**ORIGINAL ARTICLE**

# Designing an Intelligent Lesion Detection System Using Deep Architecture Neural Networks in the Lower Limb X-Ray Images

**Sepideh Amiri [1], Mina Akbarabadi [2], Shahnaz Rimaz [3,4], Fatemeh Abdolali [5], Reza Ahadi [6], Mohsen Afshani [7], Zahra Alaei Askarabad [8], Tahereh Kowsarirad [8], Sohrab Sakinehpour [8], Nazila Ayvazzadeh [9], Susan Cheraghi [3, 8*]**

[1] *Department of Computer Sciences, University of Copenhagen, Copenhagen, Denmark*

[2] *Department of Information Technology, Faculty of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran*

[3] *Radiation Biology Research Center, Iran University of Medical Sciences, Tehran, Iran*

[4] *Department of Epidemiology, School of Health, Iran University of Medical Sciences, Tehran, Iran*

[5] *Department of Radiology and Diagnostic Imaging, Faculty of Medicine and Dentistry, Alberta University, Edmonton, AB, Canada*

[6] *Department of Anatomical Science, Iran University of Medical Sciences, Tehran, Iran*

[7] *Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran*

[8] *Department of Radiation Sciences, Allied medicine faculty, Iran University of Medical Sciences, Tehran, Iran*

[9] *Radiation Research Center, Faculty of Paramedicine, AJA University of Medical Sciences, Tehran, Iran*

*Corresponding Author: Susan Cheraghi
Email: cheraghi.s@iums.ac.ir

Received: 14 January 2022 / Accepted: 30 May 2022

## Abstract

**Purpose:** Diagnosis of musculoskeletal abnormalities is critical because of the large number of people affected by these disorders worldwide. The recent advances in deep learning techniques show that convolutional neural networks can be a useful tool for the computer-aided detection of radiographic abnormalities. This study focuses on diagnosing musculoskeletal abnormalities in the lower extremities using X-Ray images by deep architecture neural networks.

**Materials and Methods:** The dataset contains 61,098 musculoskeletal radiographic images, including 42,658 normal and 18,440 abnormal images. Each image belongs to a single type of lower extremity radiography, including the toe, foot, ankle, leg, knee, femur, and hip joints, which were prepared with standard projection without artifacts and with high quality. A novel deep neural network architecture is proposed with two different scenarios that perform the lower extremity lesion diagnosis functions with high accuracy. The foundation of the proposed method is a deep learning framework based on the Mask Regional Convolutional Neural Network (R-CNN) and Convolutional Neural Network (CNN). The model with the best results incorporated the Mask R-CNN algorithm to produce the bounding box, followed by the CNN algorithm to detect the class based on that.

**Results:** The proposed model can identify different types of lower limb lesions by an Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve 0.925 with an operating point of 0.859 of sensitivity and a specificity of 0.893.

**Conclusion:** The results indicated that the consecutive implementation of Mask R-CNN and CNN has a higher efficiency than Mask R-CNN and CNN separately in lesion detection of lower limb X-ray images.

**Keywords:** X-Ray; Lower Limb; Deep Learning; Detection; Mask Regional Convolutional Neural Network.


DOI: https://doi.org/10.18502/fbt.v10i2.12221

Frontiers in BIOMEDICAL TECHNOLOGIES

## 1. Introduction

Over the past few decades, medical imaging techniques have helped diagnose and better treat diseases. Determining the exact location of a lesion on radiographic images is an important step in diagnosis and treatment. The interpretation of these images is the responsibility of experienced physicians and radiologists. In addition to the quality of images, which plays an important role in the accuracy of the diagnosis and depends on the imaging system and its protocols, the physician's experience has a significant impact on the accuracy of the diagnosis and can sometimes be extremely challenging for younger physicians. Diagnosis of musculoskeletal lesions is critical because more than 1.7 billion people are afflicted with musculoskeletal lesions worldwide. Also, severe pain and disability in the long term are among the most common symptoms of this lesion type [1]. In addition to significantly affecting the quality of life, these conditions impose a heavy economic burden on the health care system, resulting in total spending of 796 billion US$ in the United States between 2009 and 2011 [2]. Extensive pathological changes and human error might delay the correct diagnosis of these lesions. Computer systems and assisting software for doctors help make the right decision, especially in centers with a heavy workload and in emergency centers with few radiologists during the day and night shifts. These systems are used to improve physicians' work accuracy and reduce the time of radiographic interpretation. The recent advances in deep learning techniques show that Convolutional Neural Networks (CNNs) can be helpful in Computer-Aided Detection (CADe) of radiographic abnormalities [3-5].

There are systems for detecting and characterizing lesions, such as CADe and Diagnosis (CADx) [6]. However, despite the progress in their capabilities, these systems still have limitations due to using handmade features [7]. In recent years, numerous studies have been conducted on the presentation and application of artificial intelligence in medical diagnosis, and significant progress has been made in image classification using CNN [8, 9]. In a recent study, Pauwels *et al.* [10] compared CNNs with human observers. They concluded that CNNs were effective in detecting periapical lesions. Previous studies on deep learning methods for interpreting medical images have usually focused on identifying a single pathology in a particular body part. Gonella *et al.* [11] developed neural networks to process and classify lesions in mammograms. Savelli *et al.* [12] employed multi-depth CNNs for the detection of small lesions in medical images. However, studies that use transfer learning techniques, especially those that include trained models in the large natural image collection ImageNet, can detect abnormalities such as liver failure and osteoarthritis damage through images of body organs [13-18].

