BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
DOI: 10.24425/bpasts.2025.153233

A New Model for Anomaly Detection in Elbow and Finger X-Ray Images: Proposed Parallel DenseNet

Selahattin GÜÇLÜ[1,*], Durmuş ÖZDEMİR[2], Hamdi Melih SARAOĞLU[3]

[1] Department of Electrical and Electronics Engineering, Kütahya Dumlupınar University, Kütahya, Türkiye
[2] Department of Computer Engineering, Kütahya Dumlupınar University, Kütahya, Türkiye
[3] Department of Electrical and Electronics Engineering, Kütahya Dumlupınar University, Kütahya, Türkiye

Abstract. Image recognition is one of the essential branches of computer vision and has significant theoretical and practical importance. This study aims to enhance a deep learning model, DenseNet, by incorporating parallel structures using X-ray images from the MURA (Musculoskeletal Radiographs) dataset. X-ray images of the elbow and finger are analyzed using AlexNet, DenseNet, Parallel DenseNet, and Proposed Parallel DenseNet (PPDN) deep learning models for anomaly detection, and the results are compared. For the elbow, 1534 healthy and 1630 anomaly X-ray images; for the finger, 1965 healthy and 1938 anomaly X-ray images were used to train the deep learning models. As a result of the statistical analysis, the most successful model with the test accuracy value for the elbow part was the suggested PPDN model (78.74%). The next successful model for the elbow part was AlexNet (77.05%). The most successful model for the finger part was again the PPDN model (69.97%), and the next successful model was the Parallel DenseNet model for the finger part (68.94%). In anomaly detection of musculoskeletal elbow and finger X-ray images, the PPDN model is more successful than the classical DenseNet and Alexnet models in terms of test accuracy.

Key words: DenseNet; convolution; classification; musculoskeletal radiographs

1. INTRODUCTION

Musculoskeletal disorders are injuries or pain that occur in the human musculoskeletal system, including ligaments, joints, nerves, tendons, muscles, and structures supporting the neck, limbs, and back. Patients suffer from chronic pain and various limitations in mobility, dexterity, and functional abilities [1]. Musculoskeletal radiographic images are an essential tool in the diagnosis of anomalies. Usually, when a patient has an accident or a fracture is suspected, the patient goes to the emergency room, where his or her doctor first performs a fracture examination, and radiographs are taken to detect fractures. The misclassification rate of X-ray images in the emergency department is due to the emergency room doctor classifying the X-ray images as needing to be an experienced radiologist and rapidly taking images, causing errors. Therefore, various anomalies, including fractures, hardware, degenerative joint disease, lesions, and subluxations, may be missed depending on the doctor's experience [1,2]. An automatic classifier to help doctors classify X-ray images can significantly reduce the error rate [3]. Deep learning has critical importance in categorizing medical images.

Deep learning algorithms make life easier for radiologists and orthopedic surgeons by providing faster and more accurate real-time findings [4]. Therefore, deep learning has recently become one of the most potent and impressive learning models for image pattern recognition and classification problems [5]. Two points are important in deep learning. The first is that the data should have an extensive collection labeled; the second is to find the appropriate deep-learning approach to interpret the data accurately [6].

1.1. Related works

Examining studies aimed at musculoskeletal diagnosis revealed that deep learning techniques and patient X-ray image datasets were usually applied. Liang ve Gu in studies propose a novel multi-network architecture consisting of a multi-scale convolution neural network (MSCNN) with a fully connected graph convolution network (GCN), named MSCNN-GCN, for the detection of musculoskeletal abnormalities via musculoskeletal radiographs. The model's effectiveness was validated using the MURA dataset, comparing it to radiologists' performance and three popular

CNN models (DenseNet169, CapsNet, and MSCNN) [7]. Harini et al. compared the training results with deep learning models; Inception V3, Xception, VGG-19, DenseNet169, and MobileNet using the MURA dataset (hand, wrist, and shoulder) and showed that the performance of the VGG-19 model was the lowest [8]. Cheng et al. used the masking method they proposed on the input images obtained from the MURA data set (hand, finger, wrist, forearm, elbow, humerus, and shoulder) as input data. They compared them with the DenseNet model and achieved more successful results in their proposed model [9]. Lysdahlgaard, using elbow and wrist X-ray images in the MURA dataset, obtained analysis results with derivatives of VGG, ResNet, DenseNet, Xception, and Inception models. Successful results for the elbow part were obtained with test accuracy values between 64% and 73% with the DenseNet model. With test accuracy scores of 84% using the VGG model, the wrist portion produced successful results [10]. Solovoya and Solovyov analyzed the kappa statistic results obtained from training the DenseNet169 model on the entire MURA dataset. The highest kappa value, 0.942, was achieved for the wrist component, while the lowest, 0.395, was observed for the finger component [11]. Kandel et al. used VGG, ResNet, DenseNet, Xception, and Inception deep learning models to analyze the X-ray images in the MURA dataset using various statistical methods and a different algorithm, and the results were compared [12]. Mondol et al. designed a model combining VGG-19 architecture and ResNet-50 architecture to detect musculoskeletal anomalies in the MURA dataset. The model, which they call (CADx), was trained on four parts of the MURA dataset: elbow, finger, humerus, and wrist. They stated that the CADx model performed relatively better than the classical VGG-19 and ResNet-50 architecture [13]. Morra et al. considered a multi-stage transfer learning approach for medical image analysis. They combined color information extraction with transfer learning and used different classification models such as ResNet and DenseNet. They achieved successful results in the classification of medical images using deep-learning models with color features [14].

Studies in the literature have shown that the DenseNet deep learning model has achieved successful results in musculoskeletal disorders. Although deep learning methods utilizing parallel layers have been employed in the literature, performance analysis regarding the increase in the number of parallel layers and layer count in datasets with a large number of X-ray images, such as the MURA dataset, still needs to be improved. In this study, it was predicted from the literature that the parallel and multi-layered architecture would increase the accuracy performance and was tested by applying it to the DenseNet architecture. This study analyzed the performance of the test accuracy values of the layers of the DenseNet deep learning model in parallel connection and compared them with AlexNet. This study proposes to apply the PPDN model in a way that makes it possible to detect musculoskeletal system anomalies in MURA dataset X-ray images compared to the classical DenseNet model.

This study is structured as follows: Chapter 2 covers deep learning models and related technical procedures. Chapter 3 presents the proposed models along with their processes. Performance metrics are detailed in Chapter 4, and Chapter 5 discusses the experimental analysis and results. Finally, Chapter 6 provides conclusions, discussion points, and recommendations.

2. METHODOLOGY

2.1. Deep learning

Deep learning is a subset of the field of machine learning that deals with creating deep artificial neural networks inspired by biological neural networks in the human brain [15]. Deep learning has become crucial in healthcare, significantly enhancing diagnostic accuracy, personalized treatment, and predictive analytics. Deep learning models can assist healthcare professionals in early disease detection, reducing errors, and optimizing patient outcomes by analyzing complex medical data such as medical imaging, genomics, and patient records. For this reason, deep neural networks outperform shallow machine learning algorithms in most applications where text, image, video, speech, and audio data need to be processed [16]. We can classify deep learning architectures as Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), Autoencoders (AE), Hybrid Architectures (HA), and Deep Belief Networks (DBN) [17].

2.2. Models

Convolutional Neural Network (CNN) stands at the forefront of representative algorithms for image recognition through a neural network [18]. Recent studies have shown that convolutional neural networks have become deeper and deeper in order to obtain more accurate results [19]. This study used popular DenseNet and AlexNet deep learning methods to train CNN with a dataset.

