Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 13 NO: 1 PAGE: 358-365 YEAR: 2024
DOI:10.17798/bitlisfen.1413650

Upper and lower extremity bone segmentation with Mask R-CNN

Ayhan AYDIN[1*], Caner ÖZCAN[2]

[1]*Computer Engineering, Ondokuz Mayıs University, 55080, Samsun, Turkiye*
[2]*Software Engineering, Karabuk University, 78050, Karabuk, Turkiye.*
(ORCID: 0000-0001-9127-0951) (ORCID: 0000-0002-2854-4005)

Keywords: Mask R-CNN, Segmentation, Bone, Lower extremity, Upper extremity.

Abstract
Most medical image processing studies use medical images to detect and measure the structure of organs and bones. The segmentation of image data is of great importance for the determination of the area to be studied and for the reduction of the size of the data to be studied. Working with image data creates an exponentially increasing workload depending on the size and number of images and requires high computing power using machine learning methods. Our study aims to achieve high success in bone segmentation, the first step in medical object detection studies. In many situations and cases, such as fractures and age estimation, the humerus and radius of the upper extremity and the femur and tibia of the lower extremity of the human skeleton provide data. In our bone segmentation study on X-RAY images, 160 images from one hundred patients were collected using data compiled from accessible databases. A segmentation result with an average accuracy of 0.981 was obtained using the Mask R-CNN method with the resnet50 architecture.

1. Introduction

Many doctors use medical images to decide whether lesions, fractures, etc., occur in complaints that are thought to be caused by bone. With the development of medical imaging devices, a large number of high-quality medical imaging methods such as X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) are widely used [1]. Visual inspection of X-ray data to determine appropriate treatments is a primary means of detecting and determining the severity of bone and bone-related phenomena [2]. An experienced physician may need to spend a lot of time checking bone integrity and health status on the X-ray image. Many hospitals today have a shortage of experienced radiologists to handle these medical images. To assist doctors in detecting bone-related disorders, computer-aided diagnosis (CAD) has been widely used to analyze medical images and has received increasing attention [3]. Deep learning applications are intensively used to classify health data and can potentially provide pioneering knowledge to domain experts [4,5]. In our study, a segmentation application that provides the boundary information of the object with formal precision will be applied beyond visual classification and object identification within frames. Although it is translated into our language as segmentation, the process referred to as segmentation is to divide a digital image into regions or objects in the image [6]. Segmentation of four long bones in the lower and upper extremities will support experts' decision-making processes, from bone integrity to age estimation and detection of different diseases. Our study aims to show the bone structure on the image using direct radiograph images. Direct radiographs are the most intensively used diagnostic method among medical imaging methods, which is the most cost-effective diagnostic method with the highest data access. It is known that radiographs, which are examined and reported by radiologists in the first step and then by specialized physicians, create a workload for more than one physician. Today, radiologists use medical imaging methods to examine and report images in different areas and frequencies. There are systems in which

[*]Corresponding author: *ayhan.aydin@omu.edu.tr*

Received: 02.01.2024, Accepted: 15.03.2024

medical images are collected in a pool for interpretation and waiting for the appointment of a specialist. Automating the preliminary information in the radiographs interpreted by more than one specialist physician will save much time and labor by marking the preliminary information about different structures and presenting it to the specialist on the image. In image processing, determining the point of interest is an important stage and is one of the first steps in almost all studies. The term region of interest in the literature consists of visualizing bone and cartilage structure on direct radiographs. In addition to being a pioneering procedure, our study has an innovative aspect, with the masking success reaching 0.99 and above and possibly being integrated into radiological imaging devices. The segmentation of this region will be provided by masking produced with Mask R-CNN, as explained in the following sections.

Recently, thanks to the higher computational power of graphics processing units (GPUs), many new works on CAD based on deep learning have been presented. Current deep learning-based segmentation methods utilize some fully convolutional network (FCN) derivatives to estimate the class labels of all pixels in an image in parallel. Due to the pooling layers and up-sampling process, spatial information may be lost during the prediction, so there are some inaccuracies in the predicted segmentation map, especially in sharp regions such as boundaries. Therefore, a hopping architecture in FCN has been proposed to solve this problem [7]. Following this idea, Ronneberger et al. (2015) proposed a standard and popular medical image segmentation architecture called U-Net, which includes a symmetric encoder and decoder. The features of each encoder layer are hopingly connected to the corresponding decoder layer to recover the lost spatial information [8]. In bone structure segmentation, studies differ according to the preprocessing or masking area, and different results have been obtained with different network architectures.

