Ensemble learning-based abnormality diagnosis in wrist skeleton radiographs using densenet variants voting

Sajid Khan[1,2,*], Faiqa Arshad[2], Maryam Zulfiqar[2], Muhammad Asif Khan[2],
Sufyan Ali Memon[3]
[1]*Center of Excellence for Robotics, Artificial Intelligence, and Blockchain, Sukkur IBA University,* Sukkur, Pakistan
[2]*Dept. of Computer Science, Sukkur IBA University, Sukkur, Pakistan*
[3]*Defense systems engineering dept. Sejong university, Seoul, Korea*

**Corresponding author: sajidkhan@iba-suk.edu.pk*

Abstract

Almost one out of five people, including children, suffers from musculoskeletal disorders. It is the second leading cause of disability worldwide. It affects the musculoskeletal system's major areas, represented by the shoulder, forearm, and wrist. It causes severe pain, joint noises, and disability. To detect the abnormality, the radiologist analyzes the patient's anatomy through X-rays of different views and projections. To automatically diagnose the abnormality in the musculoskeletal system is a challenging task. Previously, various researchers detected the abnormality in the musculoskeletal system from radiographic images by using several deep learning techniques. They used a capsule network, 169-layer convolutional neural network, and group normalized convolutional neural network in musculoskeletal abnormality detection. However, to propose methods for improving abnormality detection, further work needs to be done because the accuracy of the conventional methods is far away from 90%. This paper presents an ensemble learning-based classification system for detecting abnormality in wrist radiographs. Tags in radiographs may result in learning noisy features hence reducing the performance. Therefore, tags are segmented and removed using UNet trained on the annotated ground truths. Segmented images are then used for voting-based diagnosis. The simulation results show that the proposed methodology improves testing accuracy by 1.5%-4.5% compared to the available wrist abnormality detection methods. The proposed methodology can be used for any kind of musculoskeletal abnormality detection.

Keywords: Abnormality diagnosis; ensemble learning; medical imaging; radiograph processing; segmentation

1. Introduction

Biomedical image processing is a revolutionized field that enfolds medical signal gathering, image formation, image processing, deep learning, and diagnosis of several medical problems based on the features extracted from the radiographic images (Nawarathne *et al.*, 2022; Saadawy *et al.*, 2020). Its application of deep learning offers better interpretations of the images and provides high accuracy. Deep learning is a powerful field that uses a range of neural networks to perform segmentation, object detection, and classification (Prajna & *Nath*, 2021; Minaee *et al.*, 2021; Pal *et al.*, 2021; Wang *et al.*, 2021; Korot *et al.*, 2021; Sungheetha & *Sharma*, 2021; Zhang *et al.*, 2021).

A prevalent deep learning method and convolutional neural network are widely used in medical imaging. A convolutional neural network is an excellent feature extractor. Therefore, it is used to classify images that can prevent complex and expensive feature engineering (Yadav & *Jadhav*, 2019).

In general, medical imaging techniques visualize the body's interior and specific organs. Medical imaging covers conventional radiography, MRI, ultrasound, endoscopy, and thermography disciplines (Zhou *et al.*, 2021). Conventional radiography uses X-ray or gamma rays to view the internal form of an object. It is used to detect the presence or absence of disorders such as musculoskeletal disorders (Aal *et al.*, 2018). Musculoskeletal disorders are injuries or disorders in muscles, joints, ligaments, and tendons that affect the movement of the human body and often cause disability (Yang et al., 2019). More than 1.7 billion people worldwide are suffering from musculoskeletal disorders. It is the second leading cause of disability worldwide (Yang *et al.*, 2019). The prevalence of these disorders varies by age, as one in two people over the age of 18 and three in four people ages 65 and over are affected. In 2009–2010, almost 105 million cases of musculoskeletal disorders were reported to physician's clinics, hospitals, and emergency departments (Chada, 2019).The early detection and treatment of musculoskeletal disorders can prevent disability. The scarcity of radiologists in primary care hospitals leads to late or no abnormality diagnosis and may result in a patient's permanent disability (Bhargavan *et al.*, 2009).

Researchers have proposed deep learning-based approaches to address the issues of efficient and effective musculoskeletal abnormality detection (Saadawy *et al.*, 2020; Aal *et al.*, 2018; Yang *et al.*, 2019; Chada, 2019; Goyal *et al.*, 2020; Solovyov & *Solovyov*, 2020; Irmakci *et al.*, 2019; Rajpurkar *et al.*, 2017; Saif *et al.*, 2019; Tantawi *et al.*, 2020). Those approaches process bone radiographs in detecting musculoskeletal abnormality. These approaches aimed to provide expert-level accuracy as false detection may increase the expense of treatment, dissatisfaction, and poor treatment of the patient hence increasing severity and disability. Therefore, it is challenging to obtain an accurate diagnosis that results in efficient and timely patient treatment.

An ensemble learning-based wrist abnormality classification method is proposed in this paper. Ensemble learning is a voting strategy from more than one model and combines them using hard voting (Sagi & *Rokach*, 2018) or soft voting (Tasci *et al.*, 2021). Hard voting counts the number of votes from each model and follows the majority's decision, whereas soft voting combines the probabilities of Softmax layers of each model. Ensemble learning trains multiple models and combines their predictions thus outperforming the predictive performance of a single model (Sagi & *Rokach*, 2018). Few researchers (Mondol *et al.*, 2019) have used ensemble learning for musculoskeletal abnormality detection to achieve better accuracy in the diagnosis.

The proposed methodology comprises some steps and follows hard voting. The explanation of the contribution requires it to break down into two parts. The first part explains how the networks are trained, whereas the second part is how trained networks are utilized for the diagnosis. For the first part, a UNet trained on annotated ground truth is used to segment the tags and remove them from the radiographs. Tags removal is essential as the classifiers may learn them noisy features. Image and edge enhancement using contrast limited adaptive histogram equalization (CLAHE), smoothing, and unsharp masking is applied to the intensity images before UNet training. A separate approach is followed as far as creating ground truth and training UNet for the segmentation. For ground truth creation, an interactive script is created using Otsu's thresholding method (Otsu, 1979), Maximally Stable Extremal Regions (MSER) (Chen *et al.*, 2011), and connectivity information instead of using available contour-based software for quick annotation. A total of 800 ground truth images are generated using that script, and the UNet is trained and tested on them to efficiently remove noisy tags. Using the radiographs with tags removed, three other networks such as DenseNet-169, DenseNet-201, and DenseNet-121, are trained. Augmentation is applied before training the UNet and the other three classifiers. All the networks are trained and tested on wrist radiographs of the MUsculoskeletal RAdiographs (MURA) dataset. However, the same methodology can be used for radiographs of other body parts as all the properties and features are similar. For the second part, this paper suggests using image enhancement followed by noisy tag removal with the help of the trained UNet. Given that the MURA dataset contains multiple orientations and views of radiographs of the same case, as a pre-classification approach, this paper suggests concatenating all these radiographs to get full advantage of various orientations. The final step is to decide whether a case is normal or abnormal based on the voting of three trained classifiers.