In some studies, mainly in chest x-ray assessments, scientists have designed models that can detect multiple pathologies with high accuracy. Nevertheless, these are usually suitable for highly homogeneous images of a single body part [4, 19, 20]. Similarly, some previously designed models can assess different types of medical images. However, they usually include the identification of defined pathologies [21, 22]. Consequently, there is limited research on developing general models that can express various body parts and diverse pathologies. This study aimed to diagnose musculoskeletal abnormalities in the X-ray images of lower extremities via deep architecture neural networks.

## 2. Materials and Methods

In the proposed model for detecting Musculoskeletal Disorders (MSDs), two scenarios were considered. In the first scenario, the Mask R-CNN technique was used to determine the bounding box, and the CNN algorithm was implemented for classification based on the bounding box. In the second scenario, both classification and bounding box generation operations were performed using the Mask R-CNN algorithm. Finally, the two scenarios were compared, and the evaluation results were presented. To the best of the authors' knowledge, no other research has used the combination of Mask R-CNN and CNN.

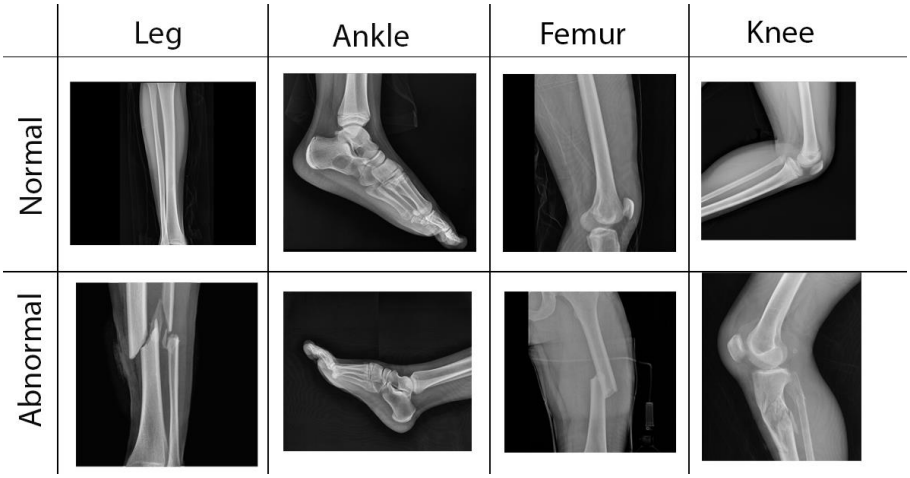
In this research, an extensive database of lower limbs (toe, foot, ankle, leg, knee, femur, and hip joint) was collected, and an intelligent model based on deep neural networks was trained. Also, the generalizability of the model in diagnosing a lesion was investigated. The musculoskeletal lesion of the lower limbs can be identified by a radiologist specializing in various fractures, dislocations, subluxation, and any other detectable abnormality appearance. This study incorporated a deep learning method to automate the bounding box of musculoskeletal lesions. A region-based Mask R-CNN with a ResNet-101 backbone was used for the bounding box. As in big data studies, collecting this labeled data

is important and valuable for this study. A large collection of radiographs, including 61,908 musculoskeletal images of the lower extremities, was presented. Each image was manually labeled as normal or abnormal by proficient radiologists because experienced radiologists are more reliable for labeling [23, 24]. The radiologists in this study have 10 to 16 years of experience.

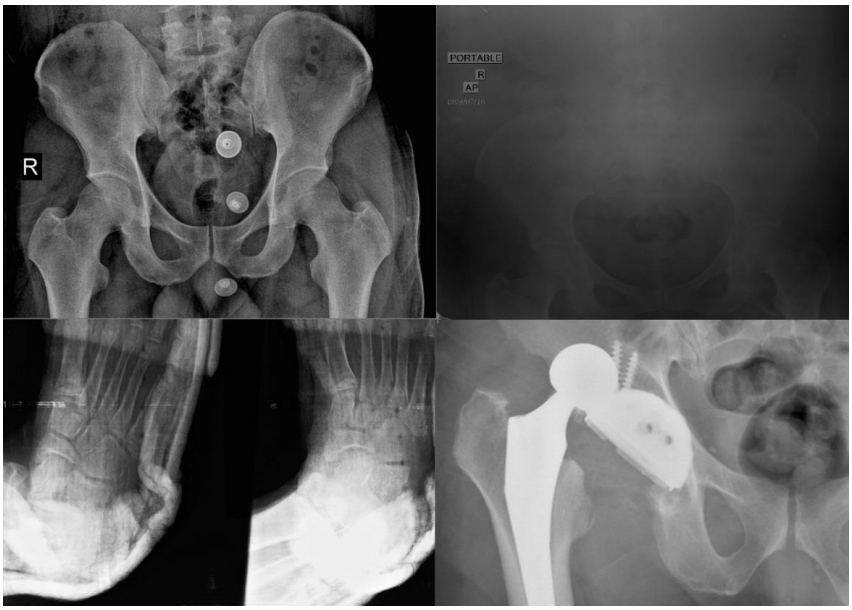
Diagnostic labeling of images as normal and abnormal was performed for images that were acceptable regarding X-ray penetration and image quality. This task required appropriate imaging processes and skilled radiographers. Images that contained any bone fracture, dislocation, sublocation, crack, tendon or ligament tear, and any lesions visible to the radiologist were labeled abnormal. Images without these symptoms were labeled normal. Images are shown in Figure 1.

