MAUN Fen Bil. Dergi., 12, 2, 64-71 Araştırma Makalesi/ Research Article
MAUN J. of Sci., 12, 2, 64-71 DOI: 10.18586/msufbd.1440119

Analyzing the Performance of Convolutional Neural Networks and Transformer Models in Automated Bone Fracture Detection

Ece BİNGÖL[2], Semih DEMİREL[1,3]✉, Ataberk URFALI[1,4], Ömer Faruk BOZKIR[1], Azer ÇELİKTEN[1,5], Abdulkadir BUDAK[1], Hakan KARATAŞ[1]

[1] Department of Artificial Intelligence and Image Processing, Akgun Computer Inc., Ankara, Türkiye
[2] Department of Information Systems Engineering, Faculty of Engineering, Atılım University, Ankara, Türkiye
[3] Department of Information Systems, Graduate School of Informatics, Gazi University, Ankara, Türkiye
[4] Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Konya Technical University, Konya, Türkiye
[5] Department of Computer Engineering, Graduate School of Science, Ege University, İzmir, Türkiye

✉: semih.demirel@akgun.com.tr [1]0009-0006-7615-1392 [2]0000-0002-3454-3631 [3]0000-0001-5709-6718
[4]0000-0002-3696-3613 [5]0000-0002-6804-737X [6]0000-0002-0328-6783 [7]0000-0002-9497-5444

Received (Geliş): 20.02.2024 Revision (Düzeltme): 20.06.2024 Accepted (Kabul): 15.08.2024

ABSTRACT

The most significant component of the skeletal and muscular system, whose function is vital to human existence, are the bones. Breaking a bone might occur from a specific hit or from a violent rearward movement. In this study, bone fracture detection was performed using convolutional neural network (CNN) based models, Faster R-CNN and RetinaNet, as well as a transformer-based model, DETR (Detection Transformer). A detailed investigation was conducted using different backbone networks for each model. This study's primary contributions are a methodical assessment of the performance variations between CNN and transformer designs. Models trained on an open-source dataset consisting of 5145 images were tested on 750 test images. According to the results, the RetinaNet/ResNet101 model exhibited superior performance with a 0.901 mAP50 ratio compared to other models. The obtained results show promising outcomes that the trained models could be utilized in computer-aided diagnosis (CAD) systems.

Keywords: Deep learning, medical imaging, object detection, convolutional neural network, vision transformer

Otomatik Kemik Kırığı Tespitinde Evrişimsel Sinir Ağları ve Transformer Modellerinin Performansının Analizi

ÖZ

İnsan varlığı için hayati önem taşıyan iskelet ve kas sisteminin en önemli bileşeni kemiklerdir. Bir kemiğin kırılması belirli bir darbeden veya şiddetli bir geriye doğru hareketten kaynaklanabilir. Bu çalışmada, kemik kırığı tespiti, evrişimli sinir ağı (ESA) tabanlı modeller olan Faster R-CNN ve RetinaNet, ayrıca bir transformer tabanlı model olan DETR (Detection Transformer) kullanılarak gerçekleştirilmiştir. Her model için farklı omurga ağları kullanılarak detaylı bir inceleme yapılmıştır. Bu çalışmanın birincil katkıları, CNN ve transformatör tasarımları arasındaki performans farklılıklarının yöntemsel bir değerlendirmesidir. 5145 görüntüden oluşan açık kaynaklı bir veri setinde eğitilen modeller, 750 test görüntüsünde test edilmiştir. Sonuçlara göre, RetinaNet/ResNet101 modeli diğer modellere göre daha üstün performans sergileyerek 0.901 mAP50 oranına ulaşmıştır. Elde edilen sonuçlar, eğitilen modellerin bilgisayar destekli tanı (BDT) sistemlerinde kullanılabilecek vaat edici sonuçlar sunmaktadır.

Anahtar Kelimeler: Derin öğrenme, medikal görüntüleme, nesne tespiti, evrişimsel sinir ağları, görüş transformatörü

INTRODUCTION

A bone is among the most fundamental components that make up our body's support and mobility system [1]. Bones have many tasks such as keeping our body upright, protecting internal organs, and walking. An adult human's body contains 206 bones in total [2]. Bones can be broken as a result of falling, trauma, or impact. Bone fractures are manually detected by radiologists after x-ray images are taken, and the result is transmitted to the orthopedic doctor. Manual detection can be time-consuming. Also, sometimes radiologists or orthopedic doctors may not be available in emergency departments late at night. In this case, other doctors can be less capable to identify minor fractures due to their lack of specialization and may prolong the patient's recovery process by applying the wrong treatment to the patient [3]. Additionally, early diagnosis plays an important role

in the selection of appropriate treatment methods and the success of treatment [4]. Computer-aided diagnosis also helps minimize human errors during treatment [5].