2.2.1. AlexNet model

AlexNet deep learning architecture is the first convolutional neural network to participate in the ImageNet competition held in 2012. It outperformed all previous low-depth algorithms with an accuracy rate of 84,6% in image classification. Since then, CNNs have become the most advanced algorithm in image classification [19]. AlexNet architecture: It has 650,000 neurons, 60,000,000 parameters, five convolution layers, and three dense layers. Two innovations made in AlexNet were using the ReLU activation function instead of the sigmoid activation function and the dropout method to overcome the overfitting problem that this deep architecture can cause. The main advantage of this network is that the training process is computationally efficient compared to other networks. On the other hand, the AlexNet deep learning method needs to be deeper to capture complex features from images [18,19]. Fig. 1 shows the architecture of the AlexNet deep learning model.

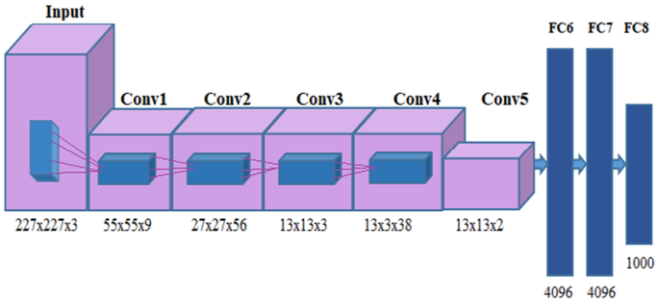


Fig.1. The architecture of the AlexNet deep learning model [20].

2.2.2. DenseNet model

Dense blocks were first proposed by Gao Huang et al. (2016). The model refers to densely connected convolutional networks. DenseNet is inspired by ResNet, but the authors propose using dense blocks instead of residual links [17,21]. The DenseNet model is a novel CNN designed for image classification. It is crafted to operate through dense blocks, utilizing densely connected layers, enabling intensive processing. Dense blocks facilitate information sharing by establishing dense connections between layers. These connections ensure enhanced information flow, utilizing the densely interconnected layers in the model [22]. The input of the DenseNet model consists of an RGB image with dimensions defined as 1 (batch size), 3 (channels), 224 (height), and 224 (width) [23,24]. This entry goes through a pile of interconnected features; this stack consists of combined attributes by combining the output of all previous layers with further layers. This form of connection is the main idea of DenseNet models. For example, the input of a layer X3 = H3 ([X0, X1, X2]) consists of the outputs of previous layers, such as X2, X1, X0, and the original input. These inputs are combined to create a single deep feature map with the exact spatial resolution but a different number of filters. Continuously connecting successive dense blocks will eventually lead to profound entrances. The architecture is divided into dense blocks using all consecutive layers in each block. This performs a shrinking process to reduce the depth of the feature map while using one-by-one convolution in successive layers to preserve spatial resolution. After this process, max pooling is used to reduce the feature map size [23]. Different types of DenseNet exist, including DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264. The numbers next to DenseNet types refer to the number of layers; for example, DenseNet-264 has 264 layers [25]. DenseNets require fewer parameters than traditional CNNs because there are no redundant feature maps. If we analyze the structure of DenseNets, the feature map sizes remain constant in blocks with different filters. This feature helps optimize the number of parameters while increasing the learning ability of the network [26]. Since architectures differ according to DenseNet types, these differences are shown in Table 1. Fig. 2 shows the architecture of the Densenet model.

TABLE 1. DenseNet architectures. Each "conv" layer shown in the table corresponds to the sequence BN (Batch Normalization)-ReLU-Conv (Convolution), respectively [27]

Layers	Output Size	DenseNet-121		DenseNet-169		DenseNet-201		DenseNet-264	
Convolution	112x112	7x7 conv, stride 2							
Pooling	56x56	3x3 max pool, stride 2							
Dense Block-1	56x56	1x1 conv 3x3 conv	x6	1x1 conv 3x3 conv	x6	1x1 conv 3x3 conv	x6	1x1 conv 3x3 conv	x6
Transition Layer-1	56x56	1x1 conv							
	28x28	2x2 average pool, stride 2							
Dense Block-2	28x28	1x1 conv 3x3 conv	x12	1x1 conv 3x3 conv	x12	1x1 conv 3x3 conv	x12	1x1 conv 3x3 conv	x12
Transition Layer-2	28x28	1x1 conv							
	14x14	2x2 average pool, stride 2							
Dense Block-3	14x14	1x1 conv 3x3 conv	x24	1x1 conv 3x3 conv	x32	1x1 conv 3x3 conv	x48	1x1 conv 3x3 conv	x64
Transition Layer-3	14x14	1x1 conv							
	7x7	2x2 average pool, stride 2							
Dense Block-4	7x7	1x1 conv 3x3 conv	x16	1x1 conv 3x3 conv	x32	1x1 conv 3x3 conv	x32	1x1 conv 3x3 conv	x48
Classification Layer	1x1	7x7 global average pool							
		1000D fully-connected, softmax							

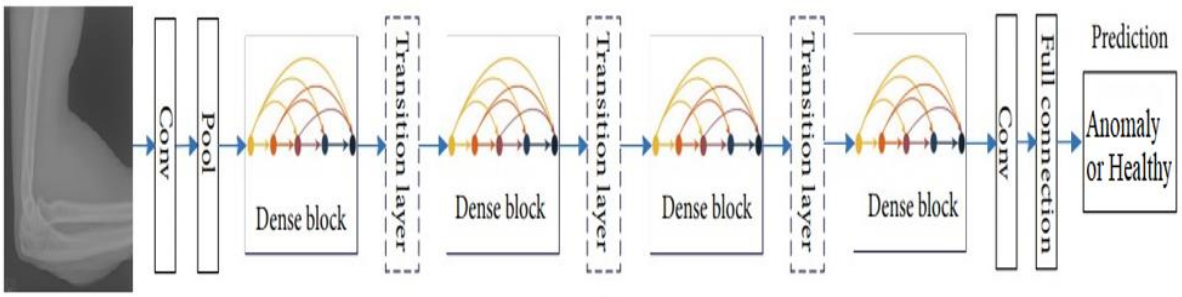


Fig.2. The architecture of the DenseNet deep learning model [28].

Fig. 3 shows the DenseNet block expansion architecture.

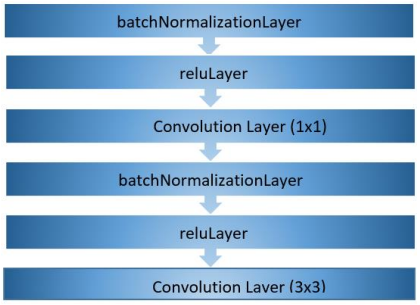


Fig.3. DenseNet block expansion [25].

Fig. 4 shows the DenseNet Transition layer expansion architecture.

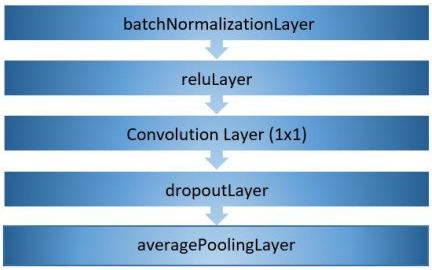


Fig.4. DenseNet Transition layer expansion [25].

The MURA dataset X-ray images were analyzed using DenseNet-264, a classical DenseNet model.