The rest of the paper is organized as follows: Section 2 describes some conventional deep learning-based solutions for musculoskeletal abnormality diagnosis. Section 3 describes the proposed voting-based approach. Section 4 shows the proposed approach's results, experiments, and discussion, followed by the conclusion and future work in Section 5.

2. Related Work

This section discusses some of the conventional state-of-the-art methods for musculoskeletal abnormality detection. This first describes the dataset used by researchers, followed by the methodologies.

Lindsey *et al.* (Lindsey *et al.*, 2018) created a dataset that comprises a total of 1,35,845 radiographs with localized fracture regions. Out of those radiographs, 34,990 radiographs projection were of the lateral wrist or posterior-anterior. However, this dataset hasn't been made publicly available. Lower Extremity RAdio graphs (LERA) is another dataset used by a few researchers that contain 1,299 radiographic images of the foot, knee, hip, and ankle. Rajpurkar *et al.* (Rajpurkar *et al.*, 2017) introduced the MURA dataset. It is the largest publicly available dataset that contains 40,561 radiographic images from 14,863 studies and is used mainly by researchers. Out of 14,863 studies, the wrist has 3,697 studies that are maximum compared to others. Among these wrist studies, 237 studies are kept for validation.

Varma *et al.* (Varma *et al.*, 2019) trained ResNet-50, DenseNet-161, and ResNet-101 on LERA dataset. They performed an abnormality diagnosis on the knee, ankle, hip, and foot. With some parameter tuning, the best accuracy they got was 85%.

Rajpurkar *et al.* (Rajpurkar *et al.*, 2017) used DenseNet-169 layers architecture in the musculoskeletal abnormality detection. They replaced the last fully connected layer of DenseNet-169 with the layer with a single output that provides the probability of belonging to either a normal or abnormal case. They applied image normalization and resizing as pre-processing steps followed by augmentation to address dataset imbalance. Solovyov and Solovyov (Solovyov & *Solovyov*, 2020)) also used DenseNet- 169 and improved the accuracy of diagnosis to 86.2% using fine-tuning of the parameters.

Saif *et al.* (Saif *et al.*, 2019)) proposed a capsule network-based classification to detect the abnormality in the musculoskeletal system. After training on 64×64, 128×128, and 224×224 sized images, they evaluated their network and observed that 224×224 provided the best training accuracy of 96% for the wrist radiographic images. However, the author did not discuss the testing results. Therefore, this methodology may not be used as a standard for evaluation as sometimes over-fitting results in highest training but lowest or moderate testing accuracy.

Goyal *et al.* (Goyal *et al.*, 2020) designed GnCNNr (Group Normalized Convolutional Neural Networks with Regularization). The GnCNNr utilizes group normalization, weight standardization, and cyclic learning rate scheduler (regularizer) for improved performance. The group normalization layer is used to form sets from channels and calculates their variance and mean for their standardization. Instead of fixing, a cyclic learning rate scheduler keeps the learning rate cycling between the upper bound and the lower bound.

Saadawy *et al.* (Saadawy *et al.*, 2020) detect the abnormality in muscles using a two-stage approach of the MobileNet model. In the first stage, they classified the skeleton into seven classes according to the various parts of the anatomy. Those parts are the shoulder, humerus, elbow, forearm, wrist, hand, and finger. In the second stage, they used seven other classifiers designed for each class of the first stage to detect whether the case is normal or abnormal. As pre-processing steps, they used adaptive histogram equalization, data augmentation, and normalization.

Mondol *et al.* (Mondol *et al.*, 2019) diagnosed the abnormalities in the elbow, wrist, humerus, and finger using soft voting-based ensemble learning. They used VGG-19 and Resnet architecture for ensemble learning to get a diagnostic accuracy of 87.86% for abnormality of the wrist. Nazim *et al.* (Nazim *et al.*, 2021) also used ensemble learning using CNN and LSTM (Long Short-Term Memory) based models. Their best training accuracy for the LERA dataset on knee abnormality detection is 99.4737%, whereas the best accuracy of detection on the MURA dataset for the wrist is 78.1086%.

Finally, we may draw the following observations from the literature review: First, the MURA dataset has been used in most of the literature discussed in this section (Rajpurkar *et al.*, 2017; Saif *et al.*, 2019; Saadawy *et al.*, 2020; Goyal *et al.*, 2020; Mondol *et al.*, 2019; Nazim *et al.*, 2021). The MURA dataset contains two or more radiographs taken from different orientations for the same case. How-

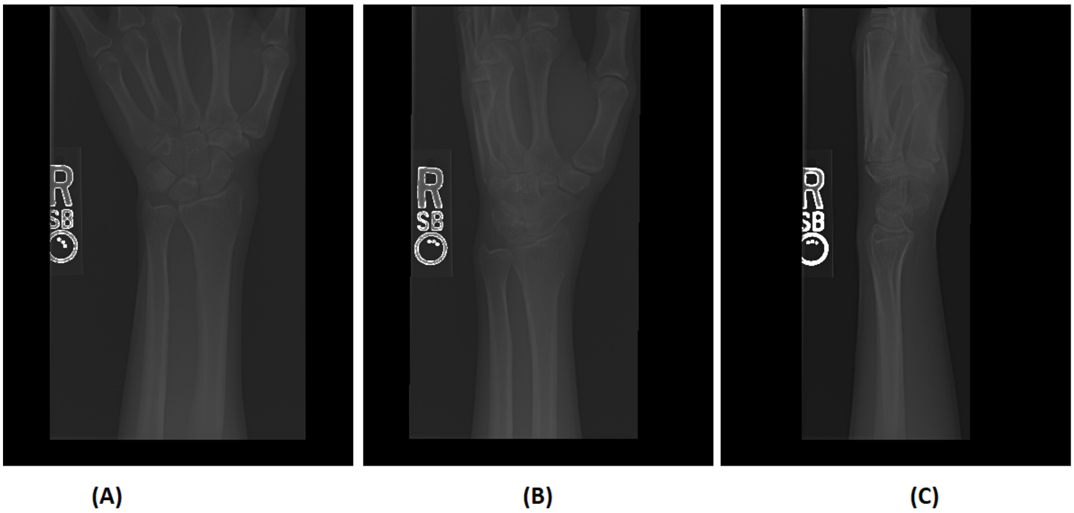

Fig. 1. An example of multiple orientation images in the MURA dataset.

ever, researchers only used a single image instead of getting advantages from the images of multiple orientations. An example is shown in Figure 1. Second, if the radiographic tags are not removed, the classifier may detect noisy and useless features thus leading to less accurate results. Lastly, during the augmentation process, rotation is performed hence reducing the performance of CNN (Saif *et al.*, 2019).