## 2.1. Data Set

In this study, 61,098 X-rays with MSDs were collected retrospectively from the diagnostic imaging service at trauma centers in Tehran, Iran. The data were gathered from the department of medical imaging in Rasool-e-Akram, Haft-e-Tir, and Shafa Yahyaian hospitals in Tehran, Iran. The dataset contains 42,658 normal images and 18,440 abnormal images from male and female patients aged 18 to 82 years. Each image belongs to a single type of lower extremity radiography, including the toe, foot, ankle, leg, knee, femur, and hip joint (Table 1).




a




b

**Figure 1.** (a) Different X-ray images of the lower limbs in both normal and abnormal positions. (b) Examples of images with the exclusion criteria: artifacts, foreign objects, platinum, poor imaging quality and technique, and no diagnostic value for the radiologist

**Table 1.** The dataset containing 42,658 normal and 18,440 abnormal cases

| Study | Normal | Abnormal | Total |
|---|---|---|---|
| **Toe** | 4991 | 2706 | 7697 |
| **Foot** | 4769 | 1294 | 6063 |
| **Ankle** | 5647 | 2031 | 7678 |
| **Leg** | 8862 | 3169 | 12031 |
| **Knee** | 5860 | 2820 | 8680 |
| **Femur** | 4902 | 2964 | 7866 |
| **Hip joint** | 7627 | 3456 | 11083 |
| **Total No. of Studies** | 42658 | 18440 | 61098 |

All lower extremity X-ray images from digital imaging systems were standardized. First, patient information was removed to ensure anonymity. Then, the data were saved and archived in the same format Digital Imaging and Communications in Medicine (DICOM), and the quality of all the images met the requirements. Finally, all the images were classified based on the lower extremity to which they belonged and stored in a secure storage system. In this research, a data augmentation strategy is performed to overcome the lack of a large dataset. The human research ethics board of the Iran University of Medical Sciences approved the study (IR.IUMS.REC.1397.537).

All algorithms were performed on Python 3.7 using a GeForce GTX 1080 Ti graphics processor with 32 GB of memory. Several preprocessing steps were applied to the images before sending them to the network. First, the mean pixel value as the average vector in all training images was subtracted from the input image. Then, each image was rescaled and turned into a tensor.

Labeling is an important step in machine learning of medical images. This process was completed manually by two expert radiologists, who used patient information to specify the class of each image. In normal images, radiologists identified the whole image as a bounding box. For abnormal images, the bounding box of the lower limb abnormality area was annotated. The inclusion criteria were all lower limb images without contrast media and artifacts classified as either normal (without damage) or abnormal (damaged) groups. This study used the random sampling method. Exclusion criteria were images with poor quality and technique and a lack of diagnostic value to the radiologist. The database was then divided into training and test data.

The model parameters were adjusted using the training data. After fixing the parameters, the model was run using the test data as the input. The results of this step were compared to correct labels of images, determining the accuracy and sensitivity of the model, which are presented in detail in the Results section.

### 2.2. Network Architecture

Network architecture has attracted significant attention, leading to technological advances in various domains of image classification. In this study, the network took the input image size as 200 × 200. DenseNet-161, ResNet-50, and ResNet-101 architectures performed the initial convolution and max-pooling using 7 × 7 and 3 × 3 kernel sizes, respectively. An extensive random hyperparameter search was performed for each model in the validation set, followed by high-performance tuned models evaluated in the test set. In contrast to a traditional CNN [25], the input vector (x) passes through the CONV-ReLU-CONV series using a residual block in ResNet. Then, the output vector (F(x)) is added to the original input block x. In a traditional CNN, x is plotted directly to F(x), which contains no information about the original input [26-34].

CNN-based strategies such as Mask R-CNN [26] have exhibited increasing success in object detection tasks. Mask R-CNN [26] is the latest CNN architecture for object detection. Its objective for an image is to return the class label and coordinate the bounding box for each object within the image. Mask R-CNN has two stages. The first stage is to suggest areas that might be object-based in the input image, and the second stage, based on the outputs of the previous stage, predicts the object class, refines the bounding box, and creates a mask at the pixel level of the object [28]. The general model of the Mask R-CNN is shown in Figure 2. The Mask R-CNN architecture can involve ResNet-101, ResNet-50, MobileNet [29], U-Net [30], or Inception V2 [31] networks. As shown in Figures 2 and 3, the Region Proposal Network (RPN) suggests an area of an object bounding box.

## 3. Results

### 3.1. Network Training and Testing

The deep network was trained and evaluated using 61,098 X-ray images. The parameters trained in this model

are as follows. The learning rate is 0.0001 for the heads mask and R-CNN, and the learning rate is 0.001 for the Region Proposal Network (RPN) plus the backbone network. The momentum during the training is set to 0.9, and a stochastic gradient descent optimizer is used. The learning rate and momentum are adjusted by monitoring the loss level during the training process. With a low learning rate, training will progress extremely slowly as the network weight does very few updates. However, high learning rates can lead to undesirable divergence behavior in loss performance. In this study, reasonable learning rate values and momentum were found for Mask R-CNN and CNN through the experiments performed. Each part in Mask R-CNN and CNN is trained for 300 epochs. Most of the model parameters were selected based on the default Mask R-CNN parameters. In the two scenarios, the 61,098 images were divided into three groups: 70% (42,769), 10% (6,110 frames), and 20% (12,219 frames) for training, validation, and testing. The 20% of training data (8,553) were augmented during training by applying flips, rotations, shifts, and scaling. The number of training data became 51,322 images after augmentation. The whole dataset was augmented, with each model having a 20% random chance to be augmented and an 80% chance of not being augmented at all. The augmentation is done during runtime, meaning that each image has an 80% chance of not being augmented at all, a 5% chance of being only flipped, a 5% chance of being only rotated, a 5% chance of being only shifted, and a 5% chance of being only scaled. During the next epoch, each image is augmented again with the above strategies. Augmenting the images during runtime and not performing the full augmentation beforehand have an advantage; to achieve the same results, multiple new datasets should be created and the model trained on the combined large dataset. This study had no problem with data limitations. Therefore, by testing different numbers and comparing the results against the cost and time spent on resources, 20% was the most appropriate amount of augmentation. Mask R-CNN training time was approximately 14 hours, and CNN training time was approximately 11 hours. The existing GPU was GeForce GTX 1080 Ti.