3. PROPOSED MODELS AND PROCESSES

The task of the Dense Block module, which is the main module of the DenseNet model, is to extract features from images. Nevertheless, this module has room for optimization as the arrangement of various functional layers may need to be revised. Using different parallel structures in the CrodenseNet architecture achieved better results in diagnosing COVID-19 disease [29]. Yin et al. obtained promising results using parallel layers on the classical DenseNet model using the CIFAR 10 and CIFAR 100 datasets, so in this study, the DenseNet deep learning method was developed by adding parallel blocks by deleting or adding some layers and adjusting the convolution layer. In contrast to Yin et al., who used three blocks by reducing the number of classic DenseNet blocks, four blocks were utilized in this study. When additional convolutional layers are superimposed on top of each other, with a larger receptive field in terms of size, it can lead to the extraction of richer features and higher computational efficiency. In contrast to Yin et al., in the developed DenseNet deep learning models, A 3x3 convolution layer was used instead of 1x1 to capture and extract features in larger areas. The 3x3 convolution layer processes each pixel in a 3x3 window around itself and its neighbors. Additionally, DenseNet processes input features through the dense block module and employs only a single convolutional kernel for feature extraction. This inevitably results in a relatively uniform structure, making it susceptible to potential loss of information in the image [18]. In order to take full advantage of the existing features and not add too many parameters, a new dilated convolution block based on the dilated convolution method is designed in parallel. Then, multipath Dense blocks are connected to combine various feature maps from different channels. This helps model the feature compatibility of channels and perform powerful feature extraction. In this study, to develop the DenseNet model, a deep learning model using X-ray images from the MURA dataset, a "Parallel DenseNet" model was first developed by adding Dense blocks parallel to the classical DenseNet architecture. By parallelizing the DenseNet architecture, better results are obtained in classification [18,29]. By optimizing this developed Parallel DenseNet architecture, the "Proposed Parallel DenseNet" model was developed. The DenseNet models used in the study are used for image recognition after training. Moreover, the accuracy of this image recognition is obtained by comparing the results with the labels of the test images.

3.1. Parallel DenseNet model

In the parallel DenseNet model, the layers and repetition numbers of the classical DenseNet-264 model were used precisely. The transition layer is the same as shown in Fig. 4. The architecture of the Parallel DenseNet deep learning model is shown in Fig. 5. The dense block expansion is shown in Figure 3.

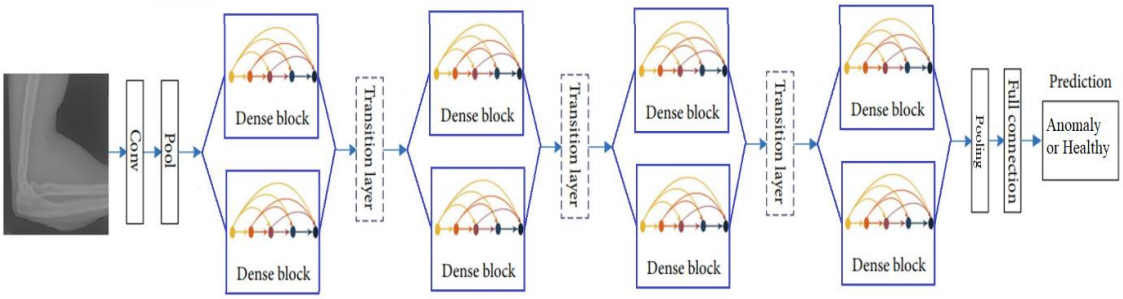


Fig.5. The architecture of the Parallel DenseNet deep learning model.

As seen in Fig. 5, based on the classical DenseNet architecture, Dense blocks with the same layers and features, connected in parallel to the classical Dense blocks, have been added. The number of classic DenseNet blocks is four, as shown in Figure 2. These block numbers can be increased and decreased [18]. In the analyses performed with three blocks, the number of blocks was chosen as four since there was a performance decrease in the test accuracy value at an accuracy value of 10%. Parallel blocks were connected with Transition layers, which have the same properties as classical DenseNet. As a result of the parallel connections, the feature extraction process is performed after the inherent structure of Dense blocks concatenates the feature maps. The 'Pooling' and 'Full Connection' layers are connected in the last layer, and the 'Anomaly- Healthy' classification is made.

3.2. Proposed parallel DenseNet (PPDN) model

In the PPDN model, the layers and repetition numbers of the classical DenseNet-264 model were used precisely. The transition layer is the same as shown in Fig. 4. In the proposed model, the architecture shown in Fig. 6 was created by connecting discrete Density blocks in parallel to the developed Density blocks. Classic DenseNet architecture consists of successively added layers. Parallel blocks were connected with Transition layers, which have the same properties as classical DenseNet. In the PPDN architecture, as seen in Fig. 6, Dense blocks combine feature maps from different channels to help the feature extraction process.

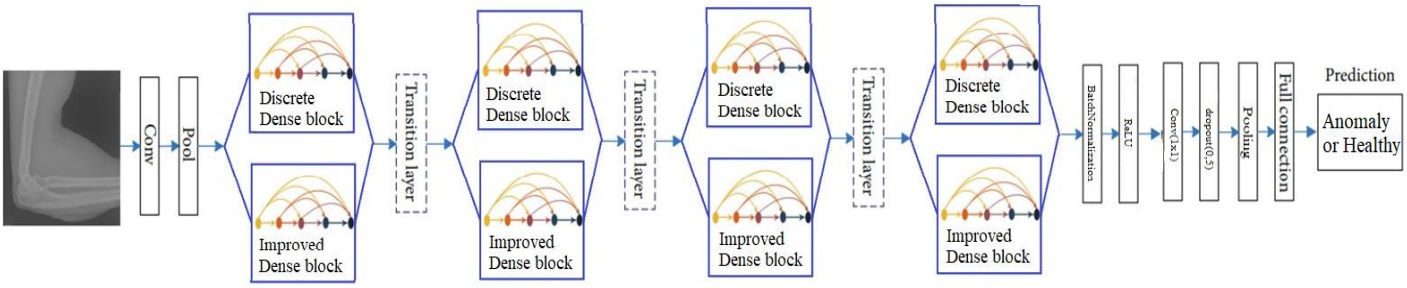


Fig.6. The architecture of the Proposed Parallel DenseNet deep learning model.

The Improved DenseNet block expansion is shown in Fig. 7.

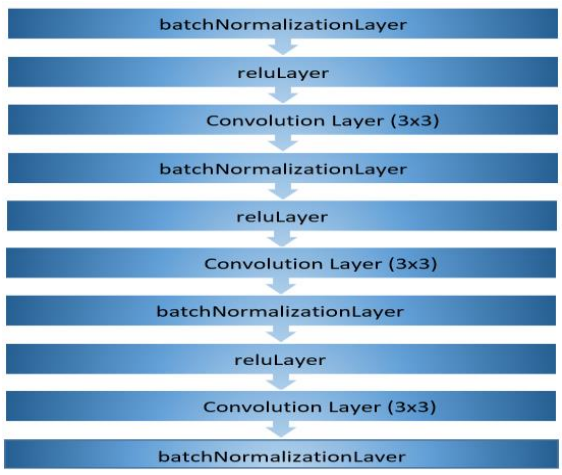


Fig.7. Improved DenseNet block expansion.

As shown in Fig. 7, in the Improved DenseNet block expansion, unlike the classical DenseNet block (Fig. 3), convolution layers are used as 3x3 instead of 1x1 so as not to lose the information in the image. Unlike the classic Dense block, Batch Normalization-ReLU-Conv (Convolution) layers have been added respectively. The Discrete DenseNet block expansion is shown in Fig. 8.

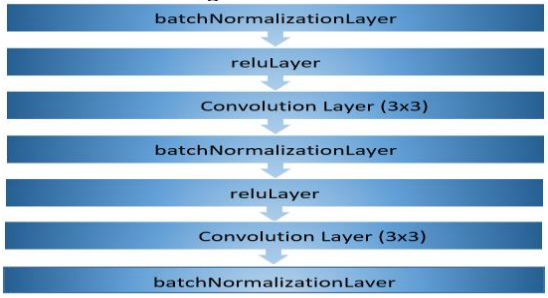


Fig.8. Discrete DenseNet block expansion.