3. Proposed Method

This paper proposes an ensemble learning-based classification method that performs better than conventional methods. The proposed methodology introduces better pre-processing, segmentation using annotated ground truths, and voting that combines results in improved classification. Wrist images from the MURA dataset are used for all the simulations, training, and testing.

This section is divided into two subsections. 3.1 discusses the flow of classification using the proposed methodology, whereas 3.2 explains the ground truth annotation for segmentation along with training the UNet and voting classifiers

3.1 Flow of classification

This subsection explains how a case should be diagnosed using the proposed methodology. Figure 2 discusses the flowchart for the diagnosis of wrist radiographs as either normal or abnormal. The MURA dataset contains multiple radiographs taken from different orientations for a single case, as shown in Figure 1. Therefore, the expected input to the system is n number of images described using $I_1, I_2, \ldots, I_n$.

Image pre-processing is applied to the input image to enhance the image, particularly the edges that represent possible fractures. To enhance the overall image, CLAHE should be applied first. However, CLAHE enhanced the unwanted details, as can be seen in Figure 3(B). Applying edge enhancement techniques such as unsharp masking to such an image will also sharpen those unwanted details. Therefore, the CLAHE result is subjected to the application of smoothing using a 5×5 box filter in suppressing the minor noisy details. Figure 3(C) shows the smoothing application results. After the smoothing filter is applied, unsharp masking with σ=3 is applied to obtain an enhanced image as shown in Figure 3(D). Figure 3(E) shows the magnified absolute difference between Figures 3(A) and 3(D). The magnification factor is kept to 5 to make the change visible. It can be seen from the figure that the unsharp masking boosted the edges of the image more than the other unwanted details.

The binary semantic segmentation is applied after applying unsharp masking to the image using the trained UNet. The process of training the UNet is explained in the following subsections. Segmentation is applied to remove the tags that may lead the classifier to learn noisy features, resulting in misclassification. Results of the application of the UNet-based noisy tag removal are shown in Figure 4. For each detected segment by UNet, a rectangular region surrounding it is assigned the intensity values of zero, as

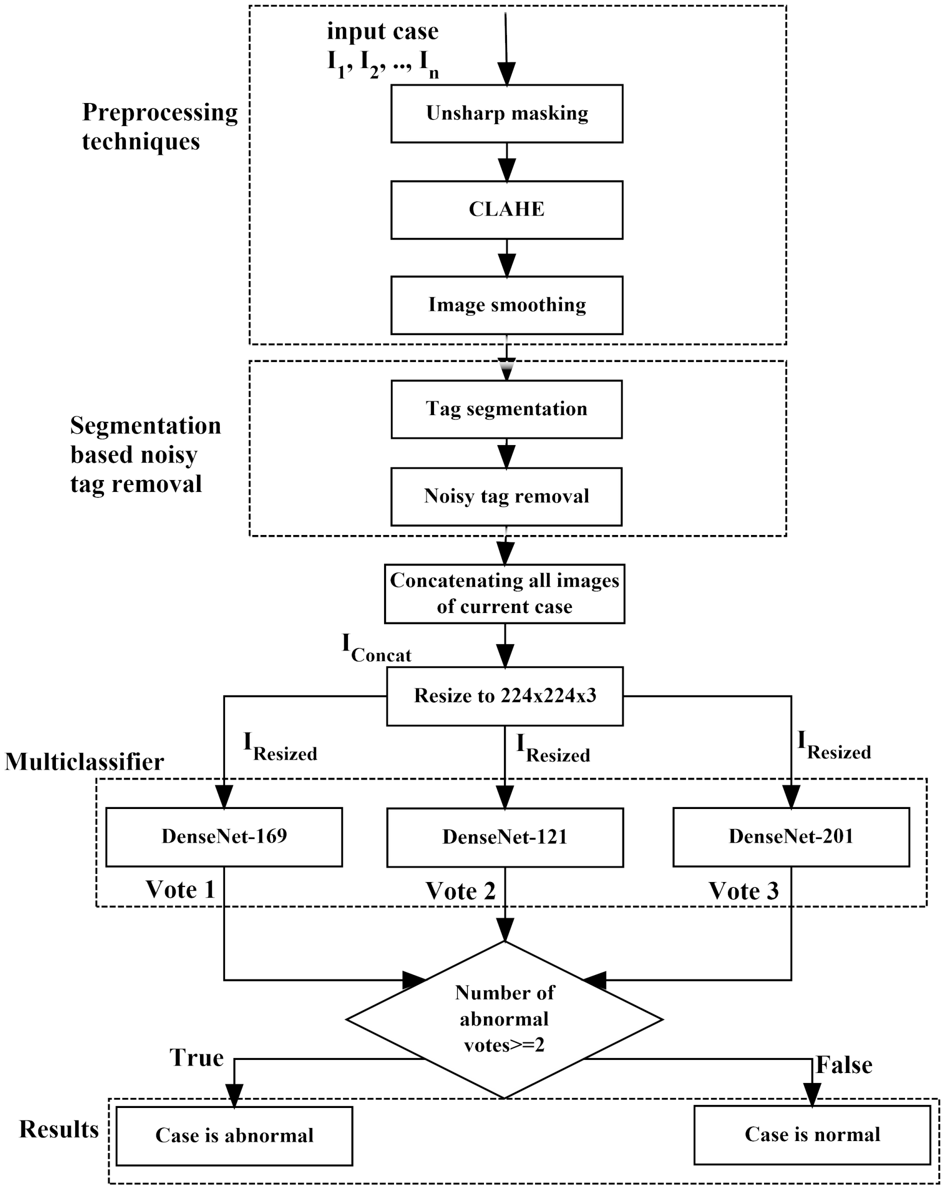

Fig. 2. Flowchart of the proposed ensemble learning-based abnormality diagnosis.

can be seen from Figure 4(B, D, F).

For the classifiers to benefit from radiographs of multiple orientations, all n radiographs are horizontally concatenated. The concatenated image is then resized to 224×224×3 as it is used as input size to the classifiers.