Previously, the usefulness of diagnostic computers for bone fracture detection was very limited in practical clinical settings, mainly due to its low accuracy. Recently, thanks to the availability of extensive, annotated image datasets, many new studies on CAD based on deep learning have been presented [6]. CNNs used in image analysis are an important network within the scope of deep learning [7]. CNNs represent a specialized branch of artificial neural networks designed particularly for tasks involving visual data processing, such as image recognition and analysis. CNN algorithms can perform object detection and classification on medical images as well as on many images. In such cases, they can be utilized to comprehend the content of the image or determine the location of objects by extracting features from complex structures within the image. In the study conducted by Ozdemir et al. [8], CNN models were comprehensively compared on augmented images, demonstrating the success of CNN models.

Faster R-CNN [9] and RetinaNet [10] are common architectures of CNNs used for object detection. Faster R-CNN performs object detection in two stages, first suggesting the regions where the objects may be, and then the location is determined from the suggested region [9]. In RetinaNet, focal loss [10] solves the class imbalance problem and detects objects of different sizes using anchor boxes [10]. The differences between Faster R-CNN and RetinaNet are that Faster R-CNN uses regional proposals, and RetinaNet uses anchor boxes. In addition, cross-entropy is used as a generality loss function in Faster R-CNN, and focal loss is used in RetinaNet. Vision transformer [11] is used for object identification inspired by the transformer [12] architecture. DETR [13] is a model based on the transformer architecture used for object detection. DETR also uses the transformer structure that performs all object detection at once.

In this study, CNN-based architectures Faster R-CNN and RetinaNet, and the DETR model based on transformer architecture, were used. These models were employed to detect the fractured region in broken bone images in x-ray images and compared to measure which model performs better.

The primary findings of this research include:

- Our study comprehensively examines the advantages and disadvantages of CNN-based architectures compared to transformer-based architectures.
- A comprehensive comparison of Faster R-CNN, RetinaNet, and DETR models was conducted.
- Our comprehensive study provides a different approach to bone fracture detection, contributing significantly to the development of computer-aided diagnostic systems.

The remaining portions of this study's content are categorized into the following sections: In the literature review section, existing studies are examined. The material and method section contains the models that were utilized in this investigation. Results and discussion include performance metrics, model experimental results, and a discussion of the findings. Future research and a conclusion are provided in the section conclusion.

LITERATURE REVIEW

Numerous studies have been conducted using state-of-the-art models of convolutional neural networks in the literature. Some studies have focused on bone fracture detection.

In the study by Warin et al. [14], a total of 1710 mandibular images were used, and 855 of these images had fractures. First, the images were classified in binary as fracture and nonfracture. DenseNet-169 [15] and ResNet50 models were used for classification. Meanwhile, the outcomes that the specialists discovered, and residents were compared. According to the results, DenseNet-169 performed 100% classification. In broken bone detection, the Faster R-CNN model outperformed the YOLOv5 model with an f1 score of 90.67%.

Kim et al. [16] proposed an assessment method utilizing a stacked autoencoder (SAE) for bone fracture investigation that builds upon prior research in the fields of unsupervised learning, medical imaging, and structural health monitoring. This novel approach aligns with recent efforts to enhance the accuracy and efficiency of fracture analysis without relying solely on traditional imaging techniques. Additionally, the use of virtual spectrograms and a short-time Fourier transform in image-based training signifies a departure from conventional methodologies.

The study conducted by Tao et al. [17] applied an automated segmentation method in medical imaging, focusing specifically on the segmentation of zygomatic bones in cone-beam computed tomography (CBCT) images. The utilization of attention maps generated by gradient-weighted class activation mapping (Grad-CAM) and guided Grad-CAM algorithms for improved interpretability represents a novel contribution to the field. Comparisons with human dentists highlight the efficiency and accuracy gains achieved by the proposed model, setting it apart as a promising tool for 3D modeling in preoperative planning scenarios. While achieving a 99.64 accuracy rate, a dice coefficient score of 92.34 was obtained.

In their study, Ahmed and Hawezi [18] used the integration of machine learning algorithms in medical imaging, particularly for bone fracture detection, representing a significant advancement in enhancing diagnostic accuracy. The proposed system, encompassing pre-processing, edge detection, feature extraction, and machine learning classifications, underscores the multidimensional approach employed to refine the diagnostic process. The findings, with support vector machine (SVM) exhibiting an accuracy rate of 0.92 among the algorithms.