In the PPDN architecture, the parallel connected Discrete DenseNet block expansion is shown in Fig. 8; unlike the classical Dense block (Fig. 3), convolution layers are used as 3x3 instead of 1x1 so as not to lose information in the image.
Unlike the classic Dense block, a Batch Normalization layer has been added for the normalization process. Before the classification layer, unlike the classical DenseNet architecture, the 'Anomaly-Healthy' classification is made using Batch Normalization-ReLU-Conv (Convolution) and dropout layers.

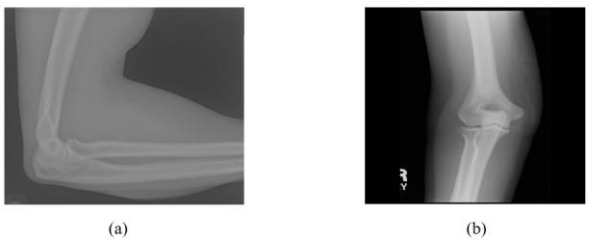
3.3. Dataset and Image preprocessing

The dataset features and source used in the study are as follows. The MURA dataset was collected from HIPAA (Health Insurance Portability and Accountability Act) compliant images from Stanford Hospital's Picture Archive and Communication System (PACS). MURA is a large radiography dataset containing 14,863 musculoskeletal studies and a total of 40,561 multi-view radiographic images of 12,173 patients between 2001 and 2012. This is one of the most extensive publicly available radiographic image datasets. The dataset was manually labeled as healthy or an anomaly by radiologists. The dataset consists of 9,045 healthy and 5,818 anomalies. It includes the radiographic study of the musculoskeletal system of the humerus, shoulder, forearm, elbow, wrist, finger, and hand [6,30]. Table 2 shows the distribution of the dataset.

TABLE 2. Distribution of Stanford MURA dataset for upper body studies [31]

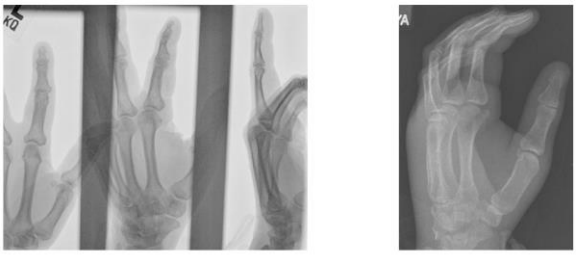
	Train		Validation	
Part	Healthy	Anomaly	Healthy	Anomaly
Elbow	2925	2006	234	230
Finger	3138	1968	214	247
Hand	4059	1484	271	189
Humerus	673	599	148	140
Forearm	1164	661	150	151
Shoulder	4211	4168	285	278
Wrist	5765	3987	364	295

Anomaly detection is a binary classification task that determines whether a study is healthy or an anomaly. Determining whether a radiographic study is healthy or an anomaly is critical; it can eliminate the requirement for patients to undergo further diagnostic tests, procedures, and interventions. Anomalies include fractures, hardware, degenerative joint diseases, lesions, and subluxations [32]. Figures 9 and 10 show the X-ray images for the elbow and finger parts of the MURA dataset.



(a) (b)

Fig.9. MURA dataset elbow part (a) healthy, (b) anomaly example.



(a) (b)

Fig.10. MURA dataset finger part (a) healthy, (b) anomaly example.

The original dimensions of the images in the MURA dataset are not fixed and vary between 512x512 pixels and 97x512 pixels. The file extension of the images is '.png' [12]. Since the input data in deep learning must have the same pixel value, all variable-size images were resized to 320x320 pixels [33]. After this resizing, images with '.png' extensions were centered by trimming the excess edges or spaces, as shown in Fig. 9 and Fig. 10; the image was centered. This process removes unnecessary or empty areas around the edges to bring the image closer to the focal point or area of interest. The bit depth of the images in the MURA dataset varies between 8 and 24. In order to make the training in deep learning more efficient, the bit depth of all input image data in the study was converted to 8 [34]. The input image data was reproduced by randomly rotating it horizontally and vertically between -30^0 and $+30^0$ and reflecting it on both axes [1]. In addition, the input image data was increased by scaling the input image data between 0,9 and 1,1 [35].
The number of radiography data used in experimental analyses is shown in Table 3.

TABLE 3. Summary of some studies using the MURA radiography dataset

	Part	Healthy	Anomaly
In this study	Elbow	1534	1630
Harini et al.[8]	Elbow	2925	2006
Kumar and Cutsuridis [30]	Elbow	162	160
In this study	Finger	1965	1938
Harini et al.[8]	Finger	3138	1968
Kumar and Cutsuridis [30]	Finger	175	164

4. PERFORMANCE METRICS

Performance of modeling for healthy-anomaly detection in MURA data set; evaluated using clinically meaningful statistical measures such as accuracy, precision, recall, specificity, F1-score, k-fold cross validation, Cohen's kappa statistic and area under the curve (AUC). These criteria are briefly defined as follows:

4.1. Accuracy

It is a parameter that evaluates the capacity of a model by measuring the proportion of correctly predicted cases out of the total number of cases. It is expressed mathematically as:

$$\text{Accuracy} = (TP + TN)/ (TP + FP + FN + TN) \qquad (1)$$

Here, TP is the number of positive cases correctly predicted by the model; TN is the number of negative cases correctly predicted by the model; FP is the number of positive cases incorrectly predicted by the model; FN refers to the number of negative cases incorrectly predicted by the model. However, accuracy may only sometimes be an excellent metric to evaluate the performance of the model, especially in the case of asymmetric data sets. Therefore, it is necessary to evaluate other performance metrics to test the model.

4.2. Precision

The ratio of correctly predicted positive cases to total positive cases. A high precision value is associated with a low FP rate. Precision is calculated as follows:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

4.3. Specificity

The ratio of correctly predicted negative observations to all true negative observations.

$$\text{Specificity} = \text{TN}/(\text{FP} + \text{TN}) \quad (3)$$

4.4. Recall

Recall is a metric that shows how many trades we should predict as positive we predict as positive.

$$\text{Recall}=\text{TP}/(\text{TP}+\text{FN}) \quad (4)$$

4.5. F1-Score

The F1 Score is measured primarily in the case of uneven class distribution with many accurate negative observations. F1-score provides a balance of precision and recall [36].

$$\text{F1-Score}=2(\text{PrecisionxRecall})/(\text{Precision}+\text{Recall}) \quad (5)$$

4.6. K-fold cross-validation

K-fold is a cross-validation method in which we iterate k times over a dataset. In standard k-fold cross-validation, we partition the data into k subsets called folds. Then, the algorithm is iteratively trained on k-1 folds while using the remaining fold as the test set (called the "holdout fold") [37]. The literature chooses the most suitable k values as 3, 5, and 10. Cross-validation is used to prevent overfitting problems [38].