For the classifiers to benefit from radiographs of multiple orientations, all n radiographs are concatenated horizontally. The concatenated image is then resized to 224×224×3 as it is used as input size to the classifiers.

The final step is the voting-based classification. As the MURA dataset is labeled using the votes of three radiologists, a similar hard voting-based approach is used for the classification. A case is labeled as normal if two or more classifiers among DenseNet-169, DenseNet-201, and DenseNet-121 vote it as normal and vice versa.

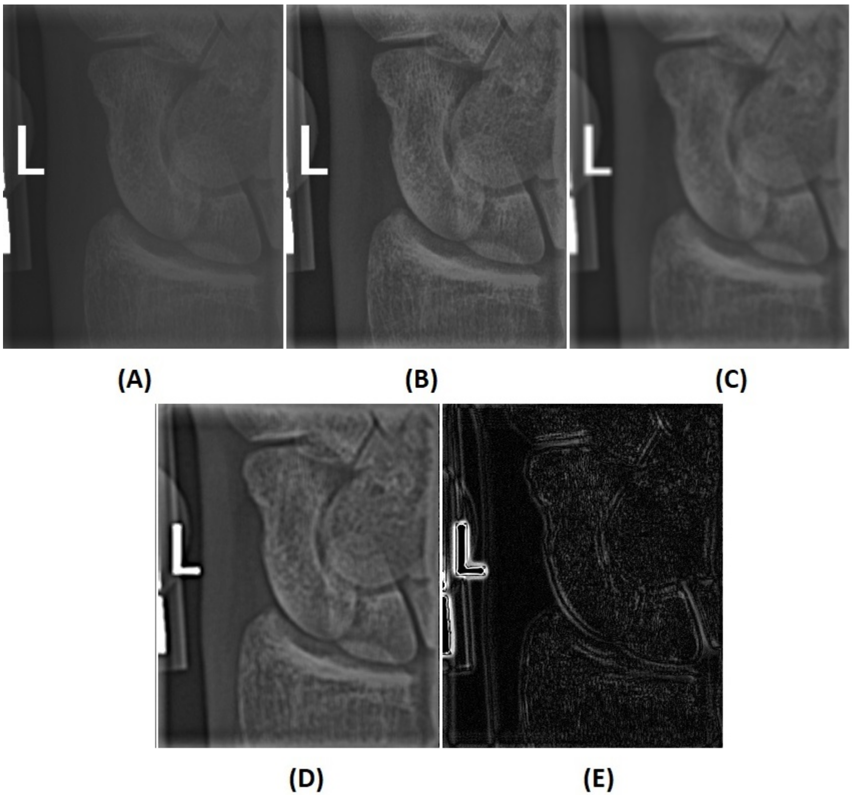

Fig. 3. Results of different steps of pre-processing: (A) input image, (B) CLAHE, (C) smoothing, (D) unsharp masking, (E) magnified absolute difference of (A) and (D).

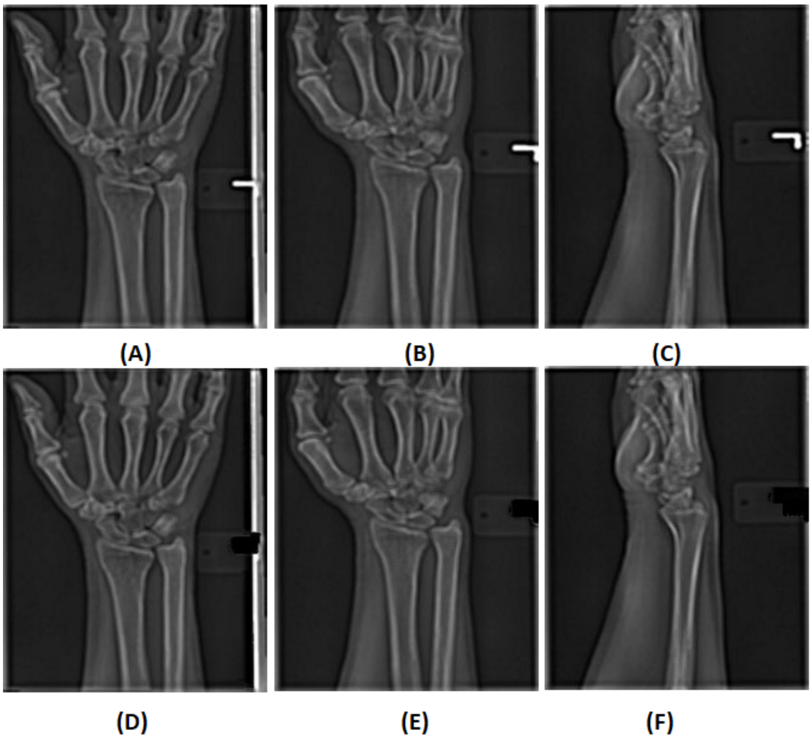

Fig. 4. Results of noisy tags removal using UNet: (A-C) input images, (D-F) images after tag removal.