Du et al. [19] contributes to the field of skeletal bone age assessment (BAA) by proposing an innovative two-stage segmentation method for hand bone X-ray images. The importance of accurate segmentation in BAA is emphasized, given the intricate structure and small features of hand bones. The utilization of the OSA-YOLOv5 network for initial extraction and the subsequent application of GRU-UNet for separation mark a novel approach to enhance accuracy and completeness in segmentation. The GRU-UNet segmentation model demonstrates a significant improvement, achieving a 14.70% higher accuracy than the conventional Unet [20].

The study conducted by Karanam et al. [2] contributes to the burgeoning field of fracture detection by presenting a comprehensive overview of various techniques and methodologies. Notably, the study addresses a critical gap in existing literature by emphasizing the importance of not only detecting but also classifying bone fractures. This work serves as a valuable resource for researchers aiming to develop models that can automatically detect and classify fractures, supporting the construction of fracture detection.

In their study, conducted by Caron et al. [3] contributes to the advancing understanding of osteoporosis by presenting a novel approach to studying microdamage development in trabecular bone under mechanical loading. The study is centered on the usage of YOLOv4 [21] and Unet. The proposed two-step approach showcases the potential of YOLOv4 for microdamage detection and Unet for segmentation, offering promising results in accurately identifying and delineating microdamage regions. With average intersection over unions (IoUs) of 45.32% and 51.12% and mean average precisions (mAPs) of 28.79% and 46.22% for samples 1 and 2, respectively, the YOLOv4p5 model performed the best.

In their study, Zheng et al. [22] contributes to medical imaging by presenting a novel two-stage method designed for the automatic identification and localization of complex pelvic fractures. The proposed method stands out by harnessing the symmetry properties of pelvic anatomy and capturing symmetric feature differences caused by fractures on both sides, addressing limitations observed in existing methods that focus solely on image or geometric features. Leveraging supervised contrastive learning with a siamese deep neural network, incorporating a structural attention mechanism and a structure-focused attention (SFA) module, the method demonstrates superior mean accuracy and sensitivity as opposed to cutting-edge contrastive learning methods and advanced classification networks.

Our study distinguishes itself in the literature by meticulously focusing on fracture detection in medical imaging, with a distinct emphasis on CNN models and transformer models. Unlike earlier works, we adopted state-of-the-art detection models for our investigation, carefully selected for their potential advantages in our targeted domain. Our comparative analysis involves a thorough comparison between CNN and transformer models for fracture detection, along with a comprehensive evaluation. Ultimately, our research contributes to the burgeoning field of medical imaging, addressing critical gaps and paving the way for advancements in fracture detection that hold significant implications for clinical applications.

MATERIAL AND METHOD

In this section, dataset and object detection models employed for fracture detection are introduced.

Dataset

In this study, the dataset used for bone fracture detection is open source and obtained from the Roboflow platform [23]. The information regarding the utilized dataset is provided in Table 1.

Table 1. Dataset for bone fracture

	Train	Val	Test	Total
Dataset	5145	750	750	6645

There are 6645 images in all in the entire dataset, which encompass fractures from different types of bones. The dataset contains images of hand, finger, arm, leg, toe, and hip fractures. Figure 1 shows some of the images taken from the dataset.

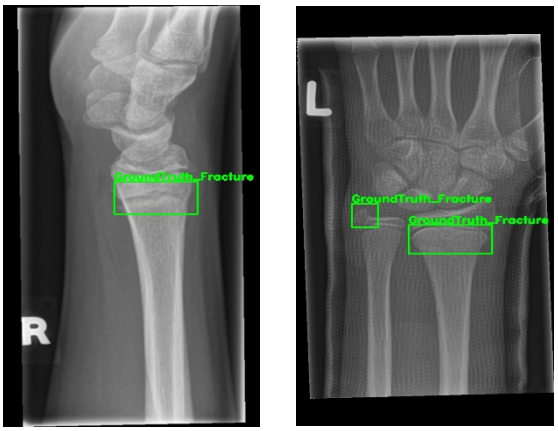

Figure 1. Sample images for fractures. The green-colored bounding boxes represent the ground truth.

Faster R-CNN

In order to extract features from the images, Faster R-CNN uses models like visual geometry group (VGG) [24] and residual network (ResNet) [25] as a backbone in the initial stage of the process [26]. Simultaneously, the VGG architecture is used as the backbone in the article on Faster R-CNN.