4.7. Cohen's kappa statistic

The kappa statistic measures how well two different assessors or tests agree with each other. The formula for Cohen's kappa statistic (κ) is as follows [39]:

$$\kappa = (\text{Accuracy} - P_e)/(1 - P_e) \quad (6)$$

where,

$$P_e = \frac{(TP+FP)\times(TP+FN)+(TN+FN)\times(TN+FP)}{(TP+TN+FP+FN)^2} \quad (7)$$

4.8. Area under the curve (AUC)

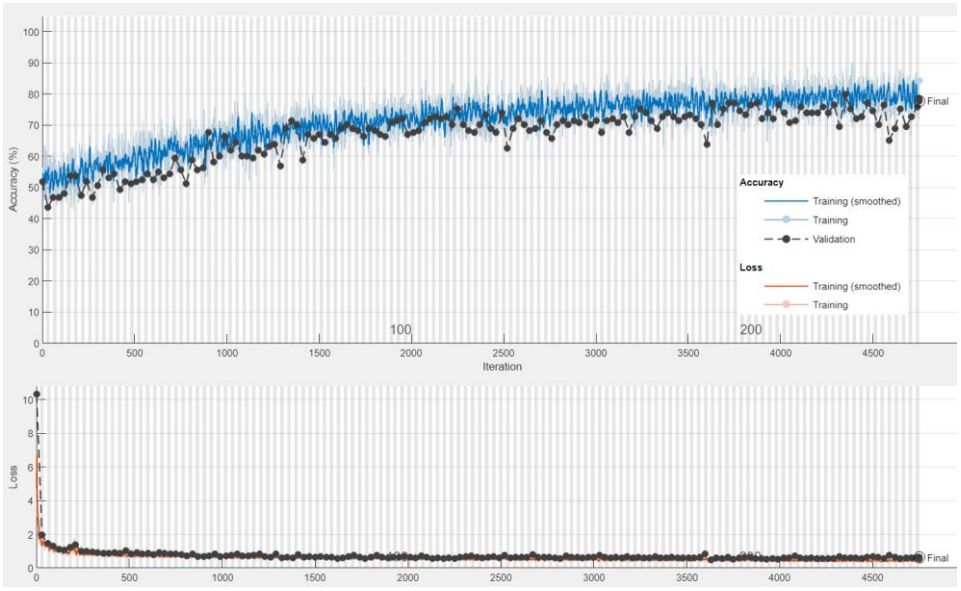
The ROC is a probability curve, and the area under the AUC represents the degree or measure of separability. As the area under the curve increases, the discrimination performance between classes increases [40]. Formally, the formula for calculating AUC is

$$AUC = \int_0^1 f(x)dx \quad (8)$$

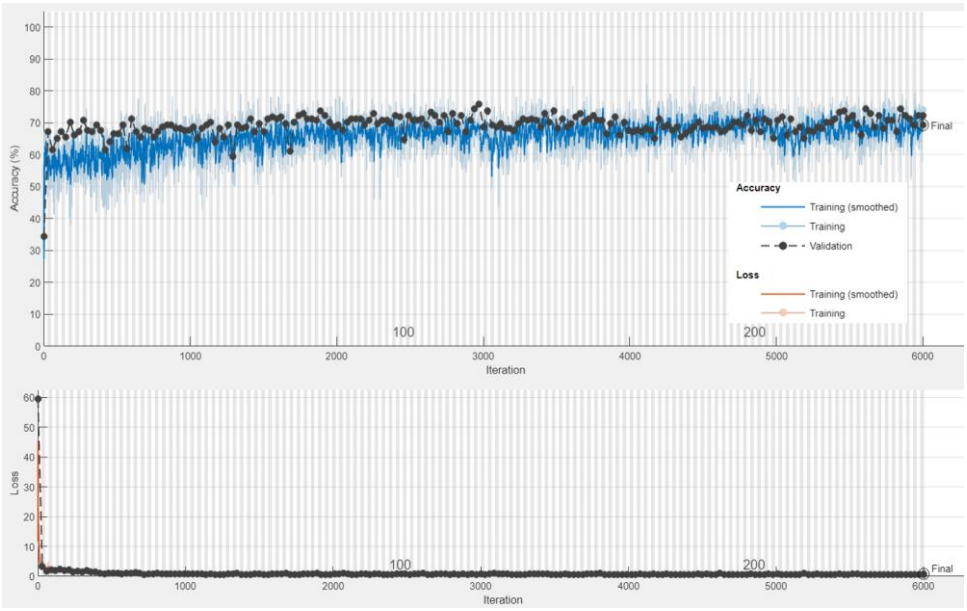
5. EXPERIMENTAL ANALYSIS AND RESULTS

The experimental environment is defined as follows. All network models in this experiment are based on deep learning frameworks, and each model trial was conducted using a computer with 12th Gen Intel(R) Core(TM) i7-12650H 2.30 GHz, Windows 11 Pro operating system, 16 GB memory, and NVIDIA GeForce RTX 4060 Laptop GPU. The experiments were completed using the MATLAB program.

Parameter settings are as follows. The networks in this experiment used the same parameter settings. In the data set, the learning rate was set to 0.001, the minibatch size to 128, and the number of epochs to 250. In order to minimize the amount of error between the output value produced by the Model Network and the actual value, the categorical cross-entropy function was adjusted using the ADAM (Adaptive Moment Estimation) algorithm. In this way, the desired goal was tried to be achieved by minimizing the difference between the output value produced by the model and the actual value. Data input images were randomly trained in all modeling with 80% training, 5% validation, and 15% testing rates. Figure 11 shows the Accuracy and Loss graphs for the elbow (a) and finger (b) parts.



(a)



(b)

Fig.11. Training Progress for the elbow (a) and finger (b) parts.

In training, 1534 healthy and 1630 anomaly images were used for the elbow part of the MURA radiography dataset. Confusion matrices of the test data obtained using the AlexNet, DenseNet, ParallelDenseNet, and PPDN models are shown in Tables 4, 5, 6, and 7, respectively.

For the elbow part AlexNet model, out of 246 images labeled as an anomaly, 191 were successful (TP), and 55 were unsuccessful (FN). Of the 229 images labeled healthy, 54 gave unsuccessful (FP) results, and 175 gave successful (TN) results.

TABLE 4. Elbow part AlexNet confusion matrix

Output Class		Anomaly	Healthy	
	Anomaly	**191** 40,2%	**55** 11,6%	77,6% 22,4%
	Healthy	**54** 11,4%	**175** 36,8%	76,4% 23,6%
		78,0% 22,0%	76,1% 23,9%	**77,1%** **22,9%**
		Anomaly	Healthy	
		Target Class		

TABLE 5. Elbow part DenseNet confusion matrix

Output Class		Anomaly	Healthy	
	Anomaly	**161** 33,9%	**36** 7,6%	81,7% 18,3%
	Healthy	**84** 17,7%	**194** 40,8%	69,8% 30,2%
		65,7% 34,3%	84,3% 15,7%	**74,7%** **25,3%**
		Anomaly	Healthy	
		Target Class		

For the elbow part DenseNet model, out of 197 images labeled as an anomaly, 161 were successful (TP), and 36 were unsuccessful (FN). Of the 278 images labeled as healthy, 84 gave unsuccessful (FP) results, and 194 gave successful (TN) results.

TABLE 6. Elbow part Parallel DenseNet confusion matrix

Output Class		Anomaly	Healthy	
	Anomaly	167 35,2%	33 6,9%	83,5% 16,5%
	Healthy	78 16,4%	197 41,5%	71,6% 28,4%
		68,2% 31,8%	85,7% 14,3%	76,6% 23,4%
		Anomaly	Healthy	
		Target Class		

For the elbow part of the Parallel DenseNet model, out of 200 images labeled as an anomaly, 167 were successful (TP), and 33 were unsuccessful (FN). Out of the 275 images labeled healthy, 78 gave unsuccessful (FP) results, and 197 gave successful (TN) results.

TABLE 7. Elbow part PPDN confusion matrix

Output Class		Anomaly	Healthy	
	Anomaly	179 37,7%	35 7,4%	83,6% 16,4%
	Healthy	66 13,9%	195 41,1%	74,7% 25,3%
		73,1% 26,9%	84,8% 15,2%	78,7% 21,3%
		Anomaly	Healthy	
		Target Class		

For the elbow part PPDN model, out of 214 images labeled as an anomaly, 179 were successful (TP), and 35 were unsuccessful (FN). Out of the 261 images labeled as healthy, 66 gave unsuccessful (FP) results, and 195 gave successful (TN) results. Figure 12 shows the ROC curve plot for the elbow part trained with the PPDN model.