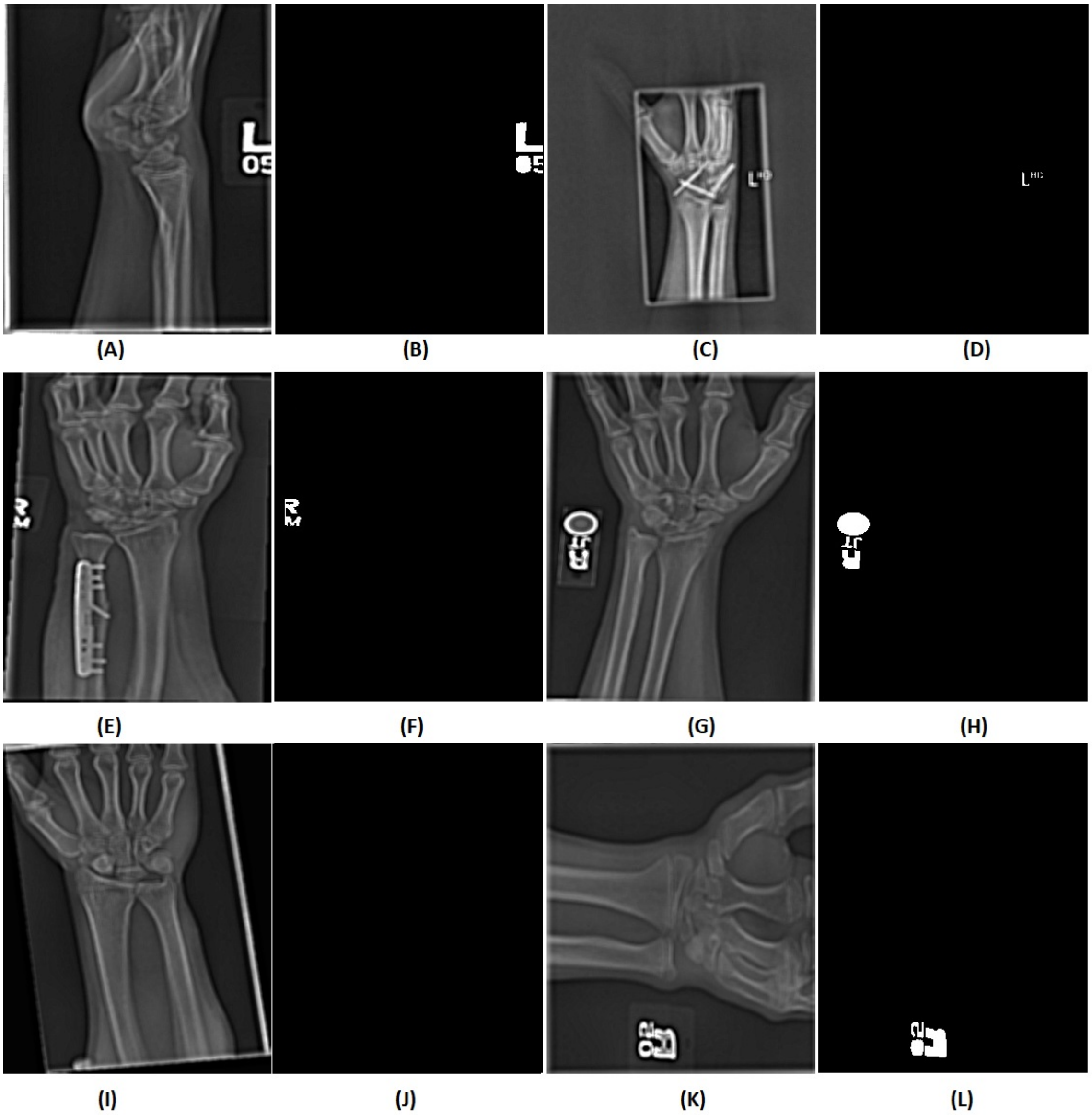

Fig. 5. Few examples of radiograph tags with different properties along with their generated grouth truths for segmentation: (A, C, E, G, I, K) intensity images, (B, D, F, H, J, L) corresponding annotated ground truths.

3.2 Training the networks

3.2.1 Training the UNet

This subsection explains the training of the UNet and three classifiers. It is observed that tags in the radiographic images act as noisy features that may result in false diagnoses. Those tags are in different sizes, shapes, and orientations and have different intensities, as can be seen from Figure 1(a–f). Automatic segmentation of all cases is not possible using conventional approaches such as Otsu's (Otsu, 1979), Maximally Stable Extremal Regions (MSER) (Saadawy *et al.*, 2020), or OCR. Therefore, an interactive Matlab script is generated to select the text region based on connectivity from the result of these three algorithms and stored as a ground truth image. A total of 800 ground truths are created using the trained UNet. A few examples of the annotated ground truths can be seen in Figure 5(B, D, F, H, J, L).

Before training the UNet, all of the intensity of the images that correspond to the annotated ground truths are enhanced using methods shown in Figure 2. To train the UNet, the ground truth images are augmented using elastic transform, grid distortion, optical distortion, motion blur, rotation with a jump of 30°, CLAHE, and vertical flip. The UNet is trained on 50 epochs with a learning rate of 0.0001 and a batch size of 12. The mean Intersection over Union (IoU) of the trained UNet obtained 0.6112.

3.2.2 Training the classifiers

All the radiographic images selected for the training of classifiers are first subjected to the application of UNet segmentation. Once the UNet segments the tags in the image, a rectangular region consisting of the segmented tag is replaced by 0 intensities as CNN ignores zero intensities when learning the features. After the noisy tag is removed, all n orientation images for the same case are concatenated, then resizing them to 224×224×3. Training the classifiers is done as follows:

1. Enhancing the image and edges using CLAHE, smoothing, and unsharp masking as discussed in 3.1.
2. Noisy tags removal using UNet.
3. Training the classifiers
    (a) Data normalization
    (b) Augmentation
    (c) Transfer learning

The first two steps are similar to as discussed in 3.1. However, as far as the third step is concerned, it comprises some parts represented by data normalization, augmentation, and transfer learning.

All the radiographic images are normalized for training. The horizontal flip and translation are done for data augmentation to balance the two classes. Rotation is avoided as CNN sometimes fails to perform well on rotated images (Saif *et al.*, 2019). In addition, rotation results in several images having similar properties and hence can produce overfitting.

After augmentation, transfer learning is applied to train the DenseNet-169, DenseNet-201, and DenseNet-121 classifiers. These classifiers are fully trained on the ImageNet dataset before transfer learning is applied. All the layers except for the last convolutional layer are frozen before training the classifiers for transfer learning. Optimizers such as Adam, SGD, Adadelta, and RMSprop are evaluated. The RMSprop is selected because it provides the best results for the testing dataset. A total of 500 test cases from the training are randomly isolated using the train test split method of the Sckit-learn library for Python. The LERA dataset would have been used for testing, however, it doesn't contain wrist cases. The batch size used is 32, whereas classifiers are trained on 200 epochs. Figure 6 shows the plot of training and validation accuracies of classifiers for the discussed parameters.

Furthermore, to choose between hard and soft voting for ensemble learning, the experiment was conducted on 500 training radiographic images. Figure 7 shows the receiving operating characteristics (ROC) plot for all three classifiers, whereas Figure 8 shows the confusion matrix comparison between the hard and soft voting. If the number of abnormal votes for hard voting is more than one, a case is considered abnormal. For all three classifiers for the soft voting, the probabilities were obtained from the Softmax layer. Afterward, the probabilities of both classes were added, and a case was diagnosed based on the maximum sum. Hard voting-based ensemble learning performed better in diagnosing the true negatives as seen in Figure 8. However, there is no significant difference among the true positives.

4. Experiments, Results and Discussions

This section discusses some of the performed experiments to finalize the methodology and results. Therefore, this section is divided into two subsections which are the "Experiments" and "Results and Discussions".

