Bone X-ray abnormality detection using MURA dataset

Author

Sana Batool

Regn Number

00000317633

Supervisor

Dr. Syed Omer Gilani

DEPARTMENT BIOMEDICAL ENGINEERING AND SCIENCES

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

JANUARY 2023

Bone X-ray abnormality detection using MURA dataset

Author

Sana Batool

Regn Number

00000317633

A thesis submitted in partial fulfillment of the requirements for the degree of

MS BIOMEDICAL SCIENCES

Thesis Supervisor:

Dr. Syed Omer Gilani

Thesis Supervisor's Signature: _____

DEPARTMENT BIOMEDICAL ENGINEERING AND SCIENCES

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

JANUARY 2023

Declaration

I certify that this research work titled "*Bone X-ray abnormality detection using MURA dataset"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged / referred.

Signature of Student

Sana Batool

00000317633

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms. **Sana Batool** (Registration No. **00000317633**), of **School of Mechanical and Manufacturing Engineering** has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is within the similarity indices limit and is accepted as partial fulfillment for the award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: Dr. Syed Omer Gilani

Date: _____

Signature (HoD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Certificate for Plagiarism

It is certified that MS Thesis Titled **Bone X-ray abnormality detection using MURA dataset** by **Sana Batool** has been examined by us. We undertake the follows:

a. Thesis has significant new work/knowledge as compared to already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.
b. The work presented is original and own work of the author (i.e., there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.
c. There is no fabrication of data or results which have been compiled/analyzed.
d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name & Signature of Supervisor

Dr. Syed Omer Gilani

Signature**:**_____

Copyright Statement

Acknowledgements

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought which You setup in my mind to improve it. Indeed, I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual, was Your will, so indeed none be worthy of praise but You.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout every department of my life.

I would also like to express special thanks to my supervisor Dr. Syed Omer Gilani for his help throughout my thesis and also for the Medical Image Analysis course which he has taught me. I can safely say that I haven't learned any other engineering subject in such depth than the one which he has taught. I would also like to pay special thanks for his tremendous support and cooperation. Each time I got stuck in something; he came up with the solution. Without his help I wouldn’t have been able to complete my thesis. I appreciate his patience and guidance throughout the whole thesis.

I would also like to thank Dr. Asim Waris and Dr. Aneeqa Noor for being on my thesis guidance and evaluation committee.

Finally, I would like to express my gratitude to my Classmate and friend Ramsha Abbasi who has rendered valuable assistance to my study.

Dedicated to my exceptional parents and adored siblings whose tremendous support and cooperation led me to this wonderful accomplishment.

Abstract

Musculoskeletal abnormalities along with bone fractures are a wide range of abnormalities that account for most visits of patients to Emergency department of hospitals. According to an estimate, more than 1.7 billion people are affected by musculoskeletal disorders each year. Bone X-rays are the first line imaging modality for imaging of fractured bones. Radiologists then undergo reporting of X-rays for detection of fractures and pathologies. Classification of bone X-rays into normal and abnormal is a time-taking process and is also subjected to variability between different radiologists. Therefore, the use of automatic classifiers incorporating deep learning algorithms is currently in use in clinical diagnostics. MURA is a large publicly available dataset released by the machine learning group of Stanford university. MURA dataset consists of 40,895 multi-view images of upper limb that belong to seven regions namely shoulder, humerus, elbow, forearm, wrist, hand, and fingers. In this study we propose the use of the single DenseNet-169 model trained on complete dataset along with multiple pre-processing and data augmentation steps, based on Keras in TensorFlow. Training data was divided into 80:20 for training and validation respectively, whereas, testing of model was done on validation set. The results obtained through the proposed technique include 80% testing accuracy. This validates the effectiveness of this method for bone fractures classification.

Keywords: Image classification, Deep learning, Deep Neural Networks, MURA, Bone X-rays.

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

1.1 Musculoskeletal Abnormalities

Musculoskeletal abnormalities are a wide range of abnormalities that not only affect bones but also muscles, tendons, and ligaments. These abnormalities result from trauma, pathological disease, or degenerative changes in the body. These are the abnormalities of bones and joints that are known to cause pain as well as restrict patient's motion. These abnormalities tend to affect people's normal routine thus affecting work efficacy(Walker-Bone et al., 2004). Therefore, correct, and early diagnosis can not only restore a patient's normal healthy routine but also increase the efficacy of work. Musculoskeletal disorder affects the muscles, bones, ligaments, tendons, and nerves etc. Examples of musculoskeletal abnormalities include, Fracture, joint displacement, tendonitis, arthritis, aging process to name a few. Patients with bone abnormalities visit orthopedicians and radiologists who use Xray images of the affected bones to detect the abnormality.