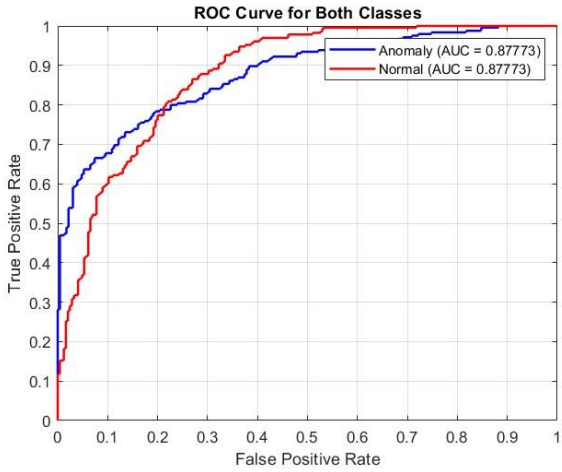


Fig.12. ROC curve for the elbow part with the PPDN model.

Table 8 shows the performances of four deep-learning models for the elbow part.

TABLE 8. Comparative performances of deep learning models for the elbow part

Models	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	F1-Score (%)	Kappa Score	5-fold Accuracy (%)	AUC	Training time for models (min)
AlexNet	77,05	77,64	76,42	77,96	77,80	0,5405	75,81	0,84720	51
DenseNet	74,74	81,73	69,78	65,71	72,85	0,4974	73,10	0,83063	175
Paralel DenseNet	76,63	83,50	71,64	68,16	75,06	0,5350	75,45	0,83972	678
PPDN	78,74	83,64	74,71	73,06	78,00	0,5761	77,95	0,87773	886

According to Table 8, the highest accuracy rate (78,74%) was seen in the PPDN model, and the lowest accuracy rates were seen in the DenseNet (74,74%) and Parallel DenseNet (76,63%) models. On the other hand, although the AlexNet model (77,05%) gives more successful results than the DenseNet and Parallel DenseNet models, it has a lower accuracy rate than the PPDN model. In their study using the AlexNet model, Yang and Ding classified the elbow part with an accuracy rate of 72,39% [41]. In their study, Karthik and Kamath classified the elbow part using the AlexNet model with an accuracy rate of 78,67% [42]. These values are close to the 77,05% value obtained as a result of the analysis in this study. The Kappa statistic value of 0,5761 was obtained in the PPDN model is more successful than all models. In the analyses performed with the 5-fold cross-validation method, the PPDN model outperformed the other models with an accuracy of 77.95%. The AUC value of 0,87773 was obtained in the PPDN model, which is more successful than all models. In training, 1965 healthy and 1938 anomaly images were used for the finger part of the MURA radiography dataset. Confusion matrices of the test data obtained using the AlexNet, DenseNet, Parallel DenseNet, and PPDN models are shown in Table 9, Table 10, Table 11, and Table 12, respectively.

TABLE 9. Finger part AlexNet confusion matrix

Output Class		Anomaly	Healthy	
	Anomaly	**224** 38,2%	**124** 21,2%	64,4% 35,6%
	Healthy	**67** 11,4%	**171** 29,2%	71,8% 28,2%
		77,0% 23,0%	58,0% 42,0%	**67,4%** **32,6%**
		Anomaly	Healthy	
		Target Class		

For the finger part AlexNet model, out of 348 images labeled as an anomaly, 224 were successful (TP), and 124 were unsuccessful (FN). Of the 238 images labeled healthy, 67 gave unsuccessful (FP) results, and 171 gave successful (TN) results.

TABLE 10. Finger part DenseNet confusion matrix

Output Class		Anomaly	Healthy	
	Anomaly	**183** 31,2%	**83** 14,2%	68,8% 31,2%
	Healthy	**108** 18,4%	**212** 36,2%	66,2% 33,8%
		62,9% 37,1%	71,9% 28,1%	**67,4%** **32,6%**
		Anomaly	Healthy	
		Target Class		

For the finger part DenseNet model, out of 266 images labeled as an anomaly, 183 were successful (TP), and 83 were unsuccessful (FN). Out of the 320 images labeled as healthy, 108 gave unsuccessful (FP) results, and 212 gave successful (TN) results.

TABLE 11. Finger part Parallel DenseNet confusion matrix

Output Class		Anomaly	Healthy	
	Anomaly	**204** 34,8%	**95** 16,2%	68,2% 31,8%
	Healthy	**87** 14,8%	**200** 34,1%	69,7% 30,3%
		70,1% 29,9%	67,8% 32,2%	**68,9%** **31,1%**
		Anomaly	Healthy	
		Target Class		

For the finger part Parallel DenseNet model, out of 299 images labeled as an anomaly, 204 were successful (TP), and 95 were unsuccessful (FN). Of the 287 images labeled healthy, 87 gave unsuccessful (FP) results, and 200 gave successful (TN) results.

TABLE 12. Finger part PPDN confusion matrix

Output Class		Anomaly	Healthy	
	Anomaly	**180** 30,7%	**65** 11,1%	73,5% 26,5%
	Healthy	**111** 18,9%	**230** 39,2%	67,4% 32,6%
		61,9% 38,1%	78,0% 22,0%	**70,0%** **30,%**
		Anomaly	Healthy	
		Target Class		

For the finger part PPDN model, out of 245 images labeled as an anomaly, 180 were successful (TP), and 65 were unsuccessful (FN). Of the 341 images labeled healthy, 111 gave unsuccessful (FP) results, and 230 gave successful (TN) results. Figure 13 shows the ROC curve plot for the finger part trained with the PPDN model.

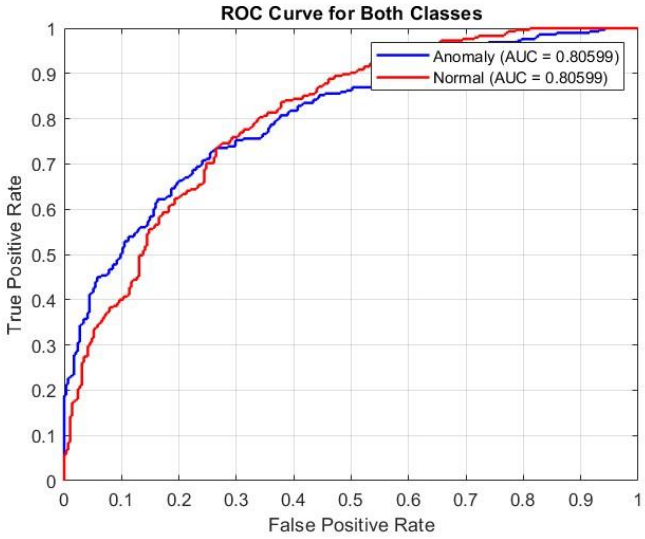


Fig.13. ROC curve for the finger part with the PPDN model.

Table 13 shows the performances of four deep-learning models for the finger part. According to Table 13, the highest accuracy rate (69,97%) was seen in the PPDN model, and the lowest accuracy rates were seen in the DenseNet (67,41%) and AlexNet (67,41%) models. On the other hand, although the Parallel DenseNet model (68,94%) gives more successful results than the DenseNet and AlexNet models, it has a lower accuracy rate than the PPDN model. A high accuracy of the PPDN model means that the number of correct predictions is high, but the F1 score may decrease when the FP or FN is high. This is often the case with imbalanced data sets or classification problems. Accuracy is the ratio of correctly classified instances to all instances, while the F1 score takes into account the imbalance between classes and focuses more on the agreement of precision and recall values. Especially when there is a large difference in the number of positive and negative samples, the model can achieve a high accuracy by biasing towards the majority class.