In the other stage of Faster R-CNN, the region proposal network (RPN) creates recommended regions for object detection, processing the image only once [27]. The purpose of the RPN is to send the extracted features to a 1x1 convolutional layer using 3x3 convolution and 2x2

max pooling layers. Two cells make up the RPN's output layer: one predicts the region bounding box, and the other determines if an item is there or not [27]. Subsequently, ROI pooling outputs the maximum feature map of each region and brings the regions proposed by RPN in different sizes to the same size [28]. After passing through convolution and fully connected layers, target class identification, and sensitive box presenter regression are performed.

Once region proposals are generated, the Fast R-CNN detector refines these proposals and classifies objects [29]. The classification loss penalizes incorrect class predictions. Cross-entropy loss is commonly used in classification problems. Equation 1 presents the cross-entropy loss.

$$H(y,\hat{y}) = -\big(y.log(\hat{y}+1-y) \times log(1-\hat{y})\big) \qquad (1)$$

where, y is the true label (ground truth), which is either 0 or 1. $\hat{y}$ is the predicted probability of the positive class.

Bounding box regression loss penalizes inaccurate bounding box regression. The Smooth L1 Loss function measures the absolute difference by comparing the ground truth with the bounding box drawn by the model [30]. The Smooth L1 loss is given in Equation 2.

$$SmoothL1(x) = \begin{cases} 0.5x^2 & if|x| < 0 \\ |x| - 0.5 & otherwise \end{cases} \qquad (2)$$

where, x is the discrepancy between the predicted bounding box coordinates and the ground truth bounding box coordinates.

Figure 2 shows the architecture of the Faster R-CNN.

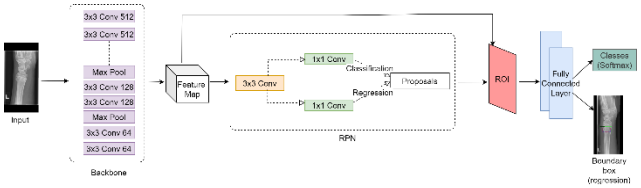

Figure 2. Architecture of Faster R-CNN.

As illustrated in Figure 2, during the initial stage of Faster R-CNN, the input image undergoes processing through a CNN backbone, such as ResNet or VGG, extracting a hierarchical set of feature maps. Concurrently, the RPN operates on these feature maps to generate region proposals.

RetinaNet

RetinaNet is a model proposed to address the issue of excessive foreground and background imbalance encountered during the training of other models in object detection [31]. RetinaNet consists of two main components: Feature Pyramid Network (FPN) [32] and focal loss.

CNN is used by FPN to capture richer information at a reduced resolution. The top-down path employs high-resolution feature maps, focusing on the general features of the image [33]. Pyramidal links in the specified paths of the FPN enable bottom-up or top-down connections, bringing together different feature levels and scales. Simultaneously, high-resolution feature maps and low-resolution feature maps merge. This merging enhances the visibility of small objects with larger feature maps, and large objects become more visible in small feature maps.

In general, loss functions assign the same importance to each sample because most samples belong to the same class. However, in some rare cases, this can cause the model to lose focus. Focal Loss, used in RetinaNet, is designed to prevent this by paying less attention to easily classifiable samples and more attention to more complex samples. Focal loss is given in Equation 3.

$$FocalLoss(p_t) = -(1-p_t)^{\gamma} log(p_t) \qquad (3)$$

where, p_t is the predicted probability of the true class, and γ is a tunable focusing parameter (typically a positive value). The term $(1-p_t)^{\gamma}$ is the focal weight, which down-weights the contribution of well-classified examples. Figure 3 shows the architecture of the RetinaNet.

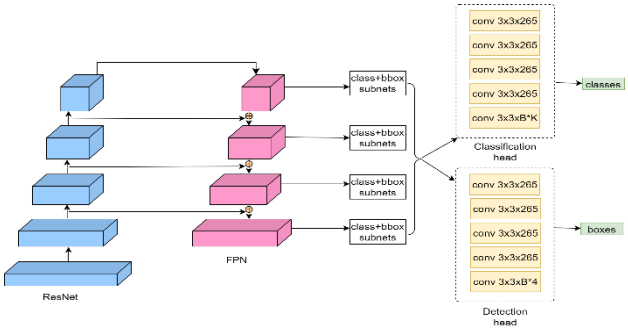

Figure 3. Architecture of RetinaNet

To attain the highest level of performance, RetinaNet combines a focal loss with an FPN, as can be observed in Figure 3.