Figure 1: Illustration of musculoskeletal disorders adopted from Domenico Albano (IRCCS Istituto Ortopedico Galeazzi) & Francesco Carrubi (University of L'Aquila: online)

1.2 Incidence of Musculoskeletal disorders

According to Journal of Pakistan Medical Association (JPMA), Pakistan has 75.8% prevalence of musculoskeletal disorders (MSD). They identified lack of rest and maintaining a difficult stature as the foremost causes of these disorders(Hameed et al., 2016) (Haroon et al., 2018). According to an estimate greater than 1.7 billion people are affected by musculoskeletal disorders each year(Solovyova & Solovyov, 2020), thus resulting in increased visits to emergency departments of hospitals and making work overload for radiologists. This increased patient load on radiologists, lack of experienced radiologists, and complicated minute deformities like hairline fractures make it difficult to report and diagnose them correctly. That outcomes in greater chances of incorrect diagnosis and overlooked abnormalities. (Fernholm et al., 2019) reported that out of all the abnormalities misdiagnosed in imaging procedures, fractures accounted for 24% the most common of which belonged to wrist and fingers, 29% (Fernholm et al., 2019). To overcome this issue artificial intelligence is incorporated to classify bone x-rays as normal or abnormal images.

1.3 Imaging modalities used to assess Musculoskeletal disorders

Different imaging modalities are used to detect bony abnormalities named as, Xray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Focused assessment with sonography in trauma (commonly abbreviated as FAST). Xray is a 2-D modality that has ionizing radiations, it is used mostly for imaging of bones. Moreover, Xray is used as a first line imaging modality in case of any musculoskeletal trauma(Al-Ayyoub et al., 2013). CT-scan is a 3-D modality used to image muscles, bones, and joints from three dimensional slices. MRI is also a 3-D imaging modality that images not only bones but also soft tissues, muscles, tendons, and ligaments. MRI is referred to as a soft tissue imaging modality as it images soft tissues better than all other modalities.

1.4 Role of deep learning in Diagnosis and Treatment of MSD

Machine learning (ML) particularly deep learning (DL) has a significant role in medical diagnosis-based problem solving. Classification is one of the problem-solving strategies. We divide classification into two types. One is binary classification, in which the given dataset is divided into two groups of normal (without abnormality) and abnormal (with abnormality)(Ishida et al., 2018). Whereas multi-class classification is a technique that divides a database into multiple sub-classes. Python is a programming language that has multiple libraries. TensorFlow is one of its basic libraries that uses Keras as an interface to give promising results in the field of medicine. This automated classification of radiographs would help in reporting prioritization, reduce radiologist's patient load, and would significantly reduce the chance of missed and incorrect diagnoses.

1.5 Research Objective

The objective of this research was the development of reliable automatic algorithm for accurate classification of musculoskeletal radiographs as normal or abnormal from MURA dataset using deep learning techniques. Therefore, we followed a relatively simple approach without any architectural variations and focused more on the training workflow and achieved comparable results.

CHAPTER 2: LITERATURE REVIEW

Limited classification studies were present on complete upper limb radiographs until the Machine learning group of Stanford university first released one of the largest publicly available datasets. This dataset was collected from the duration of 2001-2012. The radiographic images contained in the test set were labelled by a team of six radiologists having 8.83 years average experience. Many emerging methods have been proposed and used to solve different classification problems. (Rajpurkar et al., 2017) used mainly DenseNet-169 and ensemble of 5 models to classify this dataset into positive and negative groups. They calculated testing accuracy on the ensemble-based model and Cohen's kappa statistic was also used to compare model accuracy with that of certified radiologist's team. The best kappa score was obtained on wrist and finger studies. In addition to this, they also incorporated Class activation mappings (CAM) to localize the position of abnormalities such as fractures.

(Solovyova & Solovyov, 2020) used the same CNN model by adding preprocessing to improve image quality of dataset, one of the main image preprocessing techniques incorporated was cropping region of interest using threshold value from important part of images. Different types of data augmentation were used to enlarge the size of data, they trained certain epochs with freeze encoder and rested while unfreezing the system. They used ensemble of 4 DenseNet 169 models to improve mean accuracy. Kappa score calculated by (Solovyova & Solovyov, 2020) was better than the previous studies on wrist, hand and shoulder region of the datasets.

Data augmentation is a technique usually applied in training phase, but (Kandel & Castelli, 2021) incorporated this technique during prediction time and referred it as test time augmentation (TTA). TTA uses images of different varieties by incorporating 9 different augmentation techniques and calculates predictions on these images. In the end these predictions are averaged to calculate average score or majority of vote. They implemented this technique on each of VGG19, InceptionV3, ResNet50, Xception, and DenseNet121 models. They concluded that TTA can significantly increase performance of models with low accuracy. In the end they made ensembles using these models and thus selected the best of two ensembles based on final

predictions. In addition to this, kappa score was also calculated on these results. Which showed better results than previous studies.

With the advancement of ML technology, many researchers started making use of ensemble of different better performing classifiers instead of using single deep learning model. (He et al., 2021) also used one such customized ensemble and called it as calibrated ensemble model. In CE model they incorporated ConvNet, ResNet and DenseNet DL models. First, they used these models individually and then ensembled them. This new idea of CE model was based on the fact that this ensemble was created with individual models that performed better on that region of anatomy. As in case of ConvNet, it gave better results in the Elbow and forearm region, whereas ResNet outperformed in humerus and shoulder regions. Another novelty found in this research was the use of customized loss function termed as "cross-entropy loss function". This customized loss function calculated weighted loss for each of the individual region-wise sub-dataset. These weighted loss functions were also used by (de La Torre et al., 2018) and are best suitable for the datasets with class imbalance.

