

**An Efficient Way to Detect Musculoskeletal Disease through X-Ray Images
Using Layer-freezing Technique**

By

Annur Anika

ID: 192-16-452

This Report Presented in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computing and Information System

Supervised By

Mr. Israfil

Lecturer

Department of Computing & Information System

Daffodil International University



Daffodil International University

July-2023

APPROVAL

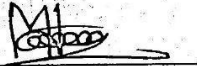
This thesis titled "An Efficient Way to Detect Musculoskeletal Disease through X-Rays Images Using Layer-freezing Technique", Submitted by Annur Anika, ID No: 192-16-452, to the Department of Computing & Information Systems, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computing & Information Systems and approved as to its style and contents. The presentation has been held on- 14-01-2023.

BOARD OF EXAMINERS



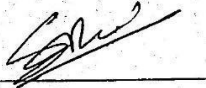
Mr. Md Sarwar Hossain Mollah
Associate Professor and Head
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Chairman



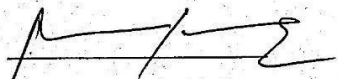
Mr. Md. Mehedi Hassan
Lecturer
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Mr. Syed Tangim Pasha
Lecturer
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



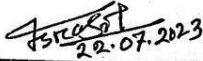
Dr. Saifuddin Md. Tareeq
Professor & Chairman
Department of Computer Science and Engineering
University of Dhaka, Dhaka

External Examiner

DECLARATION

I hereby declare that; this thesis has been done by me under supervision of Mr. Israfil, Lecturer, Department of Computing and Information System (CIS) of Daffodil International University. I am also declaring that this thesis or any part of there has never been submitted anywhere else for the award of any educational degree like, B.Sc., M.Sc., Diploma or other qualifications.

Supervised By


22.07.2023

Mr. Israfil
Designation: Lecturer
Department of CIS
Daffodil International University

Submitted By



Name: Annur Anika
ID: 192-16-452
Department of CIS
Daffodil International University

Acknowledgement

I want to express my sincere gratitude to the Almighty Allah for giving me his divine favor and enabling me to finish the senior thesis successfully.

I express my sincere gratitude and debt to Mr Israfil, Lecturer, Department of Computing & Information System, Daffodil International University. His extensive expertise, strong attention, and encouraging guidance enabled me to complete my study in data analysis, machine learning, deep learning, and computer vision. This research was made possible by his never-ending patience, academic guidance, constant encouragement, frequent and vigorous supervision, constructive criticism, invaluable advice, reviewing numerous subpar drafts and fixing them at all stages.

I want to extend my sincere gratitude to Mr Md Sarwar Hossain Mollah, Associate Professor and Head, Department of Computing & Information System, Daffodil International University, Dhaka, for his invaluable assistance and counsel in helping me complete my research. I also want to extend my sincere gratitude to the other professors and staff members of the Department of Computing & Information System, Daffodil International University, Dhaka.

Finally, I would like to thank all the well-wishers, friends, family, and elders for their support and inspiration. Hard labor, as well as all of that inspiration and help, went into this study.

Finally, I must respectfully appreciate my parents' unfailing help and tolerance.

Dedication

This thesis is dedicated to Almighty Allah and my parents, younger brother, and friends, who helped me finish my BSC program. I also praise and thank Almighty Allah for having supported, shielded, and guided me.

Abstract

Musculoskeletal diseases require immediate attention to avoid chronic issues since they impact over 1.7 billion individuals globally. Radiographic images frequently identify musculoskeletal illnesses and abnormalities in medical computer vision. Researchers have suggested many methods for this purpose. In this research, we suggest five state-of-the-art transfer learning models with a partial layer-freezing technique where initial layers of a pre-trained model were frozen, except the last ten layers, to identify such Musculoskeletal Abnormalities using the MURA dataset, one of the most extensive collections of upper extremity radiographs. This method outperforms the performance of the baseline model in finger and humerus studies by achieving Cohen's kappa score of 0.439 in the DenseNet169 model and 0.638 in the DenseNet121 model. Although the current approach faces challenges in reaching sufficient accuracy for wrist, elbow, forearm, hand, and shoulder studies, there are positive improvements. When applied to radiographic images, the DenseNet network surpasses the ResNet, Inception, and Xception networks when evaluated using five different evaluation metrics. This shows that progress is being made in enhancing the performance of neural networks for improved medical image processing and analysis.

Keywords: Binary Classification; Convolutional Neural Network; Transfer Learning; Layer-freezing; DenseNet network; ResNet; Inception; and Xception networks

Table of Contents

Acknowledgement	iv
Dedication	v
Abstract	vi
Chapter 1- Introduction	1
1.1 Introduction.....	1
1.2 Objective	2
1.3 Motivation.....	2
1.4 Rational of the study	3
1.5 Research Questions	3
1.6 Expected Outcome	3
1.7 Layout of the Report	4
Chapter 2 – Background Study	5
2.1 Introduction.....	5
2.2 Literature Review	5
2.3 Challenges.....	9
Chapter 3 - System Methodology.....	10
3.1 Methodology.....	10
3.2 Dataset.....	11
3.3 Image Preprocessing	12
3.4 CNN and Transfer Learning.....	13
3.5 CNN Architectures	14
3.6 Fine-tuning.....	15
3.7 Evaluation Metrics	18
Chapter 4 - Result Analysis and Discussion	20
4.1 Discussion.....	20
4.2 Wrist Dataset Result.....	20
4.3 Elbow Dataset Result	21
4.4 Finger Dataset Result	21
4.5 Forearm Dataset Result	22
4.6 Shoulder Dataset Result	22
4.7 Hand Dataset Result	23

4.8 Humerus Dataset Result	23
4.9 Overall Dataset Result.....	24
Chapter 5 - Conclusion.....	25
Chapter 6 - Limitation And Future Scope	26
6.1 Limitations	26
6.2 Future Scope	26
References.....	27
Plagiarism Report.....	29