Detection Transformer (DETR)

DETR is inspired by the transformer-based architecture. DETR can process the image in a single stage and estimate the object positions and class of objects. The input layer, transformer layer, and output layer are the three layers that make up DETR [34].

Firstly, in the input layer, the image passes through CNN layers using a backbone. Extracted features are used to identify the objects in the next stages. In the transformer layer of DETR, there are encoder and decoder. The encoder matches the features extracted from the backbone and the reference boxes [35]. Using the attention mechanism, the relationship between different objects in an image is learned. In the decoder part, classification and predictions are made for object detection. In the output layer, a feed-forward network is

used so that each layer produces an output for object detection and classification using the output from the previous layer.
The loss functions used in DETR are Hungarian Loss and Smooth L1 Loss [13]. Hungarian Loss calculates the absolute difference between the class labels predicted by the model and the actual class labels. The Hungarian loss is given in Equation 4.

$$HungarianLoss = -\sum_{i=1}^{N} Cost\left[i, assignment[i]\right] \quad (4)$$

where, N is the total number of predicted boxes, $assignment[i]$ is the index of the ground truth box assigned to the predicted box i according to the optimal assignment, and $Cost$ is the cost matrix representing the cost of assigning each predicted box to each ground truth box. Figure 4 displays the DETR's architecture.

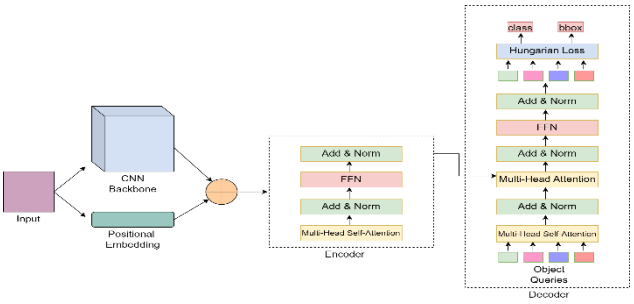

Figure 4. Architecture of DETR

DETR is an object detection architecture that uses transformer-based models to achieve accurate and effective detection tasks, as can be seen in Figure 4.

RESULTS AND DISCUSSION

Metrics are implemented to assess and enhance the model's performance by determining whether the model properly or mistakenly predicts item detections. Simultaneously, they help to take necessary actions according to the model's state by evaluating its performance. The metrics used to test object detection models are as follows:
Precision: It measures the accuracy of positive predictions made by the model. The precision is given in Equation 5.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall: It measures the ability of the model to capture all positive instances. Recall is given by Equation 6.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F1 Score: It considers both precision and recall to provide a balanced assessment of a model's performance. The F1 Score is given by Equation 7.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

mAP(Mean Average Precision): It combines precision and recall to provide a comprehensive measure of how well a model identifies and localizes objects in an image. mAP is given in Equation 8.

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \quad (8)$$

where, C is the whole count of classes. AP_c is the average precision for a class c.
Faster R-CNN, RetinaNet, and DETR models were tested using 750 test images. Both training and testing processes were conducted in the Google Colaboratory environment. Table 2 displays the results that were achieved.

Table 2. Results of the model's performances

Model/Backbone	mAP50	Precision	Recall	F1-Score
FasterR-CNN/ResNet50	0.828	0.899	0.906	0.902
FasterR-CNN/ResNet101	0.832	0.906	0.907	0.907
RetinaNet/ResNet50	0.896	0.921	0.878	0.899
RetinaNet/ResNet101	0.901	0.917	0.880	0.898
DETR/ResNet50	0.844	0.870	0.883	0.887
DETR/ResNet101	0.840	0.905	0.887	0.896

The backbone of the Faster R-CNN and RetinaNet models contains FPN. Although the same type of backbone is used in the models, the depths of the backbones are different. Faster R-CNN and RetinaNet models were trained for 20,000 iterations, while DETR models were trained for 200 epochs. A batch size of 8 and a learning rate of 0.0001 were applied to each model during training. Additionally, adam was used to optimize the models.
In Table 2, the highest mAP50 value is observed in the RetinaNet-ResNet101 model with a ratio of 0.901, indicating high success in object detection. The RetinaNet/ResNet50 model also demonstrates good performance with a high mAP50 value of 0.896. However, Faster R-CNN models have lower but still acceptable mAP50 values. The Faster R-CNN/ResNet50 model and Faster R-CNN/ResNet101 model achieved ratios of 0.828 and 0.832, respectively.
The highest Precision value is 0.921, belonging to the RetinaNet/ResNet50 model, indicating how accurately the model identifies true positive objects. Other models also have high Precision values. While the RetinaNet/ResNet101 model has a precision ratio of 0.917, slightly behind the RetinaNet/ResNet50 model, it outperforms other models.
The highest Recall value is 0.907, observed in the Faster R-CNN/ResNet101 model, indicating its ability to detect truly positive objects with a high success rate. The Faster R-CNN/ResNet50 model showed a recall ratio of 0.906,

very close to the performance of the Faster R-CNN/ResNet101 model.

