expresses the prediction performance of the model. The results show that YOLOv11 outperforms the other models in all classes and achieves the highest overall accuracy. This proves that YOLOv11 is particularly capable of detecting complex fractures, especially in terms of precision and recall.

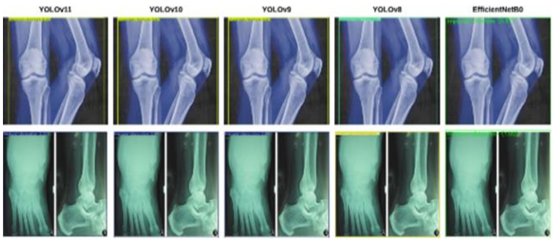


Figure 7. Sample detected images for YOLOv11, YOLOv10, YOLOv9, YOLOv8, and EfficientNetB0 respectively

5. CONCLUSION

This study introduces the YOLOv11 model for automatic bone fracture detection and compares its performance with EfficientNetB0, YOLOv8, YOLOv9, YOLOv10, DenseNet169, and ResNet50. The results show that YOLOv11 is more accurate, precise, and has a higher recall value, that is, it is faster and more reliable in detecting fractures. The model is more generalizable as the number of fracture classes increases, making it suitable for practical medical imaging. This saves radiologists time, reduces misdiagnosis, and increases bone fracture detection rates. When connected to hospital information systems or mobile diagnostic units, the model can provide rapid assessment and immediate triage even in emergency or resource-limited situations. YOLOv11 can potentially bring a tremendous impact to improve the diagnostic efficiency of radiologists, unburden the workload, and widen the use of AI in medical imaging, leading to better outcomes for the patients and thereby healthcare services. Even so, the dataset might not be representative of various clinical settings, and rare fracture types might require more work. Its deployment also faces hardware limitations in low-resource environments. Future studies will also focus on integrating the system with picture archiving and communication systems (PACS) for smooth integration into clinical workflow, as well as enhancing real-time detection capabilities and expanding to support multimodal imaging such as CT and magnetic resonance imaging (MRI).

FUNDING INFORMATION

This research is funded by Woosong University Academic Research 2026.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Shatabdi Sarker	✓			✓	✓	✓	✓	✓	✓	✓	✓		✓	
Avizit Roy		✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	
Shaila Sharmin	✓						✓			✓				
Shakila Rahman	✓			✓	✓					✓		✓	✓	
Jia Uddin					✓	✓				✓		✓		✓

C	:	**C**onceptualization	I	:	**I**nvestigation	Vi	:	**Vi**sualization
M	:	**M**ethodology	R	:	**R**esources	Su	:	**Su**pervision
So	:	**So**ftware	D	:	**D**ata Curation	P	:	**P**roject administration
Va	:	**Va**lidation	O	:	Writing - **O**riginal Draft	Fu	:	**Fu**nding acquisition
Fo	:	**Fo**rmal analysis	E	:	Writing - Review & **E**diting			

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the Mendeley Data Repository under the title “Human bone fracture C17 dataset” at https://doi.org/10.17632/2J8VVZ3J6V.1, reference number [14].

REFERENCES

[1] C. T. Chien, R. Y. Ju, K. Y. Chou, and J. S. Chiang, “YOLOv9 for fracture detection in pediatric wrist trauma X-ray images,” *Electronics Letters*, vol. 60, no. 11, Jun. 2024, doi: 10.1049/ELL2.13248.

[2] Z. Zhou, Z. Fu, J. Jia, and J. Lv, “Rib fracture detection with dual-attention enhanced U-Net,” *Computational and Mathematical Methods in Medicine*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/8945423.

[3] Y. Ma and Y. Luo, “Bone fracture detection through the two-stage system of crack-sensitive convolutional neural network,” *Informatics in Medicine Unlocked*, vol. 22, Jan. 2021, doi: 10.1016/j.imu.2020.100452.

[4] A. Alghaithi and S. A. Maskari, “Artificial intelligence application in bone fracture detection,” *Journal of Musculoskeletal Surgery and Research*, vol. 5, no. 1, Jan. 2021, doi: 10.4103/JMSR.JMSR_132_20.

[5] D. Joshi and T. P. Singh, “A survey of fracture detection techniques in bone X-ray images,” *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4475–4517, Jan. 2020, doi: 10.1007/s10462-019-09799-0.

[6] S. H. Kong *et al.*, “A novel fracture prediction model using machine learning in a community-based cohort,” *JBMR Plus*, vol. 4, no. 3, Mar. 2020, doi: 10.1002/jbm4.10337.

[7] M. A. Kassem, S. M. Naguib, H. M. Hamza, M. M. Fouda, M. K. Saleh, and K. M. Hosny, “Explainable transfer learning-based deep learning model for pelvis fracture detection,” *International Journal of Intelligent Systems*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/3281998.

[8] X. Zhang *et al.*, “Diagnostic accuracy and potential covariates of artificial intelligence for diagnosing orthopedic fractures: a systematic literature review and meta-analysis,” *European Radiology*, vol. 32, no. 10, pp. 7196–7216, Jun. 2022, doi: 10.1007/s00330-022-08956-4.