List of Figures

Figure 3. 1. Block diagram of the methodology	11
Figure 3. 2. Before preprocessing images using CLAHE	12
Figure 3. 3. After preprocessing images using CLAHE.....	13
Figure 3. 4. Trainable layers in the CNN models	17

List of Tables

Table 2. 1: Table of Related Works	6
Table 3. 1: Total number of studies in MURA dataset	12
Table 4. 1: Results of wrist dataset	20
Table 4. 2: Results of elbow dataset.....	21
Table 4. 3: Results of finger dataset.....	21
Table 4. 4: Results of forearm dataset.....	22
Table 4. 5: Results of shoulder dataset.....	22
Table 4. 6: Results of hand dataset.....	23
Table 4. 7: Results of humerus dataset.....	23
Table 4. 8: Results of overall dataset	24

Chapter 1- Introduction

1.1 Introduction

Musculoskeletal conditions are a severe concern that must be addressed immediately to avoid serious long-term issues, including obesity, chronic pain, and disability, affecting people's quality of life [1]. Musculoskeletal disability affects many people, including radiologists who spend hours detecting abnormalities in patients' radiographs. A study on radiologists showed that being a female radiologist between the ages of 30-39 and examining Computed Tomography (CT) or Ultrasound images was related to a higher risk of severe musculoskeletal problems [2]. This leads to a serious concern about reducing the workload from the radiologists and automating the detection system so that radiologists do not have to spend 7 to 9 hours reviewing the CT Scan. To address the issue, Deep Learning models, especially Convolutional Neural Networks, have achieved significant results in detecting abnormalities from X-ray images. Krizhevsky et al. proposed a CNN-based model that has drawn many researchers' attention [3]. However, to improve accuracy and avoid overfitting, CNN requires a large amount of data, which is insufficient in the medical industry. To solve this problem, Rajpurkar et al. developed the MURA dataset, which consists of 40,561 images from 14,863 upper extremity studies, including the shoulder, humerus, elbow, forearm, wrist, hand, and finger, and is an extensive collection of musculoskeletal radiographs [4]. Each study is manually classified as normal or abnormal by radiologists. The authors used DenseNet169 CNN to compare how well it performed to three radiologists and obtained 0.815 sensitivity, 0.887 specificity, and 0.929 AUCROC. To improve the accuracy of this baseline model and properly identify abnormalities in an emergency, we proposed the layer-freezing technique in the transfer learning models for the MURA dataset. Our paper focuses on pre-training several cutting-edge CNN models with ImageNet weights using transfer learning, modifying the pre-trained model by adding additional layers on top of the existing model and using regularization techniques to prevent overfitting. We trained only the last ten layers of the model to use the pre-trained generic features while still learning task-specific features.

1.2 Objective

The objectives of this paper are:

1. To analyze the effect of the partial layer-freezing technique on the medical images and whether this technique improves the performance of the baseline model.
2. To identify the best CNN model which outperforms the other models and achieves the best result.
3. To compare the results between the baseline model and our model. We use five evaluation metrics, including accuracy, Cohen's Kappa, ROCAUC score, precision, and recall/sensitivity, to measure the performance of our model.
4. To identify the best-performing dataset among the seven upper-extremity studies in the MURA dataset in our model.

1.3 Motivation

The significant effects of musculoskeletal problems on people's lives, including the possibility of obesity, persistent pain, and disability, served as the motivation for this thesis. Convolutional Neural Networks (CNNs), a type of deep learning model, have demonstrated promising results in identifying anomalies in X-ray pictures. This thesis suggests using the layer-freezing approach in transfer learning models customized to the MURA dataset to improve the baseline model's performance and reliably identify irregularities in emergency scenarios. Transfer learning will be used to pre-train state-of-the-art CNN models with ImageNet weights, and the pre-trained models are then enhanced by adding new layers and using regularization techniques to reduce overfitting. This thesis intends to create a more effective and accurate system for recognizing abnormalities in X-ray pictures by addressing the problems associated with musculoskeletal condition identification and the requirement for workload reduction for radiologists. The suggested method of layer-freezing transfer learning with regularization approaches can improve emergency diagnosis in musculoskeletal radiography and improve the performance of existing models.

1.4 Rational of the study

There has been significant research on the MURA dataset where most researchers preferred CNN to detect the abnormalities. However, the layer-freezing technique was not utilized in previous research. So, for the first time, we analyze this technique with five state-of-the-art CNN models where the training process selectively fine-tunes only the last ten layers of the model, enabling the utilization of pre-trained generic features while concurrently learning task-specific features. We will apply this technique to seven upper-extremity radiographic studies and labeled each study as normal or abnormal.

1.5 Research Questions

The MURA dataset is open-source, so we did not have to prepare the dataset from scratch. We created seven different datasets from the MURA dataset, each representing different upper-extremity studies. Moreover, there is limited research on layer-freezing techniques, so it was challenging to implement this method on a large dataset. So, before starting our research, we narrowed it down to a few topics and focused on addressing and addressing those concerns.

1. Does the layer-freezing method work on radiographic images?
2. Is it feasible to analyze the impact of CNN models?
3. Is it feasible to compare the accuracy of different models while working with the same datasets?

1.6 Expected Outcome

In order to detect abnormalities in radiographic images, as was previously said, we used the layer-freezing approach for the first time and compared the results on five cutting-edge transfer learning models.

- It will be able to detect abnormalities in seven different studies.

- It will outperform the baseline model.
- It will represent a noble technique to work on X-ray images.

1.7 Layout of the Report

The layout of this paper is organized as follows:

Chapter 1 briefly describes the motivation, importance, and method used in this study.

Chapter 2 discusses some of the research papers regarding the MURA dataset.

Chapter 3 represents the system model and methodology of the layer-freezing technique.

Chapter 4 analyzes and discusses the results of the seven upper-extremity studies in the MURA dataset.

Chapters 5 and 6 conclude the paper with the limitations and future scope.