It has been noticed from the work of (Rajpurkar et al., 2017) that the kappa agreement had the least score on finger and humerus parts of the dataset. Many researchers like (Chada, 2019), (Ghosh et al., 2021) and (Uysal et al., 2021) used only one to two regions, instead of using the complete dataset. They selected the regions in which accuracy was low and trained different models to check improvement in accuracy. (Chada, 2019) trained three deep learning-based models namely, DenseNet-169, DenseNet-201, and Inception-ResnetV2. Out of these three models, DenseNet-201 better classified humerus X-rays and showed improved kappa score than (Rajpurkar et al., 2017). Whereas (Uysal et al., 2021) used shoulder X-rays only from the MURA dataset. As (Uysal et al., 2021) stated that shoulder region because of offering a great range of motion is subjected to a great wear and tear thus gets easily fractured and dislocated. To classify normal shoulder X-rays from abnormal fractured ones they used 26 DL based pre-trained models. Out of these 26 DL models they made two ensembles namely, ensemble-1 (EL-1) and ensemble-2 (EL-2). The results of training showed that the highest AUC was achieved by EL1 whereas the highest kappa score and testing accuracy was achieved with EL2.

Further improvements in models used 10 hidden layers with adaBoost framework and trained this model for Humerus X-rays only. Their results showed that by adding 10 hidden layers to adaboost framework not only reduced the training time significantly but also it improved the kohen's cappa score and validation error better than (Rajpurkar et al., 2017). This adaBoost framework improves and accelerates the speed of model training as explained by (Freund et al., 1999), They explained that this boosting algorithm didn't even suffer overfitting problem and can also identify outliers. Thus, boosting training time and accuracy as well.

Most of the researchers used python as a training language and keras as an API on its library. Newer studies used FastAI as an API on Pytorch library. In a similar study, (Hooda & Shravankumar Bachu, 2020) worked on MURA dataset using FastAI. They trained the dataset using DenseNet-169 model twice by varying the size of input data. Once they had given the model input size of 112*112 and then used the freeze, unfreeze mode. Later they used input size of 320*320 and noted that best accuracy was achieved on Humerus region with 320*320 image size.

Table 1: Overview of literature

Author	Model/ Approach	Score
(Rajpurkar et al., 2017)	169-layered DenseNet baseline model Used ensemble of 5 DenseNet-169 models	AUROC of 0.929, 0.815 sensitivity 0.887 specificity
(Solovyova & Solovyov, 2020)	Multiple preprocessing & ensemble of 4 169-layered DenseNet models	AUC=0.870, Accuracy= 0.863
(He et al., 2021)	EL1= (ConvNet+ResNet+DenseNet) EL2= (Meta learner+ Res-DenseNet)	AUC= 0.93, Acc = 0.87, K= 0.74
(Hooda & Shravankumar Bachu, 2020)	DenseNet 169, trained model twice by varying image sizes. (112*112) freeze & unfreeze (320*320)	Accuracy With (112*112) = 0.818, with (320*320) = 0.835
(Madan et al., 2021)	Trained DenseNet-169 only on Humerus region with multiple pre-processing steps	Accuracy: 0.840 K= 0.68
(Kandel & Castelli, 2021)	Test-time augmentation	Improved results

CHAPTER 3: METHODOLOGY

Our work is based on Keras which is user friendly API based on TensorFlow framework. Python is the programming language used in this study. Keras based on TensorFlow is used in medical image analysis.