In this study, the Mask-RCNN network was trained by applying ImageNet pre-trained weights [35]. Some studies in the literature used transfer learning with fine-tuning for better outcomes in MSDs classification and detection [36, 37]. A data augmentation method was used to generalize the trained model and avoid overfitting. It should be noted that some augmentation techniques such as shearing, elastic deformations, and adding noise did not prove effective in the present model. Therefore, this study used simple augmentation methods, including flips, rotations, shifts, and scaling. Segmentation and detection occur simultaneously on the Mask RCNN applied to the dataset. The performance of this multi-task model is checked against the validation data that are kept primarily for this assessment. In case of over-fitting, the regularization hyperparameter must be tuned. Developing, validating, tuning, and repeating were duplicated until the best accuracy against the validation data was achieved. Transfer learning and data augmentation strategies were applied to mitigate overfitting. In this study, the Mask-RCNN network was trained by applying ImageNet pre-trained weights. Transfer learning with fine-tuning was used for better outcomes in MSD classification and detection. Fine-tuning consists of unfreezing the part of the obtained model and re-training it on the new data with a very low learning rate. This has potential for improvements by incrementally adapting the pre-trained features to the new data.

### 3.2.Evaluation Metrics

Performance metrics in classification are fundamental in evaluating the quality of learning methods [32]. Accuracy, sensitivity, specificity, and area under the

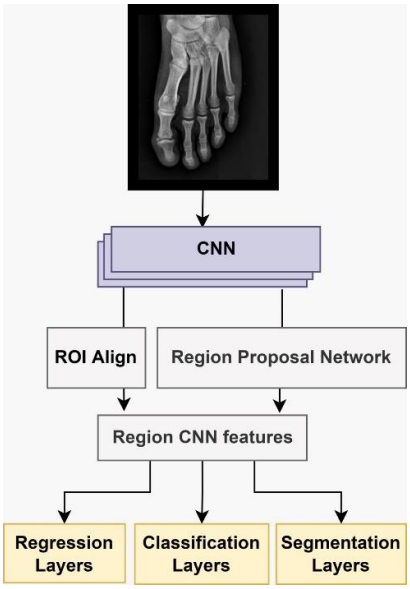



**Figure 2.** The two stages of the Mask R-CNN network. The first is the region proposal network, which predicts bounding boxes based on anchor boxes. The second stage involves an R-CNN detector classifying and generating pixel-level segmentation [21]

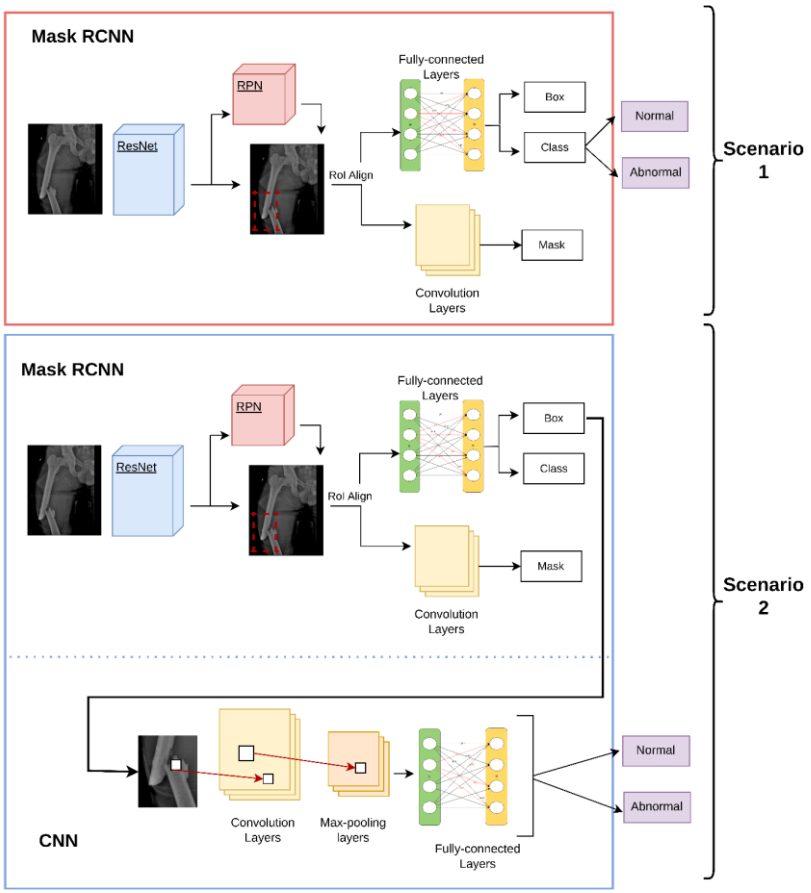



**Figure 3.** In the first scenario, both the classification and bounding box generation operations were performed using the Mask R-CNN algorithm. In the second scenario, the Mask R-CNN technique was used to determine the bounding box, and then the CNN algorithm was implemented for classification based on the bounding box

receiver operating characteristics (AUC-ROC) are metrics used for evaluating the present model. Accuracy is the most common and straightforward measure for assessing a classifier. It is defined as the degree of correct predictions of a model (or conversely, the percentage of misclassification errors) [32]. Equation 1 represents the accuracy formula.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (1)$$

Sensitivity and specificity are usually used in biomedical and medical applications and studies containing images and visual data. They evaluate the performance of a classifier in different classes [33]. Equations 2 and 3 display these two metrics.

$$Sensitivity = \frac{tp}{tp + fn} \quad (2)$$

$$Specificity = \frac{tn}{fp + tn} \quad (3)$$

As seen in the above equations, all these metrics are based on a confusion matrix that records correctly and incorrectly recognized examples for each class [33]. This matrix is shown in Table 2.