Chapter 2 – Background Study

2.1 Introduction

This part will include literature reviews, research roadblocks, and an evaluation of our findings in light of previous research. We shall contrast our study with previous research articles' methodology, efforts, results, accuracy, and other factors. In the Challenges section, we will describe how we overcame challenges to carry out our research and the lessons we took away from them.

2.2 Literature Review

After Rajpurkar et al. [4] introduced the MURA dataset, researchers published many research papers to improve the accuracy of the baseline. In 2019, Saif et al. [5] published a paper entitled "Abnormality Detection in Musculoskeletal Radiographs Using Capsule Network." For musculoskeletal radiograph abnormality detection, the authors introduced the capsule network architecture. This architecture has demonstrated auspicious properties that can contribute to overcoming CNN's limitations. Despite using less training data, this capsule network outperformed a 169-layer DenseNet by 10% in terms of kappa score. In the same year, Banga and Waiganjo [6] introduced the ensemble200 model in the paper entitled "Abnormality Detection in Musculoskeletal Radiographs with Convolutional Neural Networks (Ensembles) and Performance Optimization," which surpassed the DenseNet model on the finger studies with a Cohen Kappa score of 0.653, indicating reduced model performance variability. The following year, Tirpude et al. [7] suggested a 169-layer densely connected convolutional network for detecting abnormalities in the MURA dataset. While the model performs well on finger, hand, and wrist tests, it struggles to detect abnormalities in elbow, forearm, humerus, and shoulder studies but still achieves high accuracy. Later in 2020, Kendal et al. [8] evaluated and compared six CNN architectures to transfer learning and a network built from scratch in the paper "Musculoskeletal Images Classification for Detection of Fractures Using Transfer Learning." The author observed that transfer learning outperformed training the networks from scratch. Twenty-six deep learning-based

pre-trained models and two ensemble learning models (EL1 and EL2) were developed by Uysal et al. [9] in the paper entitled "Classification of Shoulder X-ray Images with Deep Learning Ensemble Models" in 2021. In developing EL1 and EL2 models, pre-trained models such as ResNet, ResNeXt, DenseNet, VGG, Inception, MobileNet, and their Spinal FC versions are used. Among the 28 distinct classifications, the EL2 model had the best test accuracy and Cohen's kappa values, whereas the EL1 model had the most prominent area associated with the fracture class under the receiver operating characteristic curve. For EL1, the test accuracy and Cohen's kappa values were 0.8455 and 0.6907, respectively, and 0.8472 and 0.6942 for EL2. The AUC for EL1 was 0.8862 and 0.8695 for EL2. Ibrahem Kandel and Mauro Castelli [10] proposed TTA, which can improve model prediction by giving various transformations for the same image at a low computing cost in the paper entitled, "Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset" in 2021. In this study, they evaluated the influence of TTA on image classification performance on the MURA dataset. They discovered that TTA allows for higher performance across multiple datasets and CNNs than without TTA. In a recent study in 2022, Singh et al. [11] proposed ComDNet512 model on the finger studies in MURA dataset in the paper entitled "Hybrid Deep Learning Approach for Automatic Detection in Musculoskeletal Radiographs." With an accuracy of 89.41%, this model detected anomalies in finger radiographs. The three models outperformed current models when applied to finger radiographs. The model successfully achieved an area under the ROC curve (AUC) of 0.894. Precision, recall, F1 Score, and Kappa were all 0.86, 0.94, 0.89, and 0.78, respectively.

Table 2. 1: Table of Related Works

Paper ID	Title	Year	Key technologies or Methods (Algorithm)	Contribution	Research Gap
1	Abnormality Detection in Musculoskeletal	2019	Capsule Network	Outperformed a 169-layer	A shallow network,

	Radiographs Using Capsule Network			DenseNet by 10% in terms of kappa score	explored in limited areas.
2	Abnormality Detection in Musculoskeletal Radiographs with Convolutional Neural Networks(Ensembles) and Performance Optimization	2019	ensemble200 model [DenseNet201, MobileNet, NASNETmobile]	Outperformed the baseline model on the finger studies with a Cohen Kappa score of 0.653	Cohen Kappa was lower than the DenseNet model
3	Abnormal X-Ray Detection System using Convolution Neural Network	2020	CNN: DenseNet169	In finger, hand, and wrist studies achieves a good amount of accuracy	In the elbow, forearm, humerus, and shoulder, studies find it difficult to detect the abnormality
4	Musculoskeletal Images Classification for Detection of Fractures Using Transfer Learning	2020	Trained on VGG, Xception, ResNet, GoogleNet, InceptionResNet, DenseNet, and network from scratch	less prone to overfitting	fully connected layers had a negative effect on the performance

5	Classification of Shoulder X-ray Images with Deep Learning Ensemble Models	2021	pre-trained models such as ResNet, ResNeXt, DenseNet, VGG, Inception, MobileNet, and their Spinal FC versions	test accuracy was 0.8455, 0.8472, Cohen's Kappa was 0.6907, 0.6942	Studied only shoulder images
6	Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset	2021	Test Time Augmentation	TTA can increase the classifier's performance without adding any computational cost during training	TTA requires more time to evaluate the whole test set.
7	Hybrid Deep Learning Approach for Automatic Detection in Musculoskeletal Radiographs	2022	ComDNet512 model	identified abnormalities in finger radiographs with an accuracy of 89.41%	Studied only finger images

2.3 Challenges

The most challenging part of this research paper was working with 40,561 multi-view radiographic images. The dataset took longer to train, which was not feasible for this research. We divided the huge dataset into seven different datasets according to the seven upper-extremity studies to solve this issue. Another issue was with the preprocessing of the images. We needed to experiment with various features of data augmentation techniques to analyze which features work best with the images, which took significant time in our study. However, the whole process was challenging and enjoyable, extending our knowledge about deep learning.