3.1 Dataset

MURA dataset

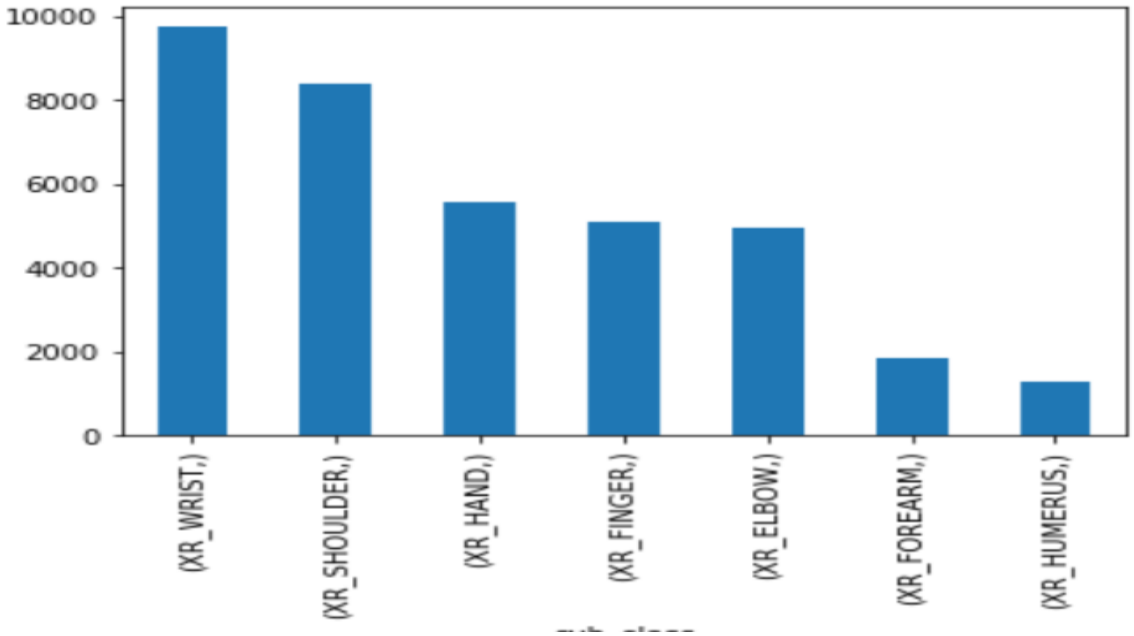
MURA large publicly available dataset was first introduced by Stanford machine learning group. The medicine unit of Stanford collected this dataset. Each radiographic study belonged to a specific patient containing more than one view/image is either normal or abnormal. This classification problem aims at splitting these radiographs into positive ~1 or negative ~0 classes. These radiographs were chosen from seven regions of upper limb including shoulder, humerus, elbow, forearm, wrist, hand, and fingers, respectively. MURA dataset consists of 40,895 multi-view images of upper limb belonging to one of those seven regions(Kandel & Castelli, 2021). These 14892 studies collected from 12,251 patients are all in 8-bit png format. This dataset is further divided into training N= 36808 and valid N= 3197 folders, respectively. Positive studies comprise of different abnormalities including Fracture, hardware placement, sclerotic lesions, and degenerative changes (Rajpurkar et al., 2017). Whereas negative studies are the normal x-rays without having any abnormality. All the images have standard format of 8-bit png. The size of these images varied significantly, width ranging from 81 to 512, and height from 132 to 512 size (Rajpurkar et al., 2017). All images belonged to a standard 8-bit png format.

On the following page, figure 2 shows the examples of X-ray images from MURA dataset.

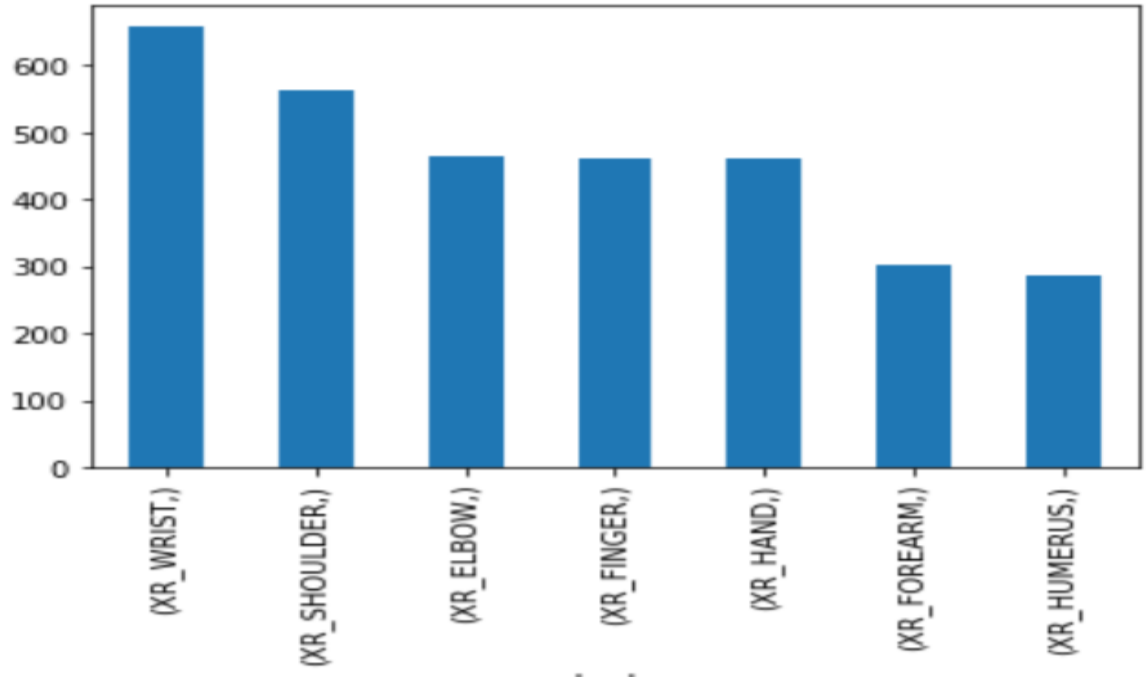
And table 2 shows the division of MURA dataset into train and valid folders first. Also the number of images in each of the seven upper limb regions is shown.