The highest F1-Score value is 0.907, belonging to the Faster R-CNN/ResNet101 model, demonstrating a good balance between Precision and Recall metrics. The Faster R-CNN/ResNet50 model also has a high F1-Score value of 0.902, indicating a well-maintained balance.

The mAP50 values for DETR/ResNet50 and DETR/ResNet101 models are 0.844 and 0.840, respectively. These values indicate that DETR models have an average performance in object detection. The Precision value for the DETR/ResNet50 model is 0.870, while for the DETR/ResNet101 model, it is 0.905. These values indicate the proportion of detected objects by DETR models that are truly positive. The Recall values for DETR/ResNet50 and DETR/ResNet101 models are 0.883 and 0.887, respectively. The DETR/ResNet101 model has a slightly higher Recall value compared to DETR/ResNet50.

RetinaNet generally outperforms Faster R-CNN and DETR in terms of mAP50, precision, indicating its superiority in object detection accuracy and reduction of false positives. Faster R-CNN achieves the highest recall, suggesting its effectiveness in capturing a higher proportion of true positive instances. DETR shows competitive performance but falls slightly behind in terms of mAP50, precision, and F1-Score compared to RetinaNet and Faster R-CNN.

Figure 5 displays the RetinaNet/ResNet101 model's prediction results.

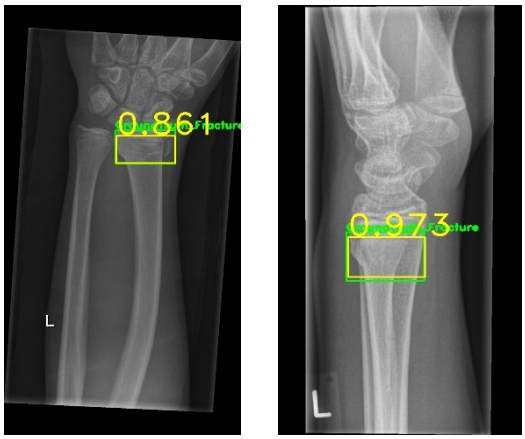

Figure 5. RetinaNet/ResNet101 prediction results.

Figure 5 displays sample images of prediction results obtained using the RetinaNet/ResNet101 model. The green color represents the ground truth, while the yellow color represents the predicted bounding box. Additionally, confidence score values are also shown.

The model exhibiting the best performance in terms of mAP50 ratio, RetinaNet/ResNet101, is compared with existing studies in Table 3. Existing studies have used similar datasets aimed at bone fracture detection.

Table 3. Evaluation of the suggested model in the light of previous research.

Paper	Model/Backbone	mAP50	Precision	Recall	F1-Score
Ma and Luo [36]	Faster R-CNN+CrackNet	-	0.897	0.9049	0.9014
Guan et al. [37]	CNN	0.620	-	-	-
Guan et al. [38]	DCFPN	0.821	-	-	-
Zou and Arshad [39]	YOLOv7-ATT	0.862	-	-	-
Caron et al. [3]	YOLOv4p5	0.462	-	-	-
Proposed Model	RetinaNet/ResNet101	0.901	0.917	0.880	0.898

The analysis of the presented table reveals notable insights into the performance of various object detection models. The Faster R-CNN+CrackNet model, as reported in Ma and Luo [36], achieved a precision of 0.897, recall of 0.9049, and an F1-Score of 0.9014, but no mAP50 value was provided. When compared to the recommended model, Faster R-CNN+CrackNet surpasses our model in terms of recall and F1-score, while our proposed model excels in precision. The CNN model proposed by Guan et al. [37] achieved a modest mAP50 of 0.6204, highlighting its limitations. Similarly, the DCFPN model proposed by Guan et al. [38] showed a mAP50 of 0.821, while YOLOv7-ATT proposed by Zou and Arshad [39] demonstrated a mAP50 of 0.862. The proposed method performs competitively with these

models in terms of mAP50. YOLOv4p5 proposed by Caron et al. [3], with a mAP50 of 0.4622, lags significantly behind the proposed model.