**Table 2.** Confusion matrix

| Class/ recognized | As positive | As negative |
|---|---|---|
| **Positive** | True positive | False negative |
| **Negative** | False positive | True negative |

The AUC-ROC is not dependent on the actual predicted values but relies only on the ordering of the cases. In this method, it is practical to configure how well the positive cases are ordered before the negative ones, and it is possible to consider this as the outline of the model performance throughout all available thresholds [38].

Figure 4 shows the efficiency of the proposed method. It should be noted that every result produced by the Mask R-CNN consists of three components: a confidence value, the coordinates of a bounding box, and a

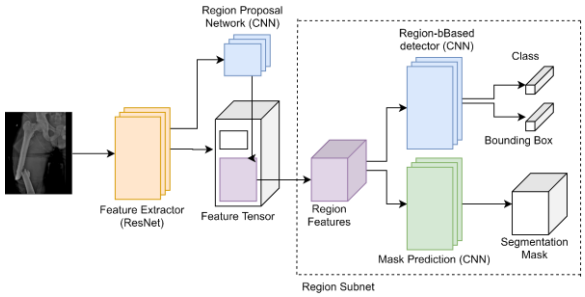



**Figure 4.** A detailed block diagram of the proposed framework

segmentation mask. Tables 3 and 4 represent the results of applying both proposed Mask R-CNN and CNN models and the evaluation metrics (accuracy, sensitivity, specificity, and AUC-ROC). The values in Table 5 demonstrate that the proposed Mask R-CNN and CNN outperformed Mask R-CNN in achieving highly accurate detection results. The multi-task loss function of Mask R-CNN is designed to combine the loss of classification, localization, and segmentation mask on each sampled RoI. The loss function for classification and box regression is the same as Faster R-CNN, that is:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i N_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{cls}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (4)$$

The first term is the classification loss, and the second term is the regression loss of bounding boxes. Additionally, regularizing the loss function improved the success of the Mask R-CNN framework.

**Table 3.** Lower limb abnormalities detection accuracy of Mask R-CNN

| Backbone | Accuracy | Sensitivity | Specificity | AUC-ROC |
|---|---|---|---|---|
| **ResNet-50** | 0.819 | 0.776 | 0.962 | 0.889 |
| **ResNet-101** | 0.841 | 0.825 | 0.867 | 0.894 |
| **DenseNet-161** | 0.791 | 0.713 | 0.980 | 0.878 |

**Table 4.** Lower limb abnormalities detection accuracy of Mask R-CNN and CNN

| Backbone | Accuracy | Sensitivity | Specificity | AUC-ROC |
|---|---|---|---|---|
| **ResNet-50** | 0.819 | 0.776 | 0.962 | 0.889 |
| **ResNet-101** | 0.841 | 0.825 | 0.867 | 0.894 |
| **DenseNet-161** | 0.791 | 0.713 | 0.980 | 0.878 |

**Table 5.** Lower limb abnormalities detection accuracy

| Method | Accuracy | sensitivity | Specificity | AUC-ROC |
|---|---|---|---|---|
| **Mask R-CNN** | 0.841 | 0.825 | 0.867 | 0.894 |
| **Mask R-CNN and CNN** | 0.891 | 0.859 | 0.893 | 0.925 |

## 4. Discussion

This study presented several supervised deep learning approaches for automated binary classification of abnormalities in lower extremity radiographs. Also, the utilization of Mask R-CNN and CNN in identifying a range of abnormalities was explored across multiple types of lower extremity radiographs. The application of Mask R-CNN and CNN for detection and classification was compared to applying only Mask R-CNN. As evidenced, the proposed Mask R-CNN and CNN exhibited higher efficiency than the other approach.

One of the models achieved an AUC-ROC of 0.925 on lower extremity radiographs. This is a promising result due to the high degree of variability in input images, the inclusion of multiple body parts, and the presence of diverse and unannotated abnormalities [28]. The results showed that using Mask R-CNN to produce the bounding box and CNN for classification based on the bounding box achieved better results than Mask R-CNN for classification and detection. A deeper network can achieve higher performance in the natural image domain. Also, as Tables 4 and 5 show, the ResNet101 backbone delivered the best performance. ResNet models define residual blocks that are made up of multiple convolution processes and have skip connections for better performance of the model. One of the advantages of the proposed framework in working with deeper layers is that the accuracy of the classification model based on the ROI is high, and the marginal features are less important. It is much easier for the model to predict the abnormality of an image. Any additional information can achieve a more accurate classification. Most existing models have multiple outputs that perform both classification and localization in parallel. However, this study used the localization output for more accurate classification, reducing the detection space, so the network needs to examine fewer features.

Related studies have CNNs with various model architectures. Despite the differences between the model structures, no statistically significant difference was observed in performance in three CNN architectures with different architecture and depth (DenseNet-161, ResNet-101, and ResNet-50). It can be concluded that a model that requires less training time and computational power can be helpful for future deep learning abnormality prediction, and the training data should also be increased. The three models were compared on the ROC curve for Mask R-CNN and CNN architectures, which plots model specificity against sensitivity (the proportion of the correctly identified negatives against the proportion of the correctly identified positives). Since the classification in this study is binary (normal and abnormal), the sigmoid layer is used. In the ROC curve in Figure 5, the vertical axis shows the rate of detected true positives, and the horizontal axis shows the rate of detected false positives. The models output the probability of abnormality in a musculoskeletal study, and the ROC curve is generated by changing the thresholds used for the classification boundary. The AUC-ROC of the DenseNet-161 model is 0.884, the ResNet-101 model is 0.925, and the ResNet-50 model is 0.906.