TABLE 13. Comparative performances of deep learning models for the finger part

Models	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	F1-Score (%)	Kappa Score	5-fold Accuracy (%)	AUC	Training time for models (min)
AlexNet	67,41	64,37	71,85	76,98	70,11	0,3490	65,17	0,73784	66
DenseNet	67,41	68,80	66,25	62,89	65,71	0,3477	65,32	0,70846	220
Paralel DenseNet	68,94	68,23	69,69	70,10	69,15	0,3789	67,89	0,75951	926
PPDN	69,97	73,47	67,45	61,86	67,16	0,3986	68,92	0,80599	1125

However, in this case, the F1 score may be low because the model fails to correctly predict the minority class samples [43].

In their study using the AlexNet model, Yang and Ding classified the finger part with an accuracy rate of 70,57% [41]. In their study, Karthik and Kamath classified the finger part using the AlexNet model with an accuracy rate of 71,13%. [42]. These values are close to the 67,41% value obtained due to the analysis using the AlexNet model in this study.

The Kappa statistic value of 0,3986 obtained in the PPDN model is more successful than all models. In the analyses performed with the 5-fold cross-validation method, the PPDN model outperformed the other models with an accuracy of 68,92%, which was more successful than all models. The AUC value of 0,80599 obtained in the PPDN model is more successful than all models.

In deep learning, hyperparameter tuning may not be necessary due to good results with default or simple settings, time and cost limitations, and problem complexity [44]. Considering the dataset's size and the performance metrics results, no hyperparameter tuning was made to compare the PPDN model's effect with the classical DenseNet and AlexNet models.

6. CONCLUSION AND DISCUSSION

The elbow and finger X-ray images from the MURA data set were categorized as healthy or anomalous using CNN models in deep learning. Shortening the treatment period by early diagnosis of musculoskeletal system disease is essential. The classical DenseNet model was developed in this study, and the test results were compared with those of other models. According to the results obtained, it is expected to positively contribute to the decision-making process regarding diagnosing musculoskeletal diseases. The proposed method offers significant advantages in matters such as test accuracy rate and personnel workload. As a result of the statistical analysis made for the elbow part, The highest test accuracy rate (78,74%) was seen in the PPDN model, and the lowest accuracy rates were seen in the DenseNet (74,74%) and Parallel DenseNet (76,63%) models. On the other hand, the AlexNet model (77,05%) achieved more successful results than the DenseNet and Parallel DenseNet models. As a result of the statistical analysis made for the finger part, The highest test accuracy rate (69,97%) was seen in the PPDN model, and the lowest accuracy rates were seen in the DenseNet (67,41%) and AlexNet (67,41%) models. In addition, for the finger part, the Parallel DenseNet model (68,94%) achieved more successful results than the DenseNet and AlexNet models. It is common for finger fractures to be difficult to diagnose and sometimes overlooked by doctors. Some fractures may not be clearly visible on radiographs and may be misdiagnosed. For example, hairline fractures or small calcifications may be more difficult to detect on finger images. Such fractures can lead to complications if ignored and may require re-evaluation for treatment. Some finger joint dislocations can also be misdiagnosed without x-rays, simply due to swelling, and may not be considered a serious ligament injury. Therefore, the accuracy values obtained in the finger section are relatively low [45, 46]. For the finger part, the test accuracy value of the classical DenseNet model used in the study by Harini et al. (49.67%) while the test accuracy value of the PPDN model, which is the DenseNet model developed in our study, is (69.97%). For the elbow part, the test accuracy value of the classical DenseNet model in the study by Cheng et al. [9] is (62,68%). In contrast, the test accuracy value of the PPDN model, which is the DenseNet model developed in our study, is (78,74%). In the study conducted by S. Lysdahlgaard [10] for the elbow part, the validation and test accuracy values for all DenseNet variants are below our study's test accuracy result values. When Kappa statistic value measurements, 5-fold cross-validation accuracy values, and AUC values are compared, the PPDN model is more successful than the classical DenseNet model for both elbow and finger parts. We propose the PPDN model developed in our study based on the results obtained, as the feature extraction process works better than the classical DenseNet model. The disadvantage of the PPDN model is that the training time is longer than that of other models. The successful results obtained in the test accuracy rates of the PPDN model are beneficial for the early diagnosis of musculoskeletal disorders in different datasets and for integrating the proposed method into computer-aided software systems used in hospitals in future studies. However, in future studies, X-ray images obtained by radiologists focusing on anomaly areas can be improved methodologically and technically.

REFERENCES

[1] M. He, X. Wang, and Y. Zhao, "A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs," *Scientific Reports*, vol. 11, no. 1, Apr. 2021, doi: 10.1038/s41598-021-88578-w.

[2] I. Kandel and M. Castelli, "Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset," *Health Information Science and Systems*, vol. 9, no. 1, Jul. 2021, doi: 10.1007/s13755-021-00163-7.

[3] R. Lindsey *et al.*, "Deep neural network improves fracture detection by clinicians," *Proceedings of the National Academy of Sciences*, vol. 115, no. 45, pp. 11591–11596, Oct. 2018, doi: 10.1073/pnas.1806905115.

[4] V. Narayan, P. K. Mall, A. Alkhayyat, K. Abhishek, S. Kumar, and P. Pandey, "Enhance-Net: An Approach to Boost the Performance of Deep Learning Model Based on Real-Time Medical Images," *Journal*