Figure 2: Examples of X-ray images from dataset

Study type	Label	Training set		Validation set	
		Studies	Images	Studies	Images
Elbow	Normal	1094	2925	92	235
	Abnormal	660	2006	66	230
Finger	Normal	1280	3138	92	214
	Abnormal	655	1968	83	247
Forearm	Normal	590	1164	69	150
	Abnormal	287	661	64	151
Hand	Normal	1497	4059	101	271
	Abnormal	521	1484	66	189
Humerus	Normal	321	673	68	148
	Abnormal	271	599	67	140
Shoulder	Normal	1364	4211	99	285
	Abnormal	1457	4168	95	278
Wrist	Normal	2134	5765	140	364
	Abnormal	1326	3987	97	295
All types	Normal	8280	21935	661	1667
	Abnormal	5177	14873	538	1530
	Total	13457	36808	1199	3197

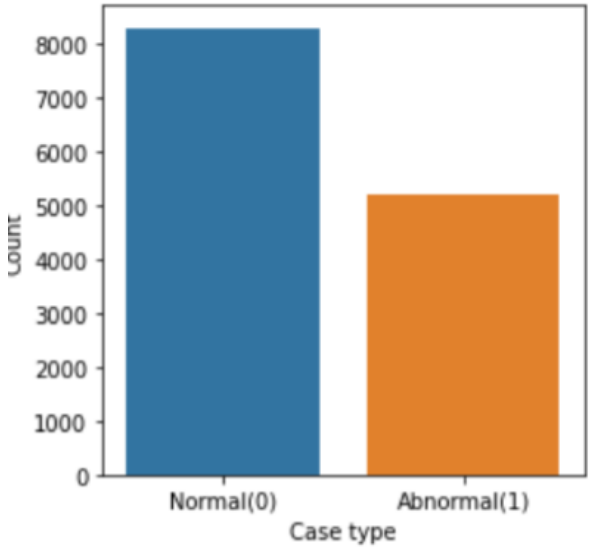


Graph 1: Distribution of Training Data in different regions



Graph 2: Distribution of Test set in different regions

There was a noticeabe class imbalance in training data of this database. Normal class in training data termed as (class 0) contained more images than the class containing abnormal images (class 1) as shown in the graph below. That was noted to affect the training efficacy.



Graph 3: Class distribution of Train studies

This graph shows that, 13,457 out of total 14892 studies belonging to train set, 5177 belong to abnormal (1) class and 8280 belong to normal (0) class, respectively.

3.2 Preprocessing

3.2.1 Resize

As already discussed, the images contained in this dataset were of varying sizes. The size of these images varied significantly, width ranging from 81 to 512, and height from 132 to 512 size. First, it was very important to resize these images and load them into the model in a single standard size. So, all the images were resized to standard 224*224 size by maintaining RGB color to detect any possible color-related features of images, which is also illustrated by (Madan

et al., 2021). Resizing size was experimentally selected to preserve image detail as well as obtain a standard size of all the images to compensate for the computational cost that is required for training of the respective model. ZCA whitening was applied to increase the brightness level of these images.

3.2.2 Cropping to Region of Interest

When the images in the dataset were closely interpreted, they appeared to be in a raw form thus having the need to be pre-processed to make them more homogenous. To obtain better quality images, Cropping to ROI function was used. This preprocessing technique was used to identify the ROI threshold value as well as contours value. When ROI value was calculated, and then this value was used to crop the image to region of interest.

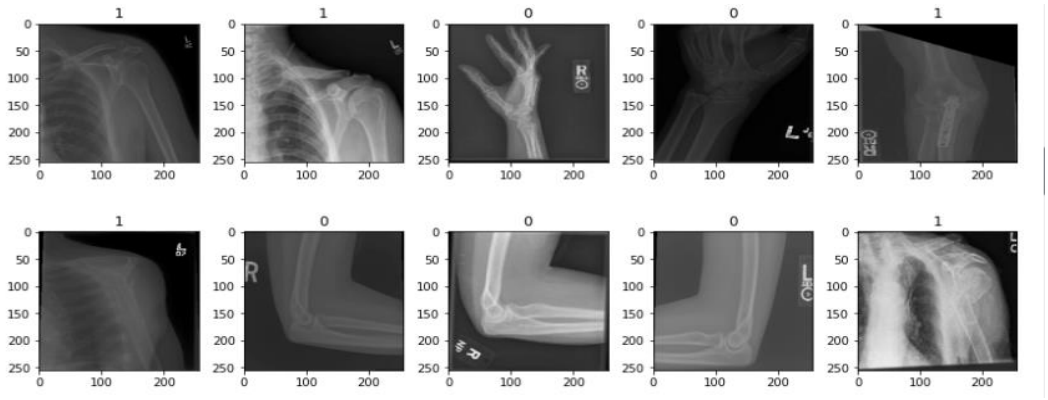


Figure 3: Random subplots of images after applying cropping to ROI function

3.3 Data Augmentation

During data augmentation steps all the images were normalized from (0, 255) to (0, 1) using min, max normalization. These images were then rotated through 45 degrees. Random as well as horizontal flips transforms were also applied to all the images to add variety shapes to be shown to model.