Bullock et al. (2019) applied XNet architecture to segment X-ray images into bones and skin and obtained results with an overall accuracy of 0.92, which surpasses classical methods [9]. Drozdzal et al. (2018) combined Cnet with the fully connected resnet architecture (FC-ResNet) applied for CT liver images to detect lesions and obtain high-accuracy results [10]. Omar (2019) used a VGG-16-based SegNet model for CT lung image segmentation and achieved an average success rate of 0.95 [11]. Deep learning is a process that allows computational models consisting of multiple processing layers to learn representations of data with multiple levels of abstraction for the automatic segmentation of different anatomical structures. It includes automatic segmentation methods that are classified as either pre-supervised or unsupervised. For supervised methods, segmentation requires operator interaction throughout the process, while unsupervised methods usually require operator intervention only after the end of the process. Unsupervised methods are preferred to obtain a reproducible result [12]. In another study, U-net detected different human bones from computed tomography images, achieving a segmentation accuracy of 0.93 [13]. Smistad et al. (2015) applied deep learning methods to MRI tomography images to detect lung tumors and, at the same time, improve MRI image quality. The segmentation processes obtained with current methods exceed the success threshold of 0.90 based on the literature review [14].

For segmentation tasks, Mask R-CNN has several advantages over Faster R-CNN. First, it can generate more accurate and fine-grained masks for each object, better at capturing the details of shape and contour than bounding boxes. Second, it can handle overlapping and occluded objects better than semantic segmentation models, which can confuse pixels from different instances of the same class. Third, it can use the existing architecture and pre-trained weights of Faster R-CNN, which can reduce training time and data requirements. Finally, adding or modifying the mask branch can extend it to other tasks, such as key point detection, pose estimation, and panoptic segmentation.

2. Material and Method

2.1. Convolutional neural networks

It is similar to traditional artificial neural networks, consisting of neurons that self-optimize through learning. Each neuron takes one input and performs a process that is the basis of numerous ANNs. From the input raw image vectors to the final output of the class score, the entire network continues to express a single weight. The final layer contains the loss functions associated with the classes and can be used to optimize objectives such as class scores. [15]. Convolutional neural networks, also known as CNNs, are a particular type of neural network, usually consisting of the following layers.

Paragraphs following the first paragraph should begin with the paragraph indentation. The general structure of CNN layers is presented in Figure 1.

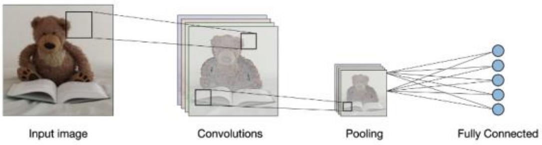

Figure 1. CNN layers [16].

When scanning images based on their size, filters perform convolution layer operations. Hyperparameters such as the filter size and step are used. The outcome of this is an activation map or feature map. This process can also be regarded as image reduction or feature extraction.

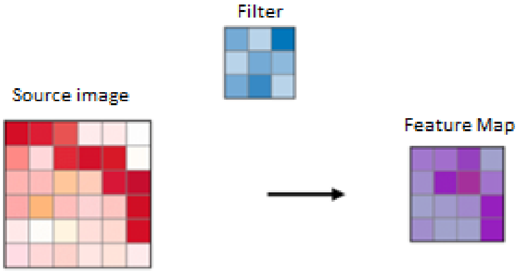

Figure 2. Convolution layer.

The full link layer represents the sampling process following the convolution layer, typically representing spatial variation. Specifically, maximum, and average co-registration represent distinct categories of co-registration, with maximum and average values taken correspondingly. Opting for maximum averaging retains the perceived features by identifying the current matrix's highest value, constituting the most preferred method. Different methods can be employed in this layer to guarantee the selection of varied features. Convolution layer sections are presented in Figure 2.

The fully linked layer works on an input where all neurons are connected to each input. As seen in Figure 3, these layers are commonly situated towards the end of a CNN architecture and can be utilized to enhance objectives such as class scores.

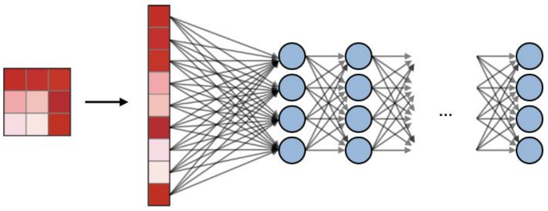

Figure 3. Fully connected layer.