Chapter 3 - System Methodology

3.1 Methodology

This study explored five state-of-the-art CNN models, ResNet152V2, DenseNet121, InceptionV3, Xception, and DenseNet169, and compared the effect on shoulder, humerus, elbow, forearm, wrist, hand, and finger studies. The radiographic images from the MURA dataset were preprocessed using data augmentation and CLAHE techniques. Transfer learning models were used for feature extraction. One of the primary reasons for using transfer learning to extract features from target images is the ability to capture information obtained by pre-trained models on large-scale datasets like ImageNet [12]. These pre-trained models have learned to extract significant characteristics from images and can be used to train a new model. For fine-tuning the model, we employed the partial layer-freezing technique, where the earlier layers of a pre-trained model are frozen while the last few layers are trained on new data. Two callbacks were used for modifying and regulating the training process. The first callback is ReduceLROnPlateau, which analyzes validation accuracy throughout training and reduces the learning rate if there is no progress in the number of epochs of patience. ModelCheckpoint is the second callback, which saves the model after each epoch if the validation loss has decreased. The five CNN models were trained and evaluated using accuracy, Cohen's Kappa, ROCAUC score, precision, and recall/sensitivity. The evaluation results of the models were measured and compared to the baseline performance.

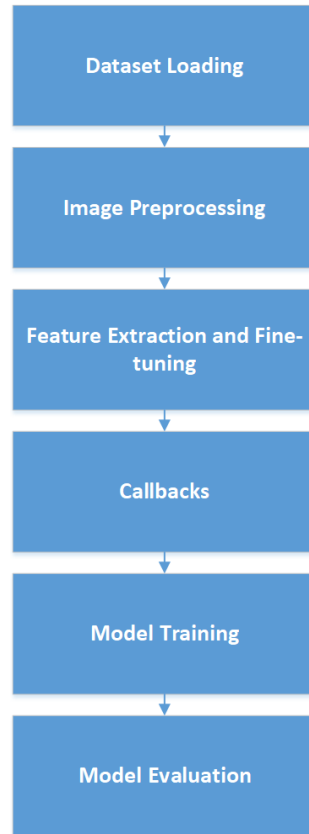


Figure 3. 1. Block diagram of the methodology

3.2 Dataset

The public musculoskeletal radiograph dataset contains 14,863 studies from 12,173 patients, totaling 40,561 multi-view radiographic images [4]. Each standard upper-extremity radiographic study type represents an elbow, finger, forearm, hand, humerus, shoulder, and wrist. Between 2001 and 2012, board-certified radiologists from Stanford Hospital manually labeled each study as normal or abnormal during clinical radiographic interpretation in the diagnostic radiology environment. We have used 36,808 images from the dataset for training (21935 normal images and 14873 abnormal images) and 3197 images for validation (1667 normal images, 1530 abnormal images).

Table 3. 1: Total number of studies in MURA dataset [4]

Study	Train		Validation		Total
	Normal	Abnormal	Normal	Abnormal	
Elbow	1094	660	92	66	1912
Finger	1280	655	92	83	2110
Hand	1497	521	101	66	2185
Humerus	321	271	68	67	727
Forearm	590	287	69	64	1010
Shoulder	1364	1457	99	95	3015
Wrist	2134	1326	140	97	3697
Total Studies	8280	5177	661	538	14656

3.3 Image Preprocessing

Image preprocessing is crucial for improving image quality and model accuracy for the classification task. As a preprocessing step, we have used CLAHE (Contrast Limited Adaptive Histogram Equalization) to enhance the contrast of images and Data Augmentation to reduce overfitting.

- CLAHE (Contrast Limited Adaptive Histogram Equalization) is a variation of AHE (Adaptive Histogram Equalization) that restricts image contrast over-amplification. It works on small regions known as tiles and then joins neighboring tiles using bilinear interpolation to erase the artificial borders. We have used this technique to improve the contrast of medical images, making them more suitable for classification and diagnosis.

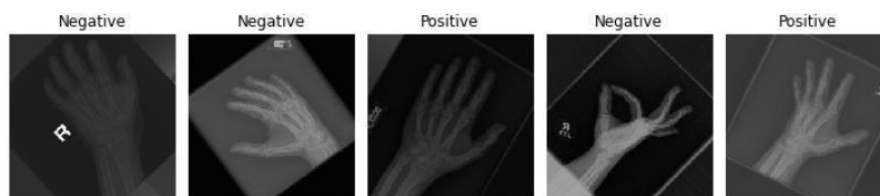


Figure 3. 2. Before preprocessing images using CLAHE

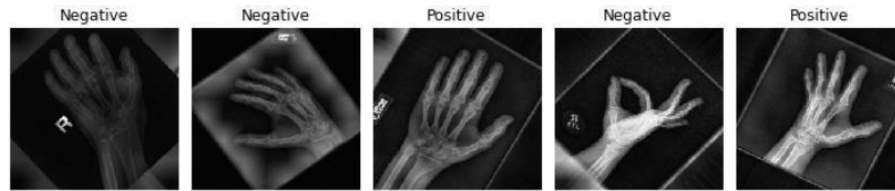


Figure 3.3. After preprocessing images using CLAHE

In our code, a function named "claheImage" takes an input image (Figure 3.2) and produces an output image that has been processed using the (CLAHE) approach (Figure 3.3).

- **Data Augmentation:** The dataset includes imbalanced studies on the training set, for example, only right elbow images. So that the model may learn high-level dataset properties that are invariant to typical affine transformations, we augmented the dataset. We used data augmentation techniques: rotation range: 45° , zoom range: 0.02, shear range: 0.02, and horizontal flip: True. Unlike other computer vision tasks, medical images have a well-known scale and defined procedures shallower the model to cope with potential lightning or further affine modification changes. As a result, much less information is required in medical radiography [13]. We only used four data augmentation strategies for the dataset as a result.