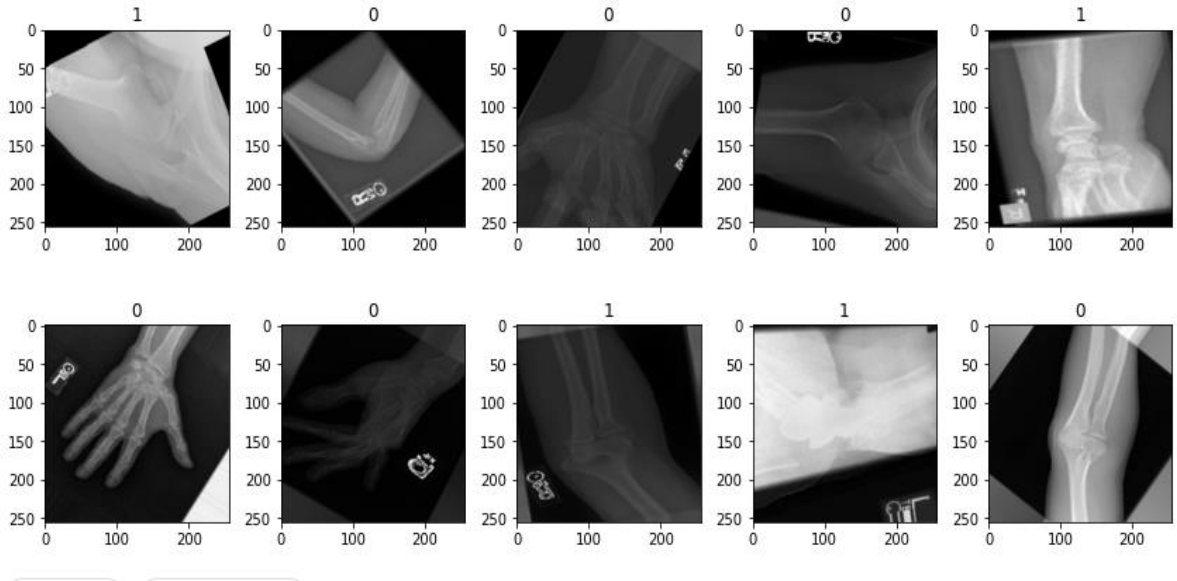


Figure 4: Random subplots after data augmentation transforms

3.4 Training Network

MURA dataset was used in this binary classification problem which classifies data into two labels i.e 0 and 1. Where 0 represents normal images and 1 represents all the abnormal images. The first step during training was to choose the best model to train our dataset. The most used networks were deep convolutional neural network (DCNN) and convolutional neural network (CNN). Choosing the best model for training was a hit and trial method in which different models were used on the dataset. The models used during experimentation purposes included, ResNet-50, VGG-19, Inception-ResNetV2, DenseNet-201 and DenseNet-169. The best model was selected based on training loss parameter. DenseNet-169 offered lowest training loss and best training accuracy.

3.4.1 Proposed Model

In this study a pretrained DenseNet-169 model was used. This model was pretrained using weights of imageNet dataset(Deng et al., 2009). As its name indicates this deep convolutional neural network contains the first convolutional layer, maxpool layers, next it relates to Dense layers which are the fully connected layers, and last are the transition layers. We have used these

layers and popped off the last transitional layer. This model was compiled to calculate predictions on a dataset. The Dense Convolutional Neural Network (DenseNet) is a new CNN yet has outperformed many CNNs like VGG16 and VGG19 by providing state-of-the-art results on highly complex problems. The fundamental idea of DenseNet is to make sure that there is maximum flow of information within layers in the network by connecting all layers directly with each other in a feed forward fashion, each layer in the dense block gets information from subsequent layers and thus transfers information to subsequent layers removing the vanishing gradient problem and need of deep nets(Huang et al., 2017). In this model we replaced the final fully connected layer with the sigmoid activation as classification layer to predict the abnormalities.