In terms of precision, the proposed model excels with a value of 0.917, demonstrating its ability to minimize false positives. The CNN model proposed by Guan et al. [37], lacks precision information, necessitating further evaluation. For recall, the Faster R-CNN+CrackNet model proposed by Ma and Luo [36] achieves the highest score at 0.9049, emphasizing its effectiveness in capturing true positives. In contrast, the proposed model exhibits a slightly lower recall of 0.880, indicative of its potential trade-off with precision. Lastly, considering the F1-Score, the proposed RetinaNet/ResNet101 remains superior with a value of 0.898, offering a balanced performance between precision and recall. Meanwhile, YOLOv4p5 again lags behind without specific F1-Score information.

Overall, the RetinaNet/ResNet101 model outperforms other models according to the mAP50 metric, while surpassing the Faster R-CNN+CrackNet model in terms of precision. The Faster R-CNN+CrackNet model outperformed our model only in the Recall and F1-Score metrics.

CONCLUSION

Faster R-CNN, RetinaNet, DETR models were compared for bone fracture detection based on mAP, precision, recall and F1-score values. The RetinaNet/ResNet101 model has the highest mAP and precision, which may be due to the fact that they are more complex samples thanks to Focal loss. The DETR models, which comes from a vision transformer-based architecture instead of a CNN-based architecture like Faster R-CNN and RetinaNet, is very close to the Faster R-CNN models in terms of F1 score. Considering all these results, it is decided that CNN-based models such as Faster R-CNN and RetinaNet are the most suitable for optimal bone fracture detection. However, it is important to recognize certain limitations in our study. Firstly, the model's dependence on extensive training data might pose challenges in scenarios where acquiring a diverse and sufficiently large dataset is difficult. The generalizability of the model to various imaging conditions and patient demographics may be compromised as a result. Secondly, the lack of interpretability in deep learning models remains a concern, as understanding the decision-making process is crucial, especially in medical applications.

In future studies, we aim to explore the potential benefits of ensemble models, amalgamating predictions from various architectures to create a synergistic and more robust framework for fracture detection. Additionally, our future endeavors will include extensive real-world validation studies across diverse clinical settings, ensuring the practical effectiveness and reliability of the bone fracture detection models in authentic healthcare environments.

ACKNOWLEDGMENT

This study was supported by AKGUN Computer Incorporated Company. We would like to thank AKGUN Computer Inc. for providing all kinds of opportunities and funds for the execution of this project.

REFERENCES

[1] Czermak E.D., Euler A., Franckenberg S., Finkenstaedt T., Villefort C., Dominic G., Guggenberger R. Evaluation of ultrashort echo-time (UTE) and fast-field-echo (FRACTURE) sequences for skull bone visualization and fracture detection – A postmortem study, Journal of Neuroradiology. 49 237-243, 2022

[2] Karanam S.R., Srinivas Y., Chakravarty S. A systematic review on approach and analysis of bone fracture classification, Materials Today: Proceedings. 80 2557-2562, 2023

[3] Caron R., Londono I., Seoud L., Villemure I. Segmentation of trabecular bone microdamage in Xray microCT images using a two-step deep learning method, Journal of the Mechanical Behavior of Biomedical Materials. 137 105540, 2023.

[4] Ozdemir C., Dogan Y. Advancing brain tumor classification through MTAP model: an innovative approach in medical diagnostics, Medical and Biological Engineering and Computing. 1-12, 2024

[5] Ozdemir C. Classification of brain tumors from MR images using a new CNN architecture." Traitement du Signal. 40(2) 611-618, 2023.

[6] Guan B., Yao J., Wang S., Zhang G., Zhang Y., Wang X., Wang M. Automatic detection and localization of thighbone fractures in X-ray based on improved deep learning method, Computer Vision and Image Understanding. 216 103345, 2022.

[7] O'Shea K., Nash R. An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458, 2015.

[8] Ozdemir C., Dogan Y., Kaya Y. RGB-Angle-Wheel: A new data augmentation method for deep learning models. Knowledge-Based Systems. 291 111615, 2024

[9] Ren S., He K., Girshick R., Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence. 39 1137-1149, 2017.

[10] Lin T.Y., Goyal P., Girshick R., He K., Dollár P. Focal loss for dense object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence. 42 318-327, 2020.

[11] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale, International Conference on Learning Representations. 2021.

[12] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin, I. Attention is all you need, Advances in neural information processing systems 30(NIPS 2017). 30, 2017.

[13] Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. End-to-end object detection with transformers, European Conference on Computer Vision. 12346 213-229, 2020.