3.4 CNN and Transfer Learning

A common kind of neural network for evaluating visual images is the convolutional neural network (CNN). Convolutional, pooling, and fully connected layers are among the layers of CNN's architecture. The architecture of CNNs, with convolutional layers, pooling layers, and fully connected layers, is specifically designed to capture the unique characteristics of images and handle the large amount of data involved in image classification. Therefore, CNNs have been shown to work better on image classification tasks compared to traditional neural networks. Several research papers have demonstrated the effectiveness of CNNs on image classification tasks, including Imagenet [3]. In deep learning, transfer learning refers to using a pre-trained model on a large dataset as a starting point for solving a new problem. By leveraging the pre-trained model's learned representations, the new model can be trained on a smaller dataset, often with

better accuracy and efficiency. By fine-tuning the pre-trained model on the new data, the model can be adapted to the new dataset while retaining the learned feature representations. It can result in improved accuracy and reduced training time compared to training a new model from scratch. Transfer learning has become an essential tool in computer vision research, with many state-of-the-art models relying on transfer learning to achieve high-performance levels on benchmark datasets [14].

3.5 CNN Architectures

We used five state-of-the-art architectures to improve the accuracy compared to the baseline model of the MURA dataset.

3.5.1 ResNet

We use the ResNet152V2 architecture, which has 152 layers, in this paper. He et al. proposed ResNet as a deep convolutional neural network architecture for image classification applications [15]. It is an enhanced version of the ResNet152 model with reduced error rates and more excellent generalization performance. The ResNet design comprises a deep stack of residual blocks that, by overcoming the vanishing gradient problem, allow for the practical training of intense neural networks.

3.5.2 DenseNet

DenseNet architecture was introduced by Huang et al. [16]. DenseNet is a convolutional neural network (CNN) that connects each layer to every other layer in a feed-forward manner, unlike traditional CNNs, where each layer is connected only to its subsequent layer. The idea behind DenseNet is that by using dense connections, the network can better use the available information and gradients throughout the network, leading to better performance with fewer parameters. DenseNet comes in different variants, including DenseNet-121 and DenseNet-169, which differ in their number of layers and complexity. DenseNet-121 has 121 layers, while DenseNet-169 has 169 layers, and both have been shown to outperform other popular CNN architectures on various image classification tasks.

3.5.3 Inception

Szegedy et al. introduced the Inception model, a CNN architecture that assists with image processing and object recognition [17]. Inception-v3 is a variant of the Inception architecture that

uses Label Smoothing, Factorized 7×7 convolutions, and an auxiliary classifier to propagate label information lower down the network and batch normalization for layers in the sidehead. In addition, an Inception module that collects local characteristics and combines them into higher-level features is a crucial building element of Inceptionv3.

3.5.4 Xception

Chollet introduced the Xception architecture, where each Inception module is replaced in this model with a depthwise separable convolution block that performs depthwise (spatial filtering) and pointwise (cross-channel filtering) convolutions independently [18]. It helps minimize the model's computational cost and memory utilization while boosting accuracy. It also uses skip connections to aid gradient propagation and avoid the vanishing gradient problem. On various picture classification benchmarks, including ImageNet and CIFAR-10, Xception outperforms earlier state-of-the-art models.

3.6 Fine-tuning

Fine-tuning is a strategy that includes training a pre-trained model on a new dataset to improve its accuracy for a particular task. We employed five different transfer learning models to predict the outcome of the MURA dataset. To increase accuracy and prevent overfitting, we adopted callbacks to automate the fine-tuning and a layer-freezing strategy, in which the acquired weights of individual layers in a pre-trained neural network are held constant while fine-tuning on a new dataset.

3.6.1 Callbacks

The first callback is 'ReduceLROnPlateau,' which checks validation accuracy ('val_accuracy') throughout training and decreases the learning rate if there is no progress for the 'patience' number of epochs. It does so by increasing the learning rate by a 'factor' value (0.5 in this paper) and enforcing a lower constraint on the learning rate of 'min_lr' ($1e-5$ in this case). This callback prevents the model from becoming trapped in a suboptimal local minimum. The 'verbose' option is set to 1, which implies that when the learning rate is lowered, the callback will output a message to the console. 'ModelCheckpoint' is the second callback, which saves the model after each epoch

if the validation loss has decreased. Again, the verbose option is set to 1, which implies that when a new best model is saved, the callback will output a message to the console.

3.6.2 Layer-freezing

In this paper, the last ten layers of the five pre-trained models are trainable, while the preceding levels are frozen. This method is frequently employed when the pre-trained model is comparable to the task at hand but needs some tweaking to accommodate the new dataset. While the frozen layers extract data, the trainable layers learn task-specific characteristics.

Activation	Post Batch Normalization	Batch Normalization	Block SepConv Batch Normalization
Mixed10 (concatenate)	Post ReLu (activation)	ReLu (activation)	Block SepConv Act (activation)
Global_average_pooling2D	Global_average_pooling2D	Global_average_pooling2D	Global_average_pooling2D
Batch Normalization	Batch Normalization	Batch Normalization	Batch Normalization
Dropout	Dropout	Dropout	Dropout
Dense	Dense	Dense	Dense
Dense	Dense	Dense	Dense
Batch Normalization	Batch Normalization	Batch Normalization	Batch Normalization
Dropout	Dropout	Dropout	Dropout
Dense	Dense	Dense	Dense

a.DenseNet

b.Inception

c.ResNet

d.Xception

Figure 3. 4. Trainable layers in the CNN models

3.7 Evaluation Metrics

The following metrics were used to assess the standard of the resulting models and compare them to the one proposed by Rajpurkar et al. while presenting the MURA dataset:

3.7.1 Accuracy

Accuracy is one of the commonly used evaluation metrics for CNN models. It measures the proportion of correctly predicted instances among all instances. For binary classification problems, accuracy can also be calculated in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

3.7.2 Cohen's Kappa

Cohen's Kappa is a metric that measures the agreement between two raters or between a rater and a classification model. It considers the possibility of random agreement, providing a more robust measure than simple percent agreement. It is commonly used in cases where the classes are imbalanced. The formula for Cohen's Kappa is:

$$K = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

Where P_o is the observed agreement, and P_e is the expected agreement.

3.7.3 ROCAUC Score

ROCAUC (Receiver Operating Characteristic Area Under the Curve) evaluates a binary classification model's performance across various classification thresholds. The area under the ROC curve, which compares the true positive rate (TPR) against the false positive rate (FPR) for various classification thresholds, is used to compute it.