2.2. Region-based CNN

The architecture used to identify objects in images and the classes of these objects was published by Girshick et al. in 2014 [17]. RCNN is run on images containing multiple objects in two different steps. The first of these steps is selective search. In this stage, the features that are candidates to be objects in the image are determined. The RCNN architecture is shown in Figure 4 below. After identifying approximately 2000 regions, each region is entered into the CNN model separately, and the boxes defining the boundaries are predicted.

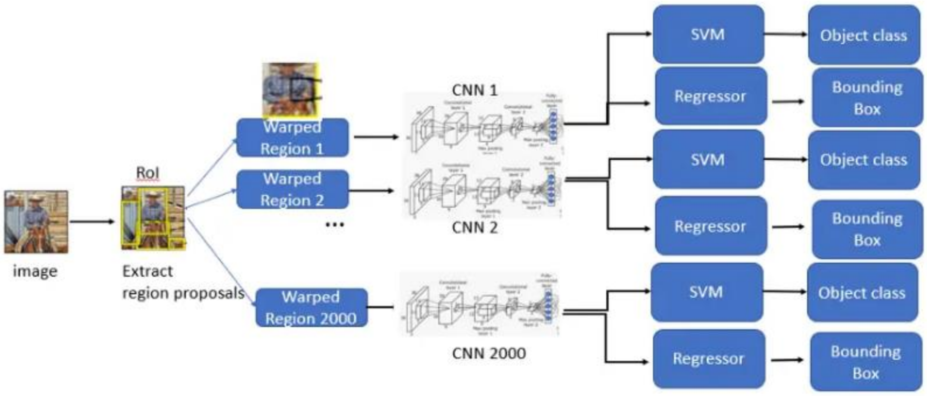

Figure 4. Region-based CNN architecture [17].

2.3. Mask R-CNN

It is a state-of-the-art example segmentation technique proposed by He et al. [18]. As shown in

Figure 5, Mask-RCNN is divided into two branches of the network: classification prediction and mask prediction. The classification prediction branch is the same as Faster-RCNN, predicts the relevant domain and produces class labels and rectangular box coordinate output. Each binary mask produced by the mask prediction branch relies on the classification prediction results to separate these objects. Mask-RCNN independently predicts a binary mask for each class to avoid competition.

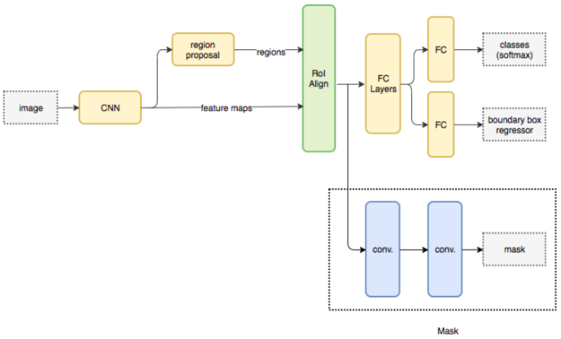

Figure 5. Mask-RCNN architecture [18].

2.3.1 Mask R-CNN Hyperparameters

Back Bone: The backbone is the Conv Net architecture. This is used in the first step of Mask R-CNN. Available backbone selection options include ResNet50, ResNet101 and ResNext 101.

Train_ROIs_Per_Image: This is the maximum number of ROIs the Region Proposal Network will generate for the image. These ROIs will be processed in the next step for classification and masking.

Detection_Min_Confidence: This is the confidence threshold above which an instance will be classified. It can be initialized by default. It can be increased or decreased depending on the model's detected instances.

Image_Min_Dim and Image_Max_Dim: These settings control the image size. The default settings resize images to 1024x1024 squares. Smaller images (512x512) can reduce memory requirements and training time.

Loss weights: Mask RCNN uses a complex loss function, calculated as the weighted sum of different losses in each model state. The hyper-parameters of the weight of the losses correspond to the weight the model should give to each state.

2.3.2 Mask R-CNN Evaluation Metric

Pixel accuracy is a common evaluation metric used in image segmentation to measure the overall accuracy of the segmentation algorithm. It is the ratio of correctly classified pixels to the total number of pixels in the image.

$$Pixel\ accuracy = \frac{Number\ of\ selected\ pixels}{Total\ number\ of\ area\ pixels} \qquad (1)$$

2.4. Resnet

In 2015, Resnet, a network structure proposed by He et al., ranked first in ILSVRC-2015 classification, ImageNet detection and localization, COCO detection, and segmentation tasks. The deepening network structure aims to solve the problem of reducing the training error by using the residual block structure. Residual blocks are added to the output by skipping one or more layers. The identity block is used if the input and output are the same [19]. If the residual block does not provide learning, it does not impose an additional load on the structure, but generally, the residual block contributes positively to the network's learning. Resnet50 architecture is presented in Figure 6 below.