[14] Warin K., Limprasert W., Suebnukarn S., Inglam S., Jantana P., Vicharueang S. Assessment of deep convolutional neural network models for mandibular fracture detection in panoramic radiographs, International Journal of Oral and Maxillofacial Surgery. 51 1488-1494, 2022.

[15] Huang G., Liu Z., Maaten L.V.D., Weinberger K.Q. Densely Connected Convolutional Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2261-2269, 2017.

[16] Kim D.Y., Park E., Ku K., Hwang S.J., Hwang K.T., Lee C.H., Yoon G.H. Application of stacked autoencoder for identification of bone fracture, Journal of the Mechanical Behavior of Biomedical Materials. 146 106077, 2023.

[17] Tao B., Yu X., Wang W., Wang H., Chen X., Wang F., Wu Y. A deep learning-based automatic segmentation of zygomatic bones from cone-beam computed tomography images, Journal of Dentistry. 135 104582, 2023.

[18] Ahmed K.D., Hawezi R. Detection of bone fracture based on machine learning techniques, Measurement: Sensors. 27 100723, 2023.

[19] Du H., Wang H., Yang C., Kabalata L., Li H., Qiang C. Hand bone extraction and segmentation based on a convolutional neural network, Biomedical Signal Processing and Control. 89 105788, 2024.

[20] Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. 9351 234-241, 2015.

[21] Bochkovskiy A., Wang C.Y., Liao H.Y.M. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[22] Zheng B., Wang H., Xu J., Tu P., Joskowicz L., Chen X. Two-Stage Structure-Focused Contrastive Learning for Automatic Identification and Localization of Complex Pelvic Fractures, IEEE Transactions on Medical Imaging. 42 2751-2762, 2023.

[23] Roboflow 100. Bone fracture dataset, Roboflow Universe. 2023.

[24] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[25] He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition. 770-778.

[26] Han S., Xiao X., Song B., Guan T., Zhang Y., Lyu M. Automatic borehole fracture detection and characterization with tailored Faster R-CNN and simplified Hough transform, Engineering Applications of Artificial Intelligence. 126 107024, 2023.

[27] Lyu H., Qiu F., An L., Stow D., Lewison R., Bohnett E. Deer survey from drone thermal imagery using enhanced faster R-CNN based on ResNets and FPN, Ecological Informatics. 79 102383, 2024.

[28] Tang Y., Chen Y., Sharifuzzaman S.A.S.M., Li T. An automatic fine-grained violence detection system for animation based on modified faster R-CNN, Expert Systems with Applications. 237 121691, 2024.

[29] Girshick R. Fast r-cnn, Proceedings of the IEEE international conference on computer vision. 1440-1448, 2015.

[30] Yang W., Xiao Y., Shen H., Wang Z. Generalized weld bead region of interest localization and improved faster R-CNN for weld defect recognition, Measurement. 222 113619, 2023.

[31] Cheng J., Wang R., Lin A., Jiang D., Wang Y. A feature enhanced RetinaNet-based for instance-level ship recognition, Engineering Applications of Artificial Intelligence. 126 107133, 2023.

[32] Lin T.Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature pyramid networks for object detection, Proceedings of the IEEE conference on computer vision and pattern recognition. 2117-2125, 2017.

[33] Tong L., Xue S., Chen X., Fang R. Artificial intelligence-based detection of posterior tibial slope on X-ray images of unicompartmental knee arthroplasty patients, Journal of Radiation Research and Applied Sciences. 16 100615, 2023.

[34] Chen Y., Zhang C., Chen B., Huang Y., Sun Y., Wang C., Fu X., Dai Y., Qin F., Peng Y., Gao Y. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases, Computers in Biology and Medicine. 170 107917, 2024.

[35] Zheng H., Wang G., Xiao D., Liu H., Hu X. FTA-DETR: An efficient and precise fire detection framework based on an end-to-end architecture applicable to embedded platforms, Expert Systems with Applications. 248 123394, 2024.

[36] Ma Y., Luo Y. Bone fracture detection through the two-stage system of Crack-Sensitive Convolutional Neural Network, Informatics in Medicine Unlocked. 22 100452, 2021.

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

[38] Guan B., Yao J., Zhang G., Wang X. Thigh fracture detection using deep learning method based on new dilated convolutional feature pyramid network, Pattern Recognition Letters. 125 521-526, 2019.

[39] Zou J., Arshad M.R. Detection of whole body bone fractures based on improved YOLOv7, Biomedical Signal Processing and Control. 91 105995, 2024.