3.7.4 Precision

The fraction of genuine positives among all positive classifications produced by the model is measured as precision. The precision formula is as follows:

$$Precision = \frac{TP}{TP + P}$$

(3)

Precision is an essential parameter for models with a high cost of false positives. For example, a false positive might lead to needless and sometimes hazardous therapies in medical diagnostics. A high-accuracy model is preferable in such instances.

3.7.5 Recall/Sensitivity

In evaluation metrics for CNN models, recall (also known as sensitivity) is a metric that measures the proportion of actual positive samples that the model correctly identifies. The formula for the recall is:

$$Recall = \frac{TP}{TP + \bar{N}}$$

(4)

Recall measures how well the model can detect positive samples.

Chapter 4 - Result Analysis and Discussion

4.1 Discussion

Except for the learning rate, all hyper-parameters remained constant throughout the studies. The models were fine-tuned properly, and the batch size was 12. In all of the tests, the Adam optimizer was employed. All of the images have been scaled to 224 x 224 pixels. Because the dataset is binarily categorized, binary cross-entropy was used as the loss function, and the models were trained for 50 epochs. As the dataset is already divided into train and validation, there was no train-test split in this experiment. Following training, the performance of each model was assessed using the evaluation metrics given in the method section and compared to the baseline model.

4.2 Wrist Dataset Result

In the wrist dataset, there were 9752 images for training (5765 normal images, 3986 abnormal images) and 659 images for testing (364 normal images, 295 abnormal images). DenseNet169 performed the best of the models, with DenseNet121 having better precision and ResNet152V2 having higher recall than DenseNet169. Nonetheless, none of the evaluated models outperformed the baseline proposed by Rajpurkar et al.

Table 4. 1: Results of wrist dataset

Model	Accuracy	Kappa	ROCAUC	Precision	Recall/sensitivity
Baseline		0.931			
ResNet152V2	76%[0.494]	0.505	0.75	0.77	0.65
DenseNet121	76%[0.502]	0.5	0.74	0.88	0.55
InceptionV3	74%[0.525]	0.461	0.72	0.82	0.54
Xception	74%[0.53]	0.465	0.73	0.75	0.63
DenseNet169	77%[0.457]	0.526	0.76	0.82	0.62

4.3 Elbow Dataset Result

In the elbow dataset, there were 4931 images for training (2925 normal images, 2006 abnormal images) and 465 images for testing (235 normal images, 230 abnormal images). In this model, ResNet152V2 and DenseNet169 performed well on all the metrics, where DenseNet169 had the best precision, and ResNet152V2 had the best sensitivity. However, the models underperformed compared to the baseline kappa score.

Table 4. 2: Results of elbow dataset

Model	Accuracy	Kappa	ROCAUC	Precision	Recall/sensitivity
Baseline		0.71			
ResNet152V2	77% [0.57]	0.53	0.76	0.82	0.67
DenseNet121	74%[0.553]	0.469	0.73	0.9	0.52
InceptionV3	73%[0.533]	0.465	0.73	0.85	0.56
Xception	75%[0.498]	0.495	0.75	0.83	0.61
DenseNet169	77% [0.507]	0.529	0.76	0.92	0.58

4.4 Finger Dataset Result

In the finger dataset, there were 5106 images for training (3138 normal images, 1968 abnormal images) and 461 images for testing (214 normal images, 247 abnormal images). In this dataset, InceptionV3 and DenseNet169 models performed best in all metrics and outperformed the kappa score of the baseline model.

Table 4. 3: Results of finger dataset

Model	Accuracy	Kappa	ROCAUC	Precision	Recall/sensitivity
Baseline		0.389			
ResNet152V2	63%[0.626]	0.279	0.64	0.75	0.47
DenseNet121	66%[0.67]	0.343	0.68	0.79	0.5
InceptionV3	71% [0.589]	0.426	0.72	0.8	0.62
Xception	69%[0.630]	0.38	0.69	0.78	0.59
DenseNet169	72% [0.591]	0.439	0.72	0.81	0.62

4.5 Forearm Dataset Result

In the forearm dataset, there were 1825 images for training (1164 normal images, 661 abnormal images) and 301 images for testing (150 normal images, 151 abnormal images). The denseNet121 model performed better than others and scored best in all assessments.

Table 4. 4: Results of forearm dataset

Model	Accuracy	Kappa	ROCAUC	Precision	Recall/sensitivity
Baseline		0.737			
ResNet152V2	74%[0.554]	0.476	0.74	0.82	0.61
DenseNet121	76%[0.513]	0.522	0.76	0.86	0.63
InceptionV3	71%[0.616]	0.416	0.71	0.86	0.5
Xception	73%[0.583]	0.456	0.73	0.82	0.58
DenseNet169	74%[0.561]	0.476	0.74	0.84	0.59

4.6 Shoulder Dataset Result

In the shoulder dataset, there were 8379 images for training (4211 normal images, 4168 abnormal images) and 563 images for testing (285 normal images, 278 abnormal images). The CNN models performed poorly compared to the baseline model in the shoulder dataset. The best kappa score was 0.449, achieved by InceptionV3. However, the recall was better than the wrist, elbow, finger, and forearm datasets because the shoulder dataset was quite balanced with normal and abnormal images.

Table 4. 5: Results of shoulder dataset

Model	Accuracy	Kappa	ROCAUC	Precision	Recall/sensitivity
Baseline		0.729			
ResNet152V2	70%[0.57]	0.407	0.7	0.69	0.72
DenseNet121	73%[0.535]	0.46	0.73	0.72	0.75
InceptionV3	72%[0.543]	0.449	0.72	0.74	0.69
Xception	68%[0.574]	0.354	0.68	0.65	0.73
DenseNet169	69%[[0.576]	0.359	0.68	0.64	0.79

4.7 Hand Dataset Result

There were 5543 images for training (4059 normal images, 1484 abnormal images) and 460 images for testing (271 normal images, 189 abnormal images) in the hand dataset. In comparison to the baseline, the models performed poorly. Despite the models' moderate precision, the kappa score and recall were deficient. However, the ResNet152V2 model outperformed the other models.