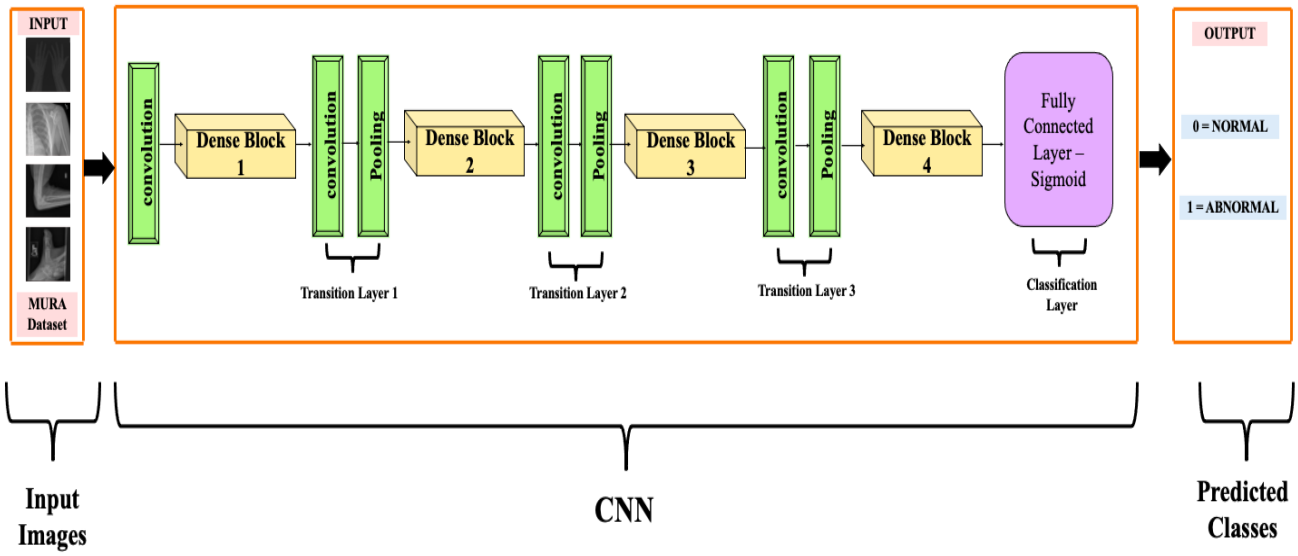


Figure 5: Model Architecture

3.4.2 Tuning of Hyperparameters

Because of the large size of the dataset, the whole of the data was first loaded into a data frame using an image data generator. And then data was trained by calling it from those data frames. Taking into consideration the large size of MURA dataset, data was loaded using mini- batches of size = 16 for each session. The number of epochs used was 30-50 epochs with Adam

optimizer. Adam optimizer used default values of learning rate as 0.0001, and β-parameters of β1= 0.9, β2 = 0.999. Rest of the hyperparameters used are listed in the table below.

Table 2: Description of Hyperparameters

Hyperparameters	**Description**
learning rate	0.0001
Epochs	50
Batch size	16
Optimizer	Adam
Activation Function	Sigmoid
Loss	Binary cross entropy

3.5 Training

In this study we experimented with different models including VGG-16, VGG-19, ResNet-50, Inception-ResnetV2, and DenseNet-121. But DenseNet-169 provided the best results with MURA dataset. This model was selected because it connects the proceeding layers in a feed forward fashion, removes gradient vanishing problem and faster speed(Huang et al., 2017; Madan et al., 2021). This data was split into 80:20. 80% was used for training purposes and 20% for validation during the model training. All images were resized to 224*224-pixel size and fed into model. Adam optimizer was used with default β-parameters of β1= 0.9, β2 = 0.999. The learning rate used was fixed to $1e^{\wedge-4}$, The batch size was fixed to 16 epochs for all epochs. Varying epochs of 30-50 were used. The best model accuracy was set as a check point and model was saved for best validation accuracy. The saved model was then evaluated on the test set. The original valid set released by Stanford ML group was used to evaluate the model performance and evaluation metrics was calculated as shown in the figure below showing the generalized pipeline of binary classification framework. The model was trained using Intel Core i7 CPU with 16 GB RAM, 512 SSD memory with NVIDIA Tesla T4 GPU with 16GB memory and RAM. Table 3 shows the specifications of the environment used.

Table 3: Specifications of the environment

Programming language	Python 3.8.8
Ubuntu version	20.04.4 LTS
RAM	32GB
GPU	16GB
CUDA version	11.4
Deep learning Framework	TensorFlow version 2.6.4 and Keras

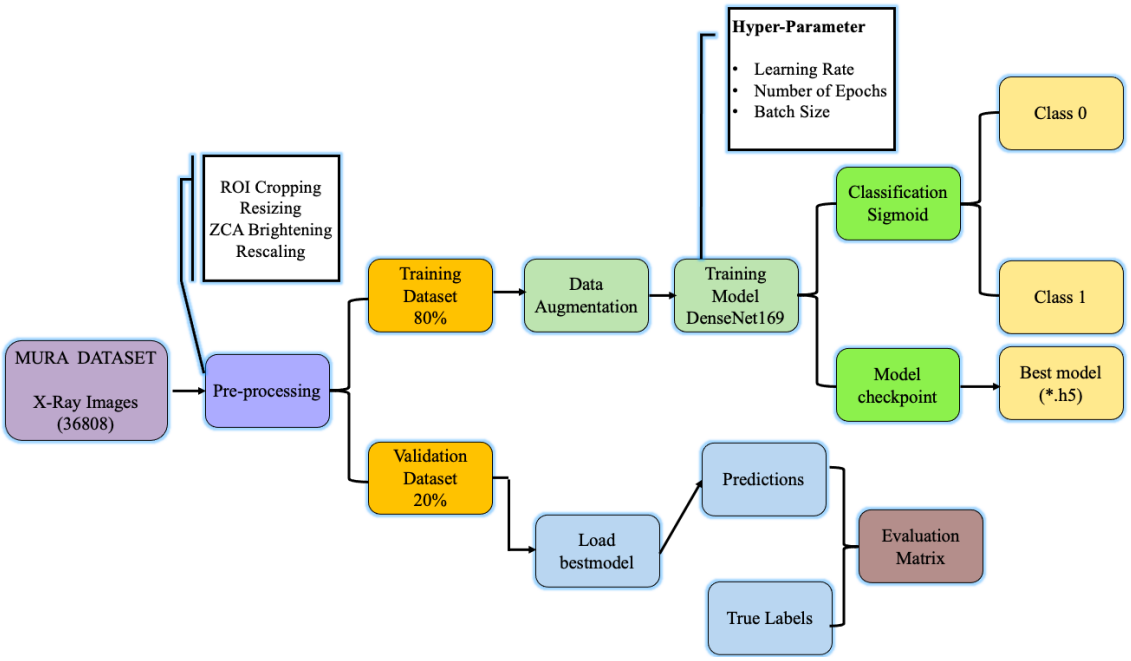


Figure 6: Generalized pipeline of Binary classification framework showing complete training and validation process

3.6 Evaluation Metrics for Binary label classification Task

Binary classification is a single label classification, in which the binary classifier places the images in dataset in one of the two classes. The following are example-based metrics used for Binary classification.