4.1 Experiments

All three classifiers are evaluated using the segmented and non-segmented datasets to verify the significance of the noisy tag removal. The segmented dataset is pre-processed for enhancement and

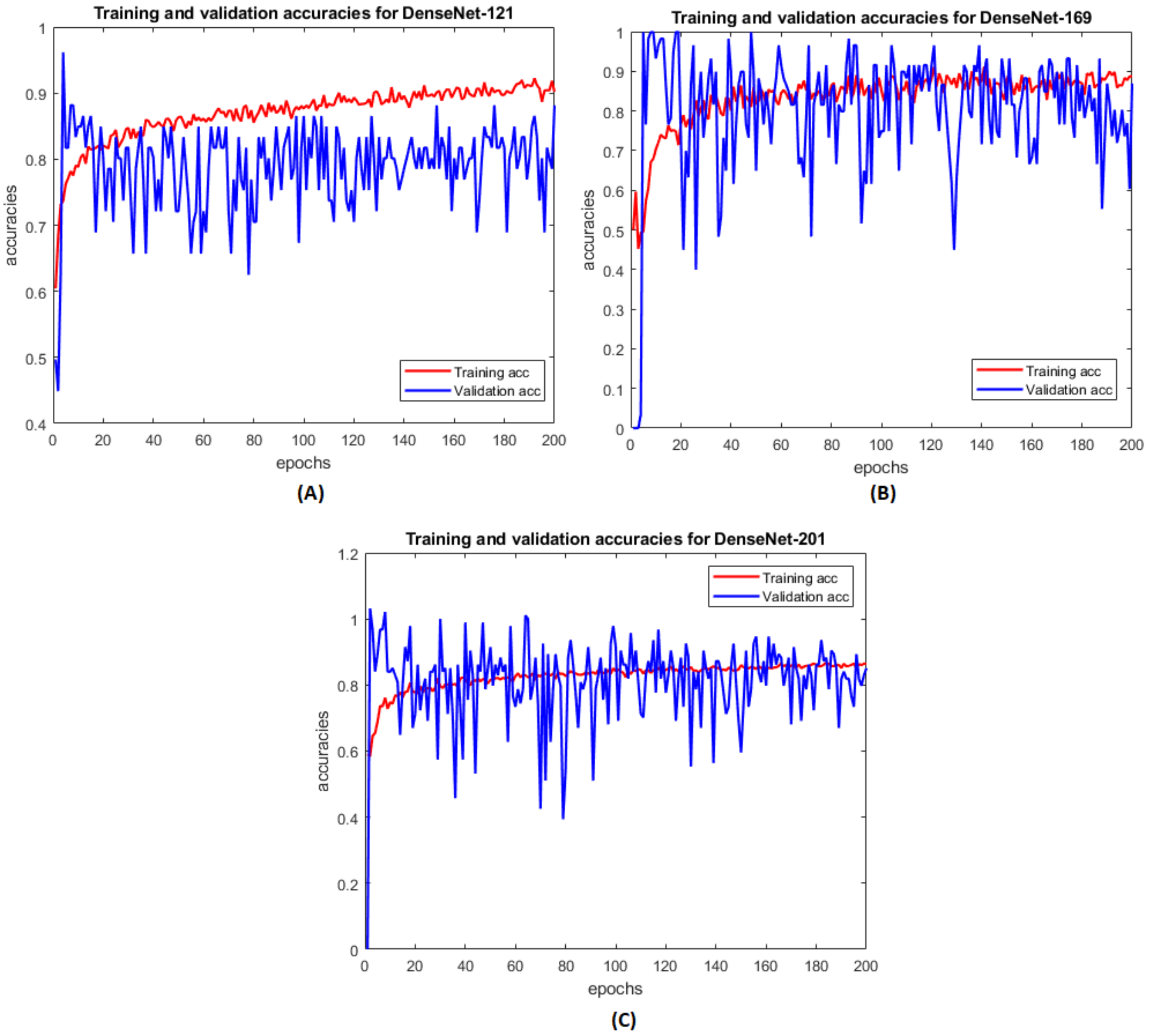

Fig. 6. Training and validation accuracies: (A) DenseNet-121, (B) DenseNet-169, (C) DenseNet-201.

removes the noisy tags, as shown in Figure 4(D–F), whereas the non-segmented dataset is only enhanced using the same pre-processed techniques.

Tables 1 and 2 show the results of the classifier training on the non-segmented and segmented datasets, respectively. The segmented dataset improved accuracies by 4%–5%, 6%–7%, and 9%–11% for the training, validation, and testing, respectively.

Before finalizing the DenseNet-121, DenseNet-201, and DenseNet-169 classifiers for the ensemble learning, experiments were performed on various networks to see which network is best fitted for the segmented dataset. Figure 9 shows the comparison of training accuracy for DenseNet-201, inceptionV3, and vgg19 as used by Mondol *et al.* (Mondol *et al.*, 2019), and MobileNet as used by Saadaway *et al.* (Saadawy *et al.*, 2020). Comparison after 80 epochs is shown to make the difference clear by reducing the vertical scale. The DenseNet-201 classifier is selected because it provides the lowest training accuracy compared to DenseNet-121 and DenseNet-169. The figure clearly shows that the DenseNet-201 provides better accuracy than other classifiers.

4.2 Results and Discussions

Table 3 compares the performance of the proposed methodology among the three voting classifiers used. The significant improvement in the performance matrices for the voting-based model compared to

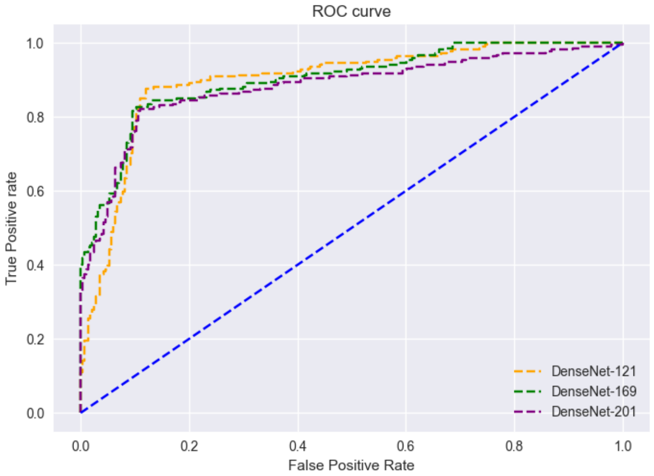

Fig. 7. ROC plot for DenseNet-121, DenseNet-169 and DenseNet-201.