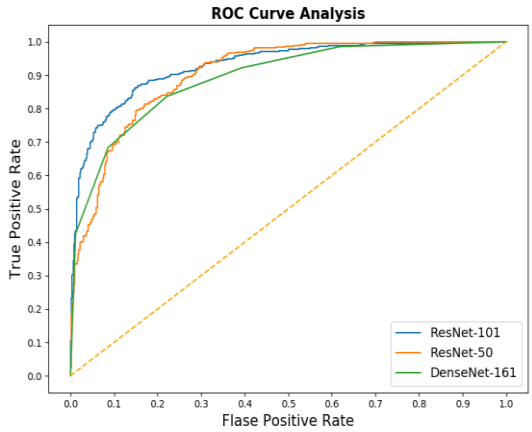



**Figure 5.** ROC curves are evaluated on sensitivity (the proportion of the correctly identified positives) and specificity (the proportion of the correctly identified negatives) for Mask R-CNN and CNN with tree models, including ResNet-101 with blue, ResNet-50 with orange, and DenseNet-161 with green

In a similar study, Varma *et al.* examined the CNN tool for diagnosing general abnormalities in lower extremity radiography on a dataset containing 93,455 radiographs. They achieved an AUC-ROC of 0.880 (sensitivity = 0.714, specificity = 0.961) in the abnormal classification [39]. Rajpurkar *et al.* presented a collection of upper limb musculoskeletal radiographs containing 40,561 images. Their deep learning model achieved an AUROC of 0.929, with a sensitivity of 0.815 and specificity of 0.887 [3]. As can be seen, the results of the present research have higher accuracy. Pradhan *et al.* worked on human upper limb bone recognition. They employed a deep convolutional neural network for recognition and achieved 91.37% accuracy for the classification of human upper limb bones [40]. Shao and Wang introduced a two-stage technique for the classification of a human upper limb bones dataset. They achieved the maximum accuracy of 88.5% for the SENet154 model in humerus images, and the highest accuracy for the DenseNet201 model was again in humerus images with 90.94% [41]. Rohrbach *et al.* used the VGG-16 model with transfer learning to classify and identify the bone destruction marking for rheumatoid arthritis. Due to their highly imbalanced dataset, they achieved 77.5% accuracy [42]. Another study examined the binomial categorization of infant elbow fractures. They revised the Xception architecture to receive monophonic gray-scale input and attained 95% AUC, with 88% accuracy in elbow classification [43].

In recent years, following the improved computational power, the entry barrier for deep learning has lowered. However, access to large and labeled radiograph datasets is still challenging [44]. Transferring learning across large, publicly available datasets is a common method for addressing this problem [45]. In this study, all models were pre-trained on ImageNet to improve the model performance. The presented results have several important clinical implications. First, the results are not limited to a single body part (e.g., only the hip) or a single pathology (e.g., fractures only) [46]. Moreover, the model can rapidly identify routine examinations with a preliminarily reading as ‘normal’ or ‘abnormal.’ These advantages can help radiologists spend more time on abnormal and complex cases and simplify an increased throughput [39]. Studies have shown that fatigue in radiologists increases by the end of the workday and can be intensified by increasing patient numbers [47]. Therefore, a model that assists the radiologist in localizing the abnormality on an image can decrease diagnostic errors.

This study had several limitations that should be addressed in future research. First, the proposed Mask R-CNN cannot handle a minimal number of MSDs. One of the primary challenges for radiologists is detecting MSDs [28]. Figure 6 shows examples of successful