of Sensors, vol. 2023, p. e8276738, May 2023, doi: 10.1155/2023/8276738.
[5] S. Panda and Mahesh Jangid, "Improving the Model Performance of Deep Convolutional Neural Network in MURA Dataset," *Smart innovation, systems and technologies*, pp. 531–541, Oct. 2019, doi: 10.1007/978-981-13-8406-6_51.
[6] G. Mehr, "Automating Abnormality Detection in Musculoskeletal Radiographs through Deep Learning," *arXiv (Cornell University)*, Oct. 2020, doi: https://doi.org/10.48550/arxiv.2010.12030.
[7] S. Liang and Y. Gu, "Towards Robust and Accurate Detection of Abnormalities in Musculoskeletal Radiographs with a Multi-Network Model," *Sensors*, vol. 20, no. 11, p. 3153, Jun. 2020, doi: 10.3390/s20113153.
[8] N. Harini, B. Ramji, S. Sriram, V. Sowmya, and K. P. Soman, "Musculoskeletal radiographs classification using deep learning," *Elsevier eBooks*, pp. 79–98, Jan. 2020, doi: 10.1016/b978-0-12-819764-6.00006-5.
[9] K. Cheng, C. Iriondo, F. Calivá, J. Krogue, S. Majumdar, and V. Pedoia, "Adversarial Policy Gradient for Deep Learning Image Augmentation," *Lecture Notes in Computer Science*, pp. 450–458, Jan. 2019, doi: 10.1007/978-3-030-32226-7_50.
[10] S. Lysdahlgaard, "Utilizing heat maps as explainable artificial intelligence for detecting abnormalities on wrist and elbow radiographs," *Radiography*, vol. 29, no. 6, pp. 1132–1138, Oct. 2023, doi: 10.1016/j.radi.2023.09.012.
[11] A. Solovoya, I. Solovyov, and Traumai, "X-Ray bone abnormalities detection using MURA dataset," 2020. Accessed: Feb. 05, 2024. [Online]. Available: https://arxiv.org/pdf/2008.03356.pdf
[12] I. Kandel, M. Castelli, and A. Popovič, "Comparing Stacking Ensemble Techniques to Improve Musculoskeletal Fracture Image Classification," *Journal of Imaging*, vol. 7, no. 6, p. 100, Jun. 2021, doi: 10.3390/jimaging7060100.
[13] T. C. Mondol, H. Iqbal, M. M. A. Hashem. Deep CNN-based ensemble CADx model for musculoskeletal abnormality detection from radiographs. In *2019 5th international conference on advances in electrical engineering (ICAEE)* (pp. 392-397), 2019, September. IEEE.
[14] L. Morra, L. Piano, F. Lamberti, T. Tommasi. Bridging the gap between natural and medical images through deep colorization. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 835-842). 2021, January. IEEE.
[15] M. Puttagunta and S. Ravi, "Medical image analysis based on deep learning approach," *Multimedia Tools and Applications*, Apr. 2021, doi: 10.1007/s11042-021-10707-4.
[16] Z. Alammar, L. Alzubaidi, J. Zhang, J. Santamaría, Y. Li, and Y. Gu, "A Concise Review on Deep Learning for Musculoskeletal X-ray Images," *IEEE Xplore*, Nov. 01, 2022. https://ieeexplore.ieee.org/abstract/document/10034618 (accessed Jul. 24, 2023).
[17] F. Piccialli, V. D. Somma, F. Giampaolo, S. Cuomo, and G. Fortino, "A survey on deep learning in medicine: Why, how and when?," *Information Fusion*, vol. 66, pp. 111–137, Feb. 2021, doi: 10.1016/j.inffus.2020.09.006.
[18] L. Yin, P. Hong, G. Zheng, H. Chen, and W. Deng, "A Novel Image Recognition Method Based on DenseNet and DPRN," *Applied sciences*, vol. 12, no. 9, pp. 4232–4232, Apr. 2022, doi: 10.3390/app12094232.
[19] R. Ghalyan, A. Singh, K. Kadian, V. Kumar, D. Yadav, "Bone X-Ray Classification For Upper Extremity Radiographs," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 1S, pp. 7082–7087, 2023, doi: 10.53555/sfs.v10i1S.2316.
[20] I. H. A. Kandel, "Deep Learning Techniques for Medical Image Classification," Doctoral dissertation, Universidade NOVA de Lisboa, Portugal, 2021.
[21] S. Lu, Z. Lu, and Y.-D. Zhang, "Pathological brain detection based on AlexNet and transfer learning," *Journal of Computational Science*, vol. 30, pp. 41–47, Jan. 2019, doi: 10.1016/j.jocs.2018.11.008.
[22] S. Li, L. Wang, J. Li, and Y. Yao, "Image Classification Algorithm Based on Improved AlexNet," *Journal of Physics: Conference Series*, vol. 1813, no. 1, p. 012051, Feb. 2021, doi: 10.1088/1742-6596/1813/1/012051.
[23] N. Zahan, M. Z. Hasan, M. S. Uddin, S. Hossain, and S. F. Islam, "A deep learning-based approach for mushroom diseases classification," *Elsevier eBooks*, pp. 191–212, Jan. 2022, doi: 10.1016/b978-0-323-90550-3.00005-9.
[24] M. K. Bohmrah and H. Kaur, "Classification of Covid-19 patients using efficient Fine-tuned Deep learning DenseNet Model," *Global Transitions Proceedings*, Aug. 2021, doi: 10.1016/j.gltp.2021.08.003.
[25] P. Podder, F. B. Alam, M. R. H. Mondal, M. J. Hasan, A. Rohan, and S. Bharati, "Rethinking Densely Connected Convolutional Networks for Diagnosing Infectious Diseases," *Computers*, vol. 12, no. 5, p. 95, May 2023, doi: 10.3390/computers12050095.
[26] B. Chen, T. Zhao, J. Liu, and L. Lin, "Multipath feature recalibration DenseNet for image classification," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 3, pp. 651–660, Sep. 2020, doi: 10.1007/s13042-020-01194-4.
[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Jul. 2017, doi: 10.1109/cvpr.2017.243.
[28] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, "Dense Convolutional Network and Its Application in Medical Image Analysis," *BioMed Research International*, vol. 2022, pp. 1–22, Apr. 2022, doi: 10.1155/2022/2384830.
[29] J.Yang, L. Zhang, X. Tang, "CrodenseNet: An efficient parallel cross DenseNet for COVID-19 infection detection", *Biomedical Signal Processing and Control*, 2022, doi:10.1016/j.bspc.2022.103775.
[30] N. Kumar and Vassilis Cutsuridis, "Deep Convolutional Neural Networks with Transfer Learning for Bone Fracture Recognition using Small Exemplar Image Datasets," Jun. 2023, doi: 10.1109/icasspw59220.2023.10193015.
[31] L. Alzubaidi, A. Salhi, M. A. Fadhel, J. Bai, F. Hollman, K. Italia, Y. Gu, (2024). Trustworthy deep learning framework for the detection of abnormalities in X-ray shoulder images. *Plos one*, vol *19. no* 3. doi: 10.1371/journal.pone.0299545.
[32] J. Vojtech, "Detecting abnormalities in X-Ray images using Neural Networks," Bachelor's thesis, Czech Technical University, Prague, 2022.
[33] A. Siddiqui, "neXt-Ray: Deep Learning on Bone X-Rays." Accessed: Feb. 05, 2024. [Online]. Available: https://cs230.stanford.edu/projects_fall_2020/reports/55773729.pdf
[34] L. Liao, W. Liu, and S. Liu, "Effect of Bit Depth on Cloud Segmentation of Remote-Sensing Images," *Remote Sensing*, vol. 15, no. 10, pp. 2548–2548, May 2023, doi: 10.3390/rs15102548.
[35] A. Karna, A. Jha, A. Dahal, A. Pandey, T. Jha, "Chest X-Ray Classification using DenseNet," *Proceedings of 13th IOE Graduate Conference*. 2023.
[36] G. Jain, D. Mittal, D. Thakur, and M. K. Mittal, "A deep learning approach to detect Covid-19 coronavirus with X-Ray images," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1391–1405, Oct. 2020, doi: 10.1016/j.bbe.2020.08.008.
[37] Ananda, C. Karabağ, A. Ter-Sarkisov, E. Alonso, and C. Carlos Reyes-Aldasoro, "Radiography Classification: A Comparison between Eleven Convolutional Neural Networks*," City Research Online (City University London)*, Oct. 2020, doi: 10.1109/mcna50957.2020.9264285.
[38] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation". Stat Comput 21, 137–146 (2011). doi: 10.1007/s11222-009-9153-8
[39] D. Normawati, D. P. Ismi, "K-fold cross validation for selection of cardiovascular disease diagnosis features by applying rule-based datamining". Signal and Image Processing Letters, 1(2), 62-72. 2019.
[40] A. J. Bowers, X. Zhou. Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, *24*(1), 20-46, 2019.
[41] F. Yang and B. Ding, "Computer Aided Fracture Diagnosis Based on Integrated Learning," *2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, Sep. 2020, doi: 10.1109/iciscae51034.2020.9236917.
[42] K. Karthik and S. Sowmya Kamath, "Correction to: MSDNet: a deep neural ensemble model for abnormality detection and classification of plain radiographs," *Journal of Ambient Intelligence and Humanized Computing*, May 2022, doi: 10.1007/s12652-022-03906-w.
[43] I. Goodfellow, Y. Bengio, A. Courville. *Deep Learning*. MIT Press.2016
[44] A. Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media. 2019.
[45] S. W. Wolfe, W. C. Pederson, S. H. Kozin, M. S. Cohen. *Green's Operative Hand Surgery*. 2010.
[46] W. P. Cooney. *The Wrist: Diagnosis and Operative Treatment.* Wolters Kluwer Health. 2011.