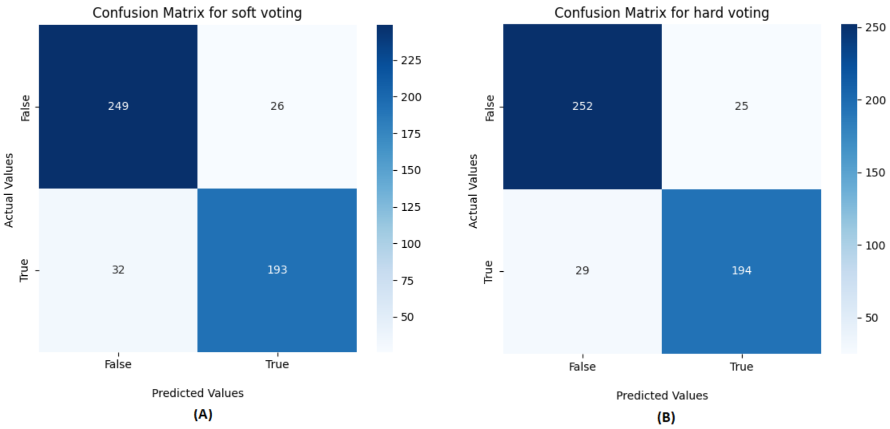

Fig. 8. Confusion matrix for: (A) soft voting, (B) hard voting.

all three classifiers used proves the ability of ensemble learning.

Table 3 also discusses a comparison of the proposed methodologies with state-of-the-art methods proposed by Saadawy *et al.* (Saadawy *et al.*, 2020), Mondol *et al.* (Mondol *et al.*, 2019), and Goyal *et al.* (Goyal *et al.*, 2020). Saadawy provided two types of accuracies when it comes to wrist abnormality classification. Accuracy is obtained when an extra classifier takes a radiograph of any part of the skeleton and classifies it as either shoulder, elbow, forearm, wrist, hand, etc., and is used as the first stage. The accuracy, in that case, is 75.61%. However, it is not appropriate to compare with it the accuracy of the proposed methodology because the accuracy of the abnormality detection depends on the accuracy of the first stage classifier. Fortunately, the accuracy of wrist abnormality detection was provided by Saadawy without using the first stage classifier. Therefore, Table 3 has the accuracy of the classifier proposed

Table 1. Evaluation of voting classifiers on the non-segmented dataset.

Model	**Training Accuracy**	**Validation Accuracy**	**Testing Accuracy**	**Training loss**	**Validation loss**	**Testing loss**
DenseNet-121	0.85	0.82	0.75	0.32	0.44	0.50
DenseNet-169	0.85	0.81	0.76	0.33	0.45	0.54
DenseNet-201	0.81	0.78	0.73	0.44	0.47	0.56

Table 2. Evaluation of voting classifiers on the segmented dataset.

Model	Training Accuracy	Validation Accuracy	Testing Accuracy	Training loss	Validation loss	Testing loss
DenseNet-121	0.9	0.88	0.86	0.23	0.35	0.36
DenseNet-169	0.89	0.87	0.85	0.31	0.37	0.38
DenseNet-201	0.86	0.85	0.84	0.34	0.4	0.42

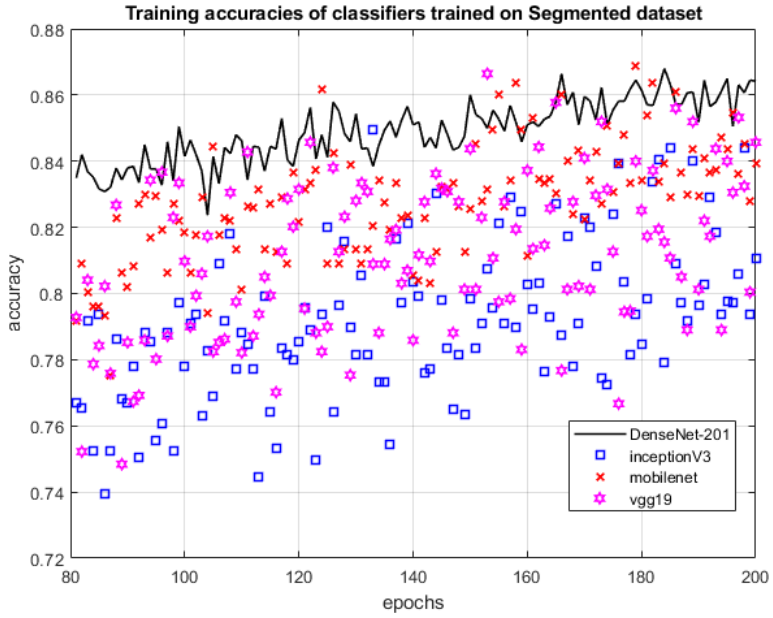

Fig. 9. Training accuracies for multiple classifiers.

by Saadawy that takes wrist radiographic image as input and classifies it as either normal or abnormal. Therefore, given that Saadawy did not provide results for other performance matrices such as specificity, sensitivity, precision, and F1 score, they are left blank. As seen from the table, Goyal *et al.* (Goyal *et al.*, 2020) provided accuracy, sensitivity, and specificity in abnormality detection. Table 3 shows that the proposed method has better accuracy in detecting abnormality than both Saadawy and Goyal's. Goyal performed better in terms of specificity. However, the proposed method has better sensitivity. Mondol proposed an ensemble learning-based method and provided testing accuracy and F1 score. For accuracy, the proposed method is 1.35% better than the Mondal method, whereas the F1 score is slightly better. Overall, the proposed method performed nearly 1.5%–4.5% better in terms of accuracy, 3.4%–6% in specificity, 1.9%–3.4% in sensitivity, 2.8%–5% in precision, and 2.71%–3.38% in F1 score.

Table 3. Comparison of proposed voting-based classification method with other classification methods.

Model	Accuracy	Specificity	Sensitivity	precision	F1 Score
DensNet-169	0.8540	0.8348	0.8696	0.8664	0.8680
DensNet-201	0.8460	0.8251	0.8628	0.8597	0.8631
DensNet-121	0.8660	0.8514	0.8777	0.8809	0.8613
Saadawy	0.8194	-	-	-	-
Goyal	0.85	0.964	0.793	-	-
Mondol	0.8786	-	-	-	0.8933
Voting base model	**0.892**	**0.8858**	**0.8968**	**0.9097**	**0.8951**

5. Conclusion and Future Work

This paper proposes an efficient classifier voting-based method for wrist abnormality detection. The results show that the proposed method introduces segmentation for noisy tags removal thus improving accuracy. For the segmentation, ground truth is created with the help of various segmentation methods and connectivity information. Edge-based image enhancement, segmentation, tags removal, and voting-based classification are good combinations thus improving matrices' performance. This paper discusses only the wrist abnormality detection; however, similar abnormality detection for other skeletal regions may be fruitful.