[9] J. R. Lex, J. D. Michele, R. Koucheki, D. Pincus, C. Whyne, and B. Ravi, “Artificial intelligence for hip fracture detection and outcome prediction: a systematic review and meta-analysis,” *JAMA Network Open*, vol. 6, no. 3, Mar. 2023, doi: 10.1001/jamanetworkopen.2023.3391.

[10] M. Kutbi and M. Kutbi, “Artificial intelligence-based applications for bone fracture detection using medical images: a systematic review,” *Diagnostics*, vol. 14, no. 17, Aug. 2024, doi: 10.3390/diagnostics14171879.

[11] K. Suen, R. Zhang, and N. Kutaiba, “Accuracy of wrist fracture detection on radiographs by artificial intelligence compared to human clinicians: a systematic review and meta-analysis,” *European Journal of Radiology*, vol. 178, Sep. 2024, doi: 10.1016/j.ejrad.2024.111593.

[12] M. Kraus, R. Anteby, E. Konen, I. Eshed, and E. Klang, “Artificial intelligence for X-ray scaphoid fracture detection: a systematic review and diagnostic test accuracy meta-analysis,” *European Radiology*, vol. 34, no. 7, pp. 4341–4351, Dec. 2023, doi: 10.1007/s00330-023-10473-x.

[13] A. Almigdad, A. Mustafa, S. Alazaydeh, M. Alshawish, M. B. Mustafa, and H. Alfukaha, “Bone fracture patterns and distributions according to trauma energy,” *Advances in Orthopedics*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/8695916.

[14] S. Sarker, A. Roy, S. Sharmin, S. Rahman, and J. Uddin, *Human Bone Fracture C17 Dataset*. Mendeley Data, 2025, doi: 10.17632/2J8VVZ3J6V.1.[Online]. Available: https://data.mendeley.com/datasets/2j8vvz3j6v/1

[15] K. D. Ahmed and R. Hawezi, “Detection of bone fracture based on machine learning techniques,” *Measurement: Sensors*, vol. 27, p. 100723, Jun. 2023, doi: 10.1016/j.measen.2023.100723.

[16] P. H. S. Kalmet *et al.*, “Deep learning in fracture detection: a narrative review,” *Acta Orthopaedica*, vol. 91, no. 2, pp. 215–220, Jan. 2020, doi: 10.1080/17453674.2019.1711323.

[17] B. S. Reddy, P. T. Chand, and C. Rajesh, “Automated detection of fractures in X-ray images using ResNet-18 and transfer learning in MATLAB,” *Medical Image Analysis*, Jun. 2025.

[18] S. Güçlü, D. Özdemir, and H. M. Saraoğlu, “A new DenseNet-based anomaly detection method using humerus and shoulder X-ray images,” *Traitement du Signal*, vol. 42, no. 2, pp. 851–864, Apr. 2025, doi: 10.18280/ts.420222.

[19] N. Hong *et al.*, “Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray radiography,” *Journal of Bone and Mineral Research*, vol. 38, no. 6, pp. 887–895, Jun. 2023, doi: 10.1002/jbmr.4814.

[20] T. Meena and S. Roy, “Bone fracture detection using deep supervised learning from radiological images: a paradigm shift,” *Diagnostics*, vol. 12, no. 10, Oct. 2022, doi: 10.3390/diagnostics12102420.

[21] Y. Jia, H. Wang, W. Chen, Y. Wang, and B. Yang, “An attention-based cascade R-CNN model for sternum fracture detection in X-ray images,” *CAAI Transactions on Intelligent Technology*, vol. 7, no. 4, pp. 658–670, Dec. 2022, doi: 10.1049/cit2.12072.

[22] J. Zou and M. R. Arshad, “Detection of whole body bone fractures based on improved YOLOv7,” *Biomedical Signal Processing and Control*, vol. 91, May 2024, doi: 10.1016/j.bspc.2024.105995.

[23] K. D. O. Nijhuis *et al.*, “AI for detection, classification and prediction of loss of alignment of distal radius fractures: a systematic review,” *European Journal of Trauma and Emergency Surgery*, vol. 50, no. 6, pp. 2819–2831, 2024, doi: 10.1007/s00068-024-02557-0.

[24] A. L. Mastro *et al.*, “Artificial intelligence in fracture detection on radiographs: a literature review,” *Japanese Journal of Radiology*, vol. 43, no. 4, pp. 551–585, Apr. 2025, doi: 10.1007/s11604-024-01702-4.

[25] H. Patel, M. Rahevar, K. Maheriya, and A. Shah, “Optimizing YOLOv5 for automated detection of cervical spine fractures in CT scans images,” in *2024 OPJU International Technology Conference on Smart Computing for Innovation and Advancement in Industry 4.0 (OTCON)*, 2024, doi: 10.1109/OTCON60325.2024.10687867.