Table 4. 6: Results of hand dataset

Model	Accuracy	Kappa	ROCAUC	Precision	Recall/sensitivity
Baseline		0.851			
ResNet152V2	70%[0.606]	0.335	0.65	0.8	0.37
DenseNet121	63%[0.763]	0.134	0.56	0.73	0.16
InceptionV3	66%[0.63]	0.224	0.6	0.78	0.25
Xception	68%[0.629]	0.281	0.63	0.77	0.32
DenseNet169	65%[0.668]	0.201	0.59	0.77	0.23

4.8 Humerus Dataset Result

The humerus dataset had 1272 images for training (673 normal images and 599 abnormal images) and 288 images for testing (148 normal images and 140 abnormal images). The models performed well in this dataset, outperforming the baseline model's kappa score. DenseNet121, InceptionV3, and DenseNet169 outscored the baseline model's kappa score. The DenseNet121 model performed the best of the three, while the Xception model performed the best in precision.

Table 4. 7: Results of humerus dataset

Model	Accuracy	Kappa	ROCAUC	Precision	Recall/sensitivity
Baseline		0.6			
ResNet152V2	76%[0.509]	0.519	0.76	0.78	0.71
DenseNet121	82%[0.430]	0.638	0.82	0.77	0.84
InceptionV3	81%[0.472]	0.611	0.81	0.8	0.8
Xception	79%[0.49]	0.574	0.79	0.84	0.69
DenseNet169	81%[0.458]	0.611	0.81	0.79	0.81

4.9 Overall Dataset Result

The MURA dataset had 36808 images for training (21935 normal images and 14873 abnormal images) and 3197 images for testing (1667 normal images and 1530 abnormal images). According to the evaluation metrics, the DenseNet169 model performed better than other models. However, the kappa score and sensitivity of the models were inferior to the reference model.

Table 4. 8: Results of overall dataset

Model	Accuracy	Kappa	ROCAUC	Precision	Recall/sensitivity
Baseline		0.705			0.815
ResNet152V2	73%[0.543]	0.447	0.72	0.8	0.57
DenseNet121	73%[0.543]	0.459	0.73	0.84	0.55
InceptionV3	70%[0.57]	0.395	0.69	0.79	0.52
Xception	71%[0.565]	0.413	0.7	0.77	0.56
DenseNet169	74%[0.536]	0.466	0.73	0.82	0.57

From the results, finger and humerus images Cohen Kappa scores outperformed the baseline model's Cohen Kappa score. In the finger studies, both InceptionV3 and DenseNet169 models performed significantly well. In the humerus studies, DenseNet121, InceptionV3, and DenseNet169 models outperformed the baseline model. However, DenseNet121 performed better than the other two models.

Chapter 5 - Conclusion

5.1 Conclusion

After analyzing the results of all the studies, we conclude that the partial layer-freezing technique improved the performance of the finger and humerus studies of the MURA dataset but did not outperform the baseline model on wrist, elbow, forearm, hand, and shoulder analyses. This could be because the pre-trained model used for transfer learning was developed on a dataset (Imagenet) entirely dissimilar to the target domain (medical images), and freezing the layers might make it challenging to capture the characteristics required for the new task adequately. However, research shows that freezing earlier layers can be advantageous since they frequently become well-trained earlier [19]. The concept is supported by the fact that we froze the previous layers and trained only the last ten layers, which improved the performance of finger and humerus studies. Furthermore, the DenseNet network showed the potential to perform well on the MURA dataset, whereas the Xception network did not achieve the expected result in any of the studies. In conclusion, while the five most advanced models discussed in this paper did not perform at the same level as the baseline, except for the finger and humerus studies, they may be crucial in providing a second view and prioritizing emergency tasks.

Chapter 6 - Limitation And Future Scope

6.1 Limitations

One of the limitations of our study is that the effect of hyperparameters was not analyzed. Our model could perform better if we focused on fine-tuning with hyperparameters. However, automated detection of musculoskeletal images in emergencies will help radiologists decide as early as possible.

6.2 Future Scope

In this scenario, the layer-freezing technique needs further research to understand its capability by tweaking parameters according to the CNN models. Furthermore, this approach works well on small datasets, which is noteworthy given the scarcity of data in the medical field.

Finally, we might create an ensemble model using the five trained CNN models. We would train a classifier so that each model could learn particular features to differentiate between normal and abnormal images.