In the future, segmentation that obtains other skeletal regions will be applied by creating another ground-truth dataset. In addition, instead of working on a specific skeletal region, a two-tier architecture will be proposed to classify the radiographs according to the skeletal region, followed by the abnormality detection. To further improve the performance, an input size of 512×512 will be used for classification as concatenated 224×224 may not contain enough details. Moreover, instead of colored images, the grayscale images will be processed to reduce the computational requirement.

References

Aal, M. M. A., Awwad, S. H., Ahmed, F. H., Wasfi, A. G., Ghanim, T. M., & Nabil, A. M. (2018). Survey: Automatic recognition of musculoskeletal disorders from radiographs. In *13th IEEE International Conference on Computer Engineering and Systems (ICCES)* (pp. 56-62).

Bhargavan, M., Kaye, A. H., Forman, H. P., & Sunshine, J. H. (2009). Workload of radiologists in United States in 2006–2007 and trends since 1991–1992. *Radiology*, **252**(2), 458-467.

Chada, G. (2019). Machine learning models for abnormality detection in musculoskeletal radiographs. *Reports*, **2**(4), 26.

Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R., & Girod, B. (2011). Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *2011 18th IEEE International Conference on Image Processing* (pp. 2609-2612).

El-Saadawy, H., Tantawi, M., Shedeed, H. A., & Tolba, M. F. (2020). A two-stage method for bone x-rays abnormality detection using MobileNet network. In *The International Conference on Artificial Intelligence and Computer Vision* (pp. 372-380).

Goyal, M., Malik, R., Kumar, D., Rathore, S., & Arora, R. (2020). Musculoskeletal abnormality detection in medical imaging using GnCNNr (group normalized convolutional neural networks with regularization). *SN Computer Science*, **1**(6), 458-467.

Irmakci, I., Anwar, S. M., Torigian, D. A., & Bagci, U. (2019). Deep learning for musculoskeletal image analysis. In *53rd IEEE Asilomar Conference on Signals, Systems, and Computers* (pp. 1481-1485).

Korot, E., Guan, Z., Ferraz, D., Wagner, S. K., Zhang, G., Liu, X., & Keane, P. A. (2021). Code-free deep learning for multi-modality medical image classification. *Nature Machine Intelligence*, **3**(4), 288-298.

LERA - lower extremity radiographs. [Online] Available: `https://aimi.stanford.edu/lera-lower-extremity-radiographs-2`.

Lindsey, R.,Daluiski, A., Chopra, S.,Lachapelle, A.,Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., & Potter, H. (2018). Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, **115**(49), 11591-11596.

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mondol, T. C., Iqbal, H., & **Hashem, M. (2019)**. Deep CNN-Based Ensemble CADx Model for Musculoskeletal Abnormality Detection from Radiographs. In *International Conference on Advances in Electrical Engineering* (pp. 392-397). Dhaka.

Nawarathne, T., Withanage, T., Gunarathne, S., Delay, U., Somathilake, E., Senanayake, J., & **Wijayakulasooriya, J. (2022)**. Comprehensive Study on Denoising of Medical Images Utilizing Neural Network-Based Autoencoder. In *Advanced Computational Paradigms and Hybrid Intelligent Computing* (pp. 159-170). Singapore.

Nazim, S., Hussain, S. S., Moinuddin, M., Zubair, M., & **Ahmad, J. (2021)**. A neoteric ensemble deep learning network for musculoskeletal disorder classification. *Neural Network World*, **37**(6), 377-393.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, **9**(1), 62-66.

Pal, S. K., Pramanik, A., Maiti, J., & **Mitra, P. (2021)**. Deep learning in multi-object detection and tracking: state of the art. *Applied Intelligence*, **51**(9), 6400-6429.

Prajna, Y., & **Nath, M. K. (2021)**. A Survey of Semantic Segmentation on Biomedical Images Using Deep Learning. In *Advances in VLSI, Communication, and Signal Processing* (pp. 347-357). Singapore.

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., ... & **Ng, A. (2017)**. MURA dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. *arXiv:1712.06957*.

Safiri, S., Kolahi, A. A., Cross, M., Hill, C., Smith, E., Carson-Chahhoud, K., & **Buchbinder, R. (2021)**. Prevalence, Deaths, and Disability-Adjusted Life Years Due to Musculoskeletal Disorders for 195 Countries and Territories 1990–2017. *Arthritis & Rheumatology*, **73**(4), 702-714.

Sagi, O., Guan, Z., & **Rokach, L. (2018)**. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **8**(4), e1249-1-e1249-18.

Saif, A. F. M., Shahnaz, C., Zhu, W. P., & **Ahmad, M. O. (2019)**. Abnormality detection in musculoskeletal radiographs using capsule network. *IEEE Access*, **7**, 81494-81503.

Solovyova, A., & **Solovyov, I. (2020)**. X-Ray bone abnormalities detection using MURA dataset. *arXiv:2008.03356*.

Sungheetha, A., & **Sharma, R. (2021)**. Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. *Journal of Trends in Computer Science and Smart technology (TCSST)*, **3**(2), 81-94.

Tantawi, M., Thabet, R., Sayed, A. M., & **El-emam, O. (2020)**. Bone X-rays classification and abnormality detection. In *Internet of Things—Applications and Future* (pp. 277-286). Springer, Singapore.

Tasci, E., Uluturk, C., & **Ugur, A. (2021)**. A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. *Neural Computing and Applications*, **33**(22), 15541-15555.

Varma, M., Lu, M., Dunnmon, J., Khandwala, N., Rajpurkar, P., Long, J., Beaulieu, C., Shpanskaya, K., Fei-Fei1, L., Lungren, M. P., & **Patel, B. N. (2019)**. Automated abnormality detection in lower extremity radiographs using deep learning. *Nature Machine Intelligence*, **1**(12), 578-583.

Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., & **Yang, R. (2021)**. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yadav, S. S., & **Jadhav, S. M. (2019)**. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, **6**(1), 1-18.

Yang, S., Lu, J., Zeng, J., Wang, L., & **Li, Y. (2019)**. Prevalence and risk factors of work-related musculoskeletal disorders among intensive care unit nurses in China. *Workplace health* & *safety*, **67**(6), 275-287.

Zhang, M., Li, H., Pan, S., Lyu, J., Ling, S., & **Su, S. (2021)**. Convolutional neural networks based lung nodule classification: a surrogate-assisted evolutionary algorithm for hyperparameter optimization. *IEEE Transactions on Evolutionary Computation*.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., & **Summers, R. M. (2021)**. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*.

Submitted:	15/03/2022
Revised:	24/05/2022
Accepted:	02/06/2022
DOI:	10.48129/kjs.splml.19477