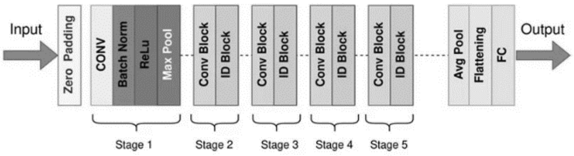

Figure 6. Resnet50 architecture [19]

With the structure shown in Figure 7, Resnet50 is used as 1x1 convolution, 3x3 convolution, and 1x1 dimensionality to recover the actual size.

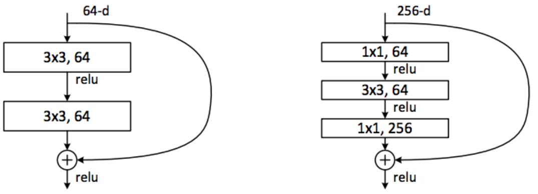

Figure 7. Resnet50 link example [19]

3. Results and discussion

LERA X-RAY images made available by Stanford University were used for educational and academic studies [20]. To compare the aims and data amounts of the studies in the literature, the information on the publications in similar fields is presented in Table 1 below.

Table 1. Similar studies in the literature and the amount of visual data used

Study	Aim	Data
Yahalomi vd. (2019) [2]	Fracture detection	38
He vd. (2020) [21]	Tumor detection	291
Eweje vd. (2021) [22]	Tumor classification	1060
Chianca vd. (2021) [23]	Tumor classification	146
Anizusman vd. (2021) [24]	Tumor classification	50
Karthik vd. (2021) [25]	Pneumonia detection	5000
Thakur ve Kumar (2021) [26]	Pneumonia detection	3877
Felfeliyan (2022) [27]	Cartilage Segmentation	500

Images with complete bone integrity and without prosthesis were selected from the set containing different anomaly and prosthesis images. In addition, a student profile was created on the Medpix page, and appropriate images of the relevant bones made available for sharing were used [28].

The dataset consisting of 160 different images was labeled as polygons with the single class 'bone' tag in Microsoft's 'Common Objects in Contects' COCO format, which is widely used in image segmentation. Examples of labeled images are shown in Figure 8. Depending on the imaging equipment, X-rays are available in different resolutions, such as 1024*817, 2436*2966, and 2021*2021. With the Mask R-CNN architecture, these images are resized to 1024*1024.

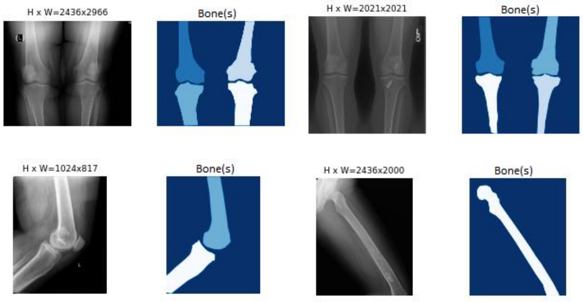

Figure 8. Labeled image samples.

The dataset was divided randomly into training and test sections at a ratio of 120/40. Data augmentation was not applied because it would produce images outside the X-ray acquisition standards. For example, if augmentation techniques were to be applied to the lower extremity image, the femur and tibia would be in different positions, which would differ from the real images. Train was performed on a Tesla K80 GPU provided by Google Colab. Experiments were performed with different resolutions, epochs, and parameters, and the optimal success value was obtained independently of the test images, with a resolution of 1024*1024 and a learning rate (lr) of 0.001. The process took 84 minutes, and the test results exceeded the success threshold set in the literature. Table 2 below presents some test results with Resnet101 and Resnet50 architectures using different parameters and average sensitivity values.