References

- [1] A. M. Briggs *et al.*, "Reducing the global burden of musculoskeletal conditions," *Bull. World Health Organ.*, vol. 96, no. 5, pp. 366–368, May 2018, doi: 10.2471/BLT.17.204891.
- [2] M. Al Shammari, A. Hassan, O. Al Dandan, M. Al Gadeeb, and D. Bubshait, "Musculoskeletal symptoms among radiologists in Saudi Arabia: a multi-center cross-sectional study," *BMC Musculoskelet. Disord.*, vol. 20, no. 1, p. 541, Nov. 2019, doi: 10.1186/s12891-019-2933-1.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
- [4] P. Rajpurkar *et al.*, "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs." arXiv, May 22, 2018. doi: 10.48550/arXiv.1712.06957.
- [5] A. F. M. Saif, C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "Abnormality Detection in Musculoskeletal Radiographs Using Capsule Network," *IEEE Access*, vol. 7, pp. 81494–81503, 2019, doi: 10.1109/ACCESS.2019.2923008.
- [6] D. Banga and P. Waiganjo, "Abnormality Detection in Musculoskeletal Radiographs with Convolutional Neural Networks(Ensembles) and Performance Optimization." arXiv, Aug. 06, 2019. doi: 10.48550/arXiv.1908.02170.
- [7] S. Tirpude, "Abnormal X-Ray Detection System using Convolution Neural Network," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, pp. 828–832, Feb. 2020, doi: 10.30534/ijatcse/2020/119912020.
- [8] I. Kandel, M. Castelli, and A. Popovič, "Musculoskeletal Images Classification for Detection of Fractures Using Transfer Learning," *J. Imaging*, vol. 6, no. 11, Art. no. 11, Nov. 2020, doi: 10.3390/jimaging6110127.
- [9] F. Uysal, F. Hardalaç, O. Peker, T. Tolunay, and N. Tokgöz, "Classification of Shoulder X-ray Images with Deep Learning Ensemble Models," *Appl. Sci.*, vol. 11, no. 6, Art. no. 6, Jan. 2021, doi: 10.3390/app11062723.
- [10] I. Kandel and M. Castelli, "Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset," *Health Inf. Sci. Syst.*, vol. 9, no. 1, p. 33, Jul. 2021, doi: 10.1007/s13755-021-00163-7.
- [11] G. Singh, D. Anand, W. Cho, G. P. Joshi, and K. C. Son, "Hybrid Deep Learning Approach for Automatic Detection in Musculoskeletal Radiographs," *Biology*, vol. 11, no. 5, Art. no. 5, May 2022, doi: 10.3390/biology11050665.
- [12] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1717–1724. doi: 10.1109/CVPR.2014.222.
- [13] B. J. Erickson, P. Korfiatis, T. L. Kline, Z. Akkus, K. Philbrick, and A. D. Weston, "Deep Learning in Radiology: Does One Size Fit All?," *J. Am. Coll. Radiol.*, vol. 15, no. 3, pp. 521–526, Mar. 2018, doi: 10.1016/j.jacr.2017.12.027.
- [14] R. Ribani and M. Marengoni, *A Survey of Transfer Learning for Convolutional Neural Networks*. 2019, p. 57. doi: 10.1109/SIBGRAP-T.2019.00010.

- [15]K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [16]G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks." arXiv, Jan. 28, 2018. doi: 10.48550/arXiv.1608.06993.
- [17]C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.
- [18]"Xception: Deep Learning with Depthwise Separable Convolutions."
<https://www.computer.org/csdl/proceedings-article/cvpr/2017/0457b800/12OmNqFJhzG>
(accessed May 11, 2023).
- [19]Y. Wang, D. Sun, K. Chen, F. Lai, and M. Chowdhury, "Egeria: Efficient DNN Training with Knowledge-Guided Layer Freezing." arXiv, Mar. 11, 2023. doi: 10.48550/arXiv.2201.06227.

Plagiarism Report

7/23/23, 9:20 AM

Tumitin - Originality Report - 192-16-452

Turnitin Originality Report	
Processed on: 23-Jul-2023 09:20 +06 ID: 2135185569 Word Count: 7088 Submitted: 1	
192-16-452 By Annur Anika	
Similarity Index 21%	Similarity by Source Internet Sources: 18% Publications: 11% Student Papers: 9%

2% match (Internet from 26-Oct-2022) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8436/181-16-255.pdf?isAllowed=y&sequence=1
2% match (student papers from 31-Mar-2019) Submitted to Daffodil International University on 2019-03-31
1% match (Internet from 21-Jul-2023) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/10006/22637.pdf?isAllowed=y&sequence=1
1% match (Internet from 21-Nov-2022) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/6844/172-15-9900%20%2815%25%29%20clearance.pdf?isAllowed=y&sequence=1
1% match (Internet from 29-Jun-2023) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/9380/21862.pdf?isAllowed=y&sequence=1
1% match (Internet from 15-Oct-2022) https://www.mdpi.com/2076-3417/11/6/2723/htm
1% match (Internet from 22-Sep-2022) https://www.researchgate.net/publication/344878290_Automating_Abnormality_Detection_in_Musculoskeletal_Radiographs_through_Deep_Learning
1% match (Internet from 22-Sep-2022) https://www.researchgate.net/publication/347114011_Musculoskeletal_Images_Classification_for_Detection_of_Fractures_Using_Transfer_Learning
1% match (student papers from 10-Dec-2021) Submitted to Liverpool John Moores University on 2021-12-10
1% match () Kandel, Ibrahim Hamdy Abdelhamid. "Deep Learning Techniques for Medical Image Classification", 2021
1% match (Internet from 05-Nov-2022) https://arxiv.org/pdf/1712.06957.pdf
1% match (Internet from 12-Jan-2023) https://mdpi-res.com/d_attachment/biology/biology-11-00665/article_deploy/biology-11-00665.pdf?vversion=1650972381
1% match (Internet from 31-Oct-2019) https://repositori.upf.edu/bitstream/handle/10230/42540/De_Abreu_2019.pdf?isAllowed=y&sequence=1
< 1% match (Internet from 25-Oct-2022) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5447/192-25-789%20%2825%25%29.pdf?isAllowed=y&sequence=1
< 1% match (Internet from 29-Jun-2023) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/9279/21602.pdf?isAllowed=y&sequence=1
< 1% match (Internet from 29-Jun-2023) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/9790/22506.pdf?isAllowed=y&sequence=1
< 1% match (student papers from 16-May-2023) Submitted to Liverpool John Moores University on 2023-05-16
< 1% match (student papers from 29-Mar-2023) Submitted to Liverpool John Moores University on 2023-03-29
< 1% match (student papers from 01-Jun-2020) Submitted to Liverpool John Moores University on 2020-06-01
< 1% match () Kandel, Ibrahim, Castelli, Mauro. "a case study using MURA dataset", Health Information Science and Systems, 2021
< 1% match (Internet from 05-Feb-2023) https://arxiv.org/pdf/2102.13233v1.pdf
< 1% match (Internet from 08-Oct-2022) https://www.warse.org/IJATCSE/static/pdf/file/ijatcse119912020.pdf
< 1% match (Internet from 20-Oct-2022)

https://www.tumitin.com/newreport_printview.asp?eq=1&eb=1&esm=10&oid=2135185569&sid=0&n=0&m=2&svr=29&r=33.408713523223184&lang=e... 1/8