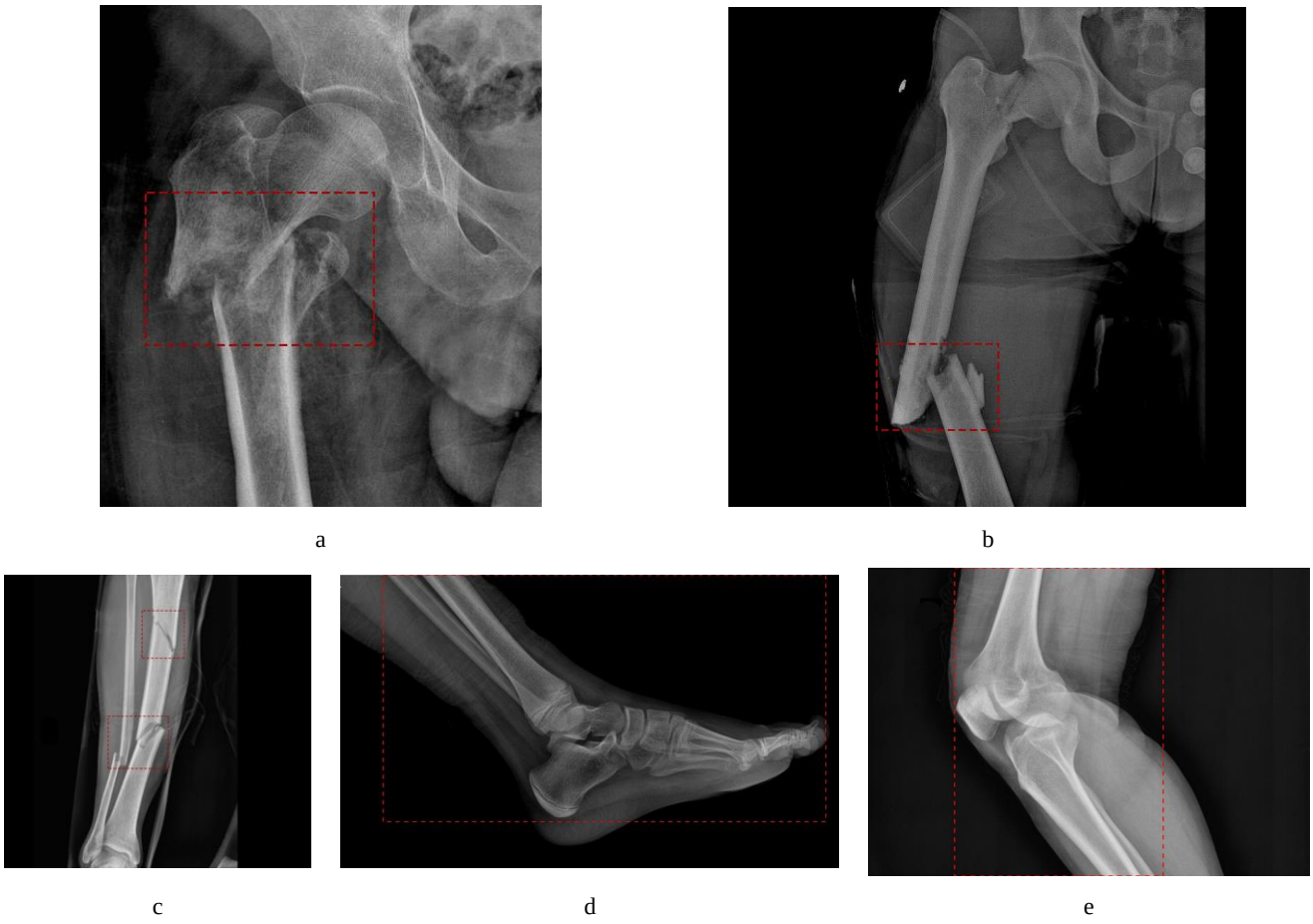



**Figure 6.** (a), (b), and (c) are examples of cases for lower limb abnormalities detection; (d) and (e) are examples of cases for lower limb normality detection. (a) Hip joint, (b) Femur, (c) Leg, (d) Foot, and (e) Ankle successfully diagnosed. Bounding boxes are denoted by red color

detection of MSDs. Minimizing the overfitting problem is a challenge in developing a deep learning MSDs detection approach. Transfer learning and data augmentation strategies were applied to mitigate overfitting.

Second, only binary classes were considered in this study, and the model's performance on important subclasses was not evaluated. However, subtle abnormalities may be obscured by downsampling images to a 224 × 224 matrix.

Future research can develop better strategies to overcome the overfitting problem. Mask R-CNN may become a favorable technique for MSD detection by having access to a better training dataset [28]. Therefore, future studies can retrain the proposed model with a larger dataset.

In conclusion, the present study demonstrated that deep learning models could identify abnormalities in lower extremity radiographs at performance levels of clinical importance. With further preclinical assessment, these approaches may eventually allow for rapid and automated triaging of patients with MSDs.

## Acknowledgement

This work was supported by the Research Chancellor of Iran University of Medical Sciences [grant number 97-02-223-33781]; the Novo Nordisk Foundation [grant Number NNF20OC0062056].

## References

1- A. D. Woolf and B. Pfleger, "Burden of major musculoskeletal conditions." (in eng), *Bull World Health Organ*, Vol. 81 (No. 9), pp. 646-56, (2003).

2- E. Yelin, S. Weinstein, and T. King, "The burden of musculoskeletal diseases in the United States." (in eng), *Semin Arthritis Rheum*, Vol. 46 (No. 3), pp. 259-60, Dec (2016).

3- Pranav Rajpurkar *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs." *arXiv preprint arXiv:1712.06957*, (2017).

4- P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists." (in eng), *PLoS Med*, Vol. 15 (No. 11), p. e1002686, Nov (2018).

5- Yee Liang Thian, Yiting Li, Pooja Jagmohan, David Sia, Vincent Ern Yao Chan, and Robby T Tan, "Convolutional neural networks for automated fracture detection and localization on wrist radiographs." *Radiology: Artificial Intelligence*, Vol. 1 (No. 1), p. e180001, (2019).

6- Macedo Firmino, Giovani Angelo, Higor Morais, Marcel R Dantas, and Ricardo Valentim, "Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy." *Biomedical engineering online*, Vol. 15 (No. 1), pp. 1-17, (2016).

7- Junji Shiraishi, Qiang Li, Daniel Appelbaum, and Kunio Doi, "Computer-aided diagnosis and artificial intelligence in clinical imaging." in *Seminars in nuclear medicine*, (2011), Vol. 41 (No. 6): *Elsevier*, pp. 449-62.

8- Kai-Lung Hua, Che-Hao Hsu, Shintami Chusnul Hidayati, Wen-Huang Cheng, and Yu-Jen Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique." *OncoTargets and therapy*, Vol. 8(2015).

9- Petros-Pavlos Ypsilantis *et al.*, "Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks." *PloS one*, Vol. 10 (No. 9), p. e0137036, (2015).

10- Ruben Pauwels *et al.*, "Artificial intelligence for detection of periapical lesions on intraoral radiographs: comparison between convolutional neural networks and human observers." *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, Vol. 131 (No. 5), pp. 610-16, (2021).

11- Gloria Gonella, Marco Paracchini, Elisabetta Binaghi, and Marco Marcon, "Breast Lesion Detection from Mammograms Using Deep Convolutional Neural Networks." in *Proceedings of the 2020 European Symposium on Software Engineering*, (2020), pp. 120-24.

12- Benedetta Savelli, Alessandro Bria, Mario Molinara, Claudio Marrocco, and Francesco Tortorella, "A multi-context CNN ensemble for small lesion detection." *Artificial Intelligence in Medicine*, Vol. 103p. 101749, (2020).