[26] D. Ari and M. Burukanli, “Real-time human bone fracture detection using YOLO models,” *ASES IX. International Scientific Research Congress*, Adiyaman University, Adiyaman, Turkiye, 2025.

[27] K. C. Santos, C. A. Fernandes, and J. R. Costa, “Feasibility of bone fracture detection using microwave imaging,” *IEEE Open Journal of Antennas and Propagation*, vol. 3, pp. 836–847, 2022, doi: 10.1109/OJAP.2022.3194217.

[28] M. S. Hossen *et al.*, “DNG metamaterial-inspired slotted stub antenna with enhanced gain, efficiency, and distributed current for early stage bone fracture detection applications,” *IEEE Sensors Journal*, vol. 24, no. 22, pp. 37932–37946, 2024, doi: 10.1109/JSEN.2024.3459794.

[29] A. N. Moqadam and R. Kazemi, “High-resolution imaging of narrow bone fractures with a novel microwave transceiver sensor utilizing dual-polarized RIS and SRR array antennas,” *IEEE Sensors Journal*, vol. 23, no. 24, pp. 30335–30344, Dec. 2023, doi: 10.1109/JSEN.2023.3328240.

[30] Y. Cha, J. T. Kim, C. H. Park, J. W. Kim, S. Y. Lee, and J. Il Yoo, “Artificial intelligence and machine learning on diagnosis and classification of hip fracture: systematic review,” *Journal of Orthopaedic Surgery and Research*, vol. 17, no. 1, Dec. 2022, doi: 10.1186/s13018-022-03408-7.

[31] R. Y. L. Kuo *et al.*, “Artificial intelligence in fracture detection: a systematic review and meta-analysis,” *Radiology*, vol. 304, no. 1, pp. 50–62, Jul. 2022, doi: 10.1148/radiol.211785.

[32] Z. Su, A. Adam, M. F. Nasrudin, M. Ayob, and G. Punganan, “Skeletal fracture detection with deep learning: a comprehensive review,” *Diagnostics*, vol. 13, no. 20, Oct. 2023, doi: 10.3390/diagnostics13203245.

[33] A. Tieu *et al.*, “The role of artificial intelligence in the identification and evaluation of bone fractures,” *Bioengineering (Basel)*, vol. 11, no. 4, Apr. 2024, doi: 10.3390/bioengineering11040338.

[34] C. Rainey, J. McConnell, C. Hughes, R. Bond, and S. McFadden, “Artificial intelligence for diagnosis of fractures on plain radiographs: a scoping review of current literature,” *Intelligence-Based Medicine*, vol. 5, Jan. 2021, doi: 10.1016/j.ibmed.2021.100033.

[35] N. Jegham, C. Y. Koh, M. Abdelatti, and A. Hendawi, “Evaluating the evolution of YOLO (you only look once) models: a comprehensive benchmark study of YOLO11 and its predecessors,” *arXiv:2411.00201*, 2024.

[36] L. Tanzi *et al.*, “X-ray bone fracture classification using deep learning: a baseline for designing a reliable approach,” *Applied Sciences*, vol. 10, no. 4, Feb. 2020, doi: 10.3390/app10041507.

BIOGRAPHIES OF AUTHORS

Shatabdi Sarker is a student in the Department of Computer Science and Engineering, Faculty of Science and Technology, American International University-Bangladesh, with a steady academic performance. Her research interests include machine learning and deep learning. She can be contacted at email: shatabdisarker1357@gmail.com.

Avizit Roy is a student in the Department of Computer Science and Engineering, Faculty of Science and Technology, American International University-Bangladesh, with a steady academic performance. His research interests include machine learning and deep learning. He can be contacted at email: contact@avizitrx.com.

Shaila Sharmin is a student in the Department of Computer Science and Engineering, Faculty of Science and Technology, American International University-Bangladesh, with a steady academic performance. Her research interests include machine learning and deep learning. She can be contacted at email: shailarichi76952@gmail.com.

Shakila Rahman is a lecturer in the Department of Computer Science and Engineering, Faculty of Science and Technology, American International University-Bangladesh (AIUB). She received an M.Sc. in AI and Computer Engineering from the University of Ulsan, South Korea, and completed a B.Sc. in Computer Science and Engineering from IIUC, Bangladesh. Her research interests include machine learning, artificial intelligence, image processing, optimization algorithms, and wireless sensor networks. She can be contacted at email: shakila.rahman@aiub.edu.

Jia Uddin is an associate professor at the Department of AI and Big Data, Endicott College, Woosong University, Korea. He received a Ph.D. in Computer Engineering from the University of Ulsan, Korea, and an M.Sc. In Telecommunications, from the Blekinge Institute of Technology, Sweden, and a B.Sc. in Computer and Communication Engineering, from the International Islamic University Chittagong, Bangladesh. He was a visiting faculty at the School of Computing, Staffordshire University, United Kingdom, Telkom University, Indonesia, and the University of Foggia, Italy. He was an associate professor at the Department of CSE, Brac University. His research interests are fault diagnosis using sensor data. He can be contacted at email: jia.uddin@wsu.ac.kr.