Table 2. Training parameters and accuracy results

mAP	Parameters			
	Validation steps	Steps per epoch	Backbone	Min. conf.
0,981	50	1000	Resnet101	0,7
0,989	50	1000	Resnet50	0,7
0,976	50	500	Resnet101	0,7
0,977	50	500	Resnet50	0,7
0,976	100	1000	Resnet101	0,7
0,977	100	1000	Resnet50	0,7

4. Conclusion and Suggestions

When diagnosing bone-related complaints, our application aims to reduce the target region by detecting the relevant area on X-ray images, which is the primary source of information for medical specialists. Our results show an average accuracy of 0.981. Figure 9 also lists the accuracy values obtained for the test radiographs, with the lowest value of 0.88. This shows that a segmentation success of 0.99 is possible by overcoming some limitations. Figure 10 shows sample images of the test set. It is important to note that this is an objective evaluation based on the resources used. The application aims to improve successful object/structure detection within the masked area. This is based on pioneering studies with Mask-RCNN in detecting anomalies, fractures, and lesions on the four long bones. The study has reached a point of success with masking and aims to create innovative software that can be added to radiological imaging devices. The Mask R-CNN architecture preferred for the study provides the segmentation function on visual data. This process provides region of interest (ROI) extraction in fracture, implant, tumor, and anomaly detections, which are intensively included in the literature and controlled depending on the bone structure. Running deep learning models on ROI-detected images for object detection and similar operations will provide a lower workload and faster results [29]. Future studies will include images with structural differences for data augmentation and sample object detection. Additionally, they will continue to explore applications with different network and architectural structures and carry out optimization trials using the existing parameters to achieve the most optimal results for the data.

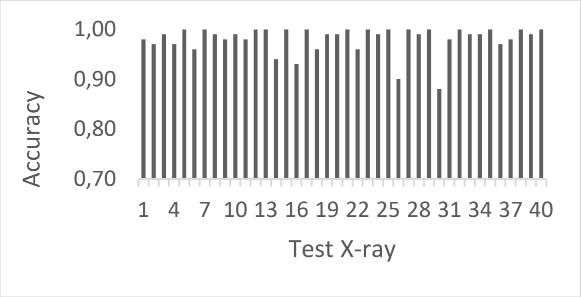

Figure 9. Test X-rays accuracy.

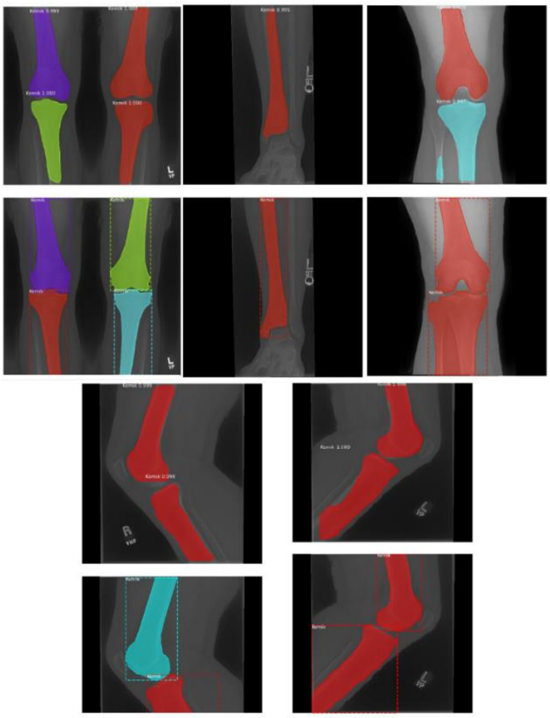

Figure 10. Prediction examples on test data.

Contributions of the authors

Ayhan Aydın carried out data preparation, deep learning modelled experiments and drafted the manuscript. Caner Özcan participated in the design of the study and helped to prepare the manuscript.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The authors of this article declare that the materials and methods used in this study do not require ethics committee approval and/or legal-special permission

References

[1] Y. Ma and Y. Luo, "Bone fracture detection through the two-stage system of Crack-Sensitive Convolutional Neural Network," *Inform. Med. Unlocked*, vol. 22, no. 100452, p. 100452, 2021.

[2] E. Yahalomi, M. Chernofsky, and M. Werman, "Detection of distal radius fractures trained by a small set of X-ray images and faster R-CNN," *in Advances in Intelligent Systems and Computing, Cham: Springer International Publishing*, 2019, pp. 971–981.

[3] T. Urakawa, Y. Tanaka, S. Goto, H. Matsuzawa, K. Watanabe, and N. Endo, "Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network," *Skeletal Radiol*., vol. 48, no. 2, pp. 239–244, 2019.

[4] H. Çetiner, "Cataract disease classification from fundus images with transfer learning based deep learning model on two ocular disease datasets," *Gümüshane Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 13, no. 2, 2023.