13- Joseph Antony, Kevin McGuinness, Noel E O'Connor, and Kieran Moran, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks." in *2016 23rd International Conference on Pattern Recognition (ICPR)*, (2016): *IEEE*, pp. 1195-200.

14- Lei Bi, Jinman Kim, Ashnil Kumar, and Dagan Feng, "Automatic liver lesion detection using cascaded deep residual networks." *arXiv preprint arXiv:1704.02703*, (2017).

15- Ruikai Zhang *et al.*, "Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain." *IEEE journal of biomedical and health informatics*, Vol. 21 (No. 1), pp. 41-47, (2016).

16- Varun Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama*, Vol. 316 (No. 22), pp. 2402-10, (2016).

17- Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique." *IEEE transactions on medical imaging*, Vol. 35 (No. 5), pp. 1153-59, (2016).

18- Daniel S Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning." *Cell*, Vol. 172 (No. 5), pp. 1122-31. e9, (2018).

19- Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang, "Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays." in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, (2018), pp. 103-10.

20- Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan, "Chest pathology detection using deep learning with non-medical training." in *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, (2015): *IEEE*, pp. 294-97.

21- Jakub Olczak *et al.*, "Artificial intelligence for analyzing orthopedic trauma radiographs: deep learning algorithms—are they on par with humans for diagnosing fractures?" *Acta orthopaedica*, Vol. 88 (No. 6), pp. 581-86, (2017).

22- Robert Lindsey *et al.*, "Deep neural network improves fracture detection by clinicians." *Proceedings of the National Academy of Sciences*, Vol. 115 (No. 45), pp. 11591-96, (2018).

23- Hieu H. Pham, Tung T. Le, Dat Q. Tran, Dat T. Ngo, and Ha Q. Nguyen, "Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels." *Neurocomputing*, Vol. 437pp. 186-94, 2021/05/21/ (2021).

24- Stephen Waite *et al.*, "Analysis of perceptual expertise in radiology–Current knowledge and a new perspective." *Frontiers in human neuroscience*, Vol. 13p. 213, (2019).

25- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional

neural networks." *Advances in neural information processing systems*, Vol. 25(2012).

26- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn." in *Proceedings of the IEEE international conference on computer vision*, (2017), pp. 2961-69.

27- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition." in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 770-78.

28- Fatemeh Abdolali, Jeevesh Kapur, Jacob L Jaremko, Michelle Noga, Abhilash R Hareendranathan, and Kumaradevan Punithakumar, "Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks." *Computers in Biology and Medicine*, Vol. 122p. 103871, (2020).

29- Andrew G Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861*, (2017).

30- Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation." in *International Conference on Medical image computing and computer-assisted intervention*, (2015): *Springer*, pp. 234-41.

31- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision." in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 2818-26.

32- César Ferri, José Hernández-Orallo, and R Modroiu, "An experimental comparison of performance measures for classification." *Pattern recognition letters*, Vol. 30 (No. 1), pp. 27-38, (2009).

33- Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." in *Australasian joint conference on artificial intelligence*, (2006): *Springer*, pp. 1015-21.

34- Jacob Cohen, "A coefficient of agreement for nominal scales." *Educational and psychological measurement*, Vol. 20 (No. 1), pp. 37-46, (1960).

35- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database." in *2009 IEEE conference on computer vision and pattern recognition*, (2009): *Ieee*, pp. 248-55.

36- Jianning Chi, Ekta Walia, Paul Babyn, Jimmy Wang, Gary Groot, and Mark Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network." *Journal of digital imaging*, Vol. 30 (No. 4), pp. 477-86, (2017).

37- Tianjiao Liu, Shuaining Xie, Jing Yu, Lijuan Niu, and Weidong Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features." in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2017): *IEEE*, pp. 919-23.

38- Sarang Narkhede, "Understanding auc-roc curve." *Towards Data Science*, Vol. 26 (No. 1), pp. 220-27, (2018).

39- Maya Varma *et al.*, "Automated abnormality detection in lower extremity radiographs using deep learning." *Nature Machine Intelligence*, Vol. 1 (No. 12), pp. 578-83, (2019).

40- Nitesh Pradhan, Vijaypal Singh Dhaka, and Himanshu Chaudhary, "Classification of human bones using deep convolutional neural network." in *IOP conference series: materials science and engineering*, (2019), Vol. 594 (No. 1): *IOP Publishing*, p. 012024.

41- Yunxue Shao and Xin Wang, "A two stage method for abnormality diagnosis of musculoskeletal radiographs." in *International Conference on Pattern Recognition and Artificial Intelligence*, (2020): *Springer*, pp. 610-21.

42- Janick Rohrbach, Tobias Reinhard, Beate Sick, and Oliver Dürr, "Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks." *Computers & Electrical Engineering*, Vol. 78pp. 472-81, (2019).

43- Jesse C Rayan, Nakul Reddy, J Herman Kan, Wei Zhang, and Ananth Annapragada, "Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making." *Radiology: Artificial Intelligence*, Vol. 1 (No. 1), p. e180015, (2019).

44- Gabriel Chartrand *et al.*, "Deep learning: a primer for radiologists." *Radiographics*, Vol. 37 (No. 7), pp. 2113-31, (2017).

45- Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, Vol. 27(2014).

46- William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer, "Detecting hip fractures with radiologist-level performance using deep neural networks." *arXiv preprint arXiv:1711.06504*, (2017).

47- Elizabeth A Krupinski, Kevin S Berbaum, Robert T Caldwell, Kevin M Schartz, and John Kim, "Long radiology workdays reduce detection and accommodation accuracy." *Journal of the American College of Radiology*, Vol. 7 (No. 9), pp. 698-704, (2010).