[5] K. A. Y. A. Volkan and İ. Akgül, "Classification of skin cancer using VGGNet model structures," *Gümüşhane Üniversitesi Fen Bilimleri Dergisi*, vol. 13, no. 1, pp. 190–198, 2023.

[6] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, Ruan Qiuqi. *Digital Image Processing*, vol. 8. Beijing: Publishing House of Electronics Industry, 2007.

[7] D. Wang et al., "A novel dual-network architecture for mixed-supervised medical image segmentation," *Comput. Med. Imaging Graph*., vol. 89, no. 101841, p. 101841, 2021.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv [cs.CV], 2015.

[9] J. Bullock, C. Cuesta-Lazaro, and A. Quera-Bofarull, "XNet: a convolutional neural network (CNN) implementation for medical x-ray image segmentation suitable for small datasets," *in Medical Imaging* 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, 2019.

[10] M. Drozdzal et al., "Learning normalized inputs for iterative estimation in medical image segmentation," *Medical image analysis*, vol. 44, pp. 1–13, 2018.

[11] A. Omar, "Lung CT Parenchyma Segmentation using VGG-16 based SegNet Model," *Int. J. Comput. Appl*., vol. 178, no. 44, pp. 10–13, 2019.

[12] H. Lee et al., "Fully automated deep learning system for bone age assessment," *J. Digit. Imaging*, vol. 30, no. 4, pp. 427–441, 2017.

[13] F. La Rosa, *A deep learning approach to bone segmentation in CT scans*, Universit` a di Bologna, Alma Mater Studiorum, 2017.

[14] E. Smistad, T. L. Falch, M. Bozorgi, A. C. Elster, and F. Lindseth, "Medical image segmentation on GPUs-A comprehensive review," *Medical image analysis*, vol. 20, no. 1, pp. 1–18, 2015.

[15] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," arXiv [cs.NE], 2015.

[16] A. A. Shervine Amidi, *Stanford Convolutional Neural Networks Handbook. Palo Alto, CA*: Stanford University.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *in 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[18] He, K., Gkioxari, G., Dollár, P., & Girshick, R, "Mask r-cnn," *in IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] Stanford University, "LERA- Lower Extremity RAdiographs," Stanford Center for Artifical Intelligence in Medicine & Imaging. [Online]. Available: https://aimi.stanford.edu/lera-lower-extremity-radiographs. [Accessed: 12-Oct-2023].

[21] Y. He et al., "Deep learning-based classification of primary bone tumors on radiographs: A preliminary study," *EBioMedicine*, vol. 62, no. 103121, p. 103121, 2020.

[22] F. R. Eweje et al., "Deep learning for classification of bone lesions on routine MRI," *EBioMedicine*, vol. 68, no. 103402, p. 103402, 2021.

[23] V. Chianca et al., "Radiomic machine learning classifiers in spine bone tumors: A multi-software, multi-scanner study," *Eur. J. Radiol*., vol. 137, no. 109586, p. 109586, 2021.

[24] D. M. Anisuzzaman, H. Barzekar, L. Tong, J. Luo, and Z. Yu, “A deep learning study on osteosarcoma detection from histological images,” *Biomed. Signal Process. Contro*l, vol. 69, no. 102931, p. 102931, 2021.

[25] R. Karthik, R. Menaka, and H. M, “Learning distinctive filters for COVID-19 detection from chest X-ray using shuffled residual CNN,” *Appl. Soft Comput*., vol. 99, no. 106744, p. 106744, 2021.

[26] S. Thakur and A. Kumar, “X-ray and CT-scan-based automated detection and classification of covid-19 using convolutional neural networks (CNN),” *Biomed. Signal Process. Control*, vol. 69, no. 102920, p. 102920, 2021.

[27] B. Felfeliyan, A. Hareendranathan, G. Kuntze, J. L. Jaremko, and J. L. Ronsky, “Improved-Mask R-CNN: Towards an accurate generic MSK MRI instance segmentation platform (data from the Osteoarthritis Initiative),” *Comput. Med. Imaging Graph*., vol. 97, no. 102056, p. 102056, 2022.

[28] An official website of the United States government, “https://medpix.nlm.nih.gov/,” MEDPIX. [Online]. Available: https://medpix.nlm.nih.gov/search?allen=true&-allt=true&alli=true&query=tibia,. [Accessed: 10-Dec-2023].

[29] A. Aslam and E. Curry, “A survey on object detection for the internet of multimedia things (IoMT) using deep learning and event-based middleware: Approaches, challenges, and future directions,” *Image Vis. Comput*., vol. 106, no. 104095, p. 104095, 2021.