3.6.1 Example Based Metrics

Accuracy: The metrics, is a ratio of positive predictions to total number of predictions, is referred to as Accuracy. Given as,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP= The examples correctly identified as positive or abnormal**,**

TN= the examples correctly classified as negative or normal**.**

FP= those examples which are classified as positive (abnormal) but are negative (normal) instead.

FN= those examples which are classified as negative (normal) but are positive (abnormal) instead.

Precision: The metrics that show how many examples of total correct predictions are actually positive. It is a ratio between Truly positive predictions and total number of positive predictions.

$$precision = \frac{TP}{TP + FP}$$

Recall: This metric is also labelled as sensitivity which shows the true positive values out of actual positives. It is a ratio between Truly positive predictions and actual positive predictions.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: This metric shows the combined results of recall and precision. As it is a harmonic mean of recall and precision. Given as:

$$F1 = 2.\frac{precision \cdot Recall}{precision + Recall}$$

3.6.2 Label Based Metrics

This category uses two types of averaging methods. Prior, is called macro- average where the binary evaluation metric is computed for each individual class and later averaged over all classes. Whereas the second metric is called micro-average binary evaluation metrics is computer for all the samples and classes. Receiver operating curve (ROC) also known as AUC is widely used in MLC task since it helps in eliminating subjectivity in the threshold selection process, as continuous probability derived scores are transformed into binary presence or absence by summarizing overall performance of the model over all possible thresholds.

CHAPTER 4: RESULTS

In this section we will discuss the results obtained on MURA dataset. The model was trained on the labels from MURA dataset released by (Rajpurkar et al., 2017). The training data composed of 36808 images that was split into 80% (29446 images) training set and 20% (7362) validation set. Pretrained DenseNet-169 model was used to train the dataset and our proposed methodology achieved 94.75% Training accuracy and 81.35% validation accuracy. Table 2 below shows the results of training MURA dataset. And graph A shows training & validation accuracy, and graph B shows training & validation loss.

Table 4: Results of training on MURA dataset

Model	Training Time	Metrics			
		Training Accuracy	Training Loss	Valid Accuracy	Valid Loss
DenseNet169	10 hours and 30 minutes	94.75	0.1744	81.35	0.541

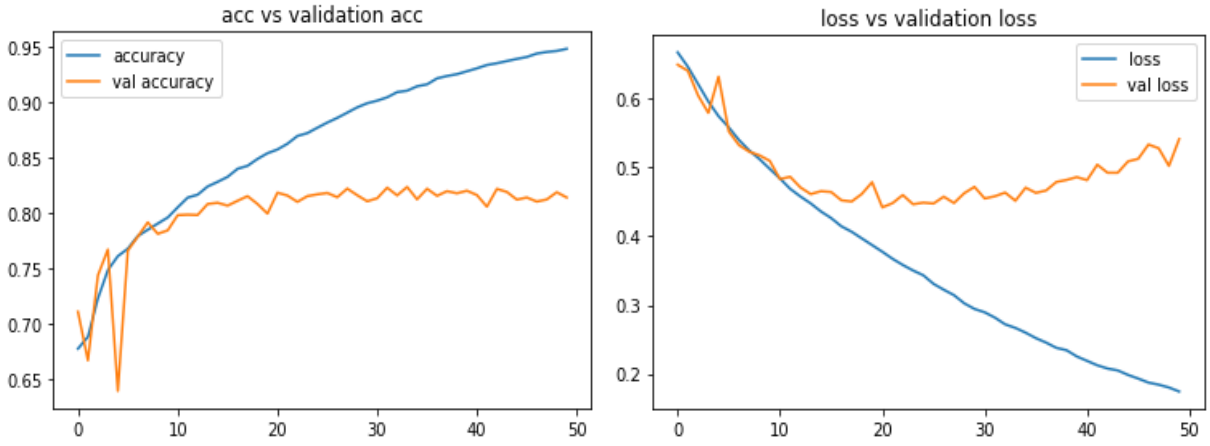


Figure 7 Graphs showing Training results. (A) shows Training and validation accuracy. (B) shows Training & Validation loss

In this research we implemented use of single DenseNet-169 model on complete dataset for the first time in literature and computed comparable results by using multiple preprocessing techniques and data augmentation steps. By making use of optimum hyperparameters and selecting the best learning rate for the provided dataset promising results were achieved upon evaluation of the model which shows competency of the proposed methodology to be used as an automatic binary classifier of musculoskeletal radiographs. Our proposed supervised learning approach has achieved the highest average AUC of 0.8481. The classification report on the testing data is obtained and printed below:

	precision	recall	f1-score	support
0	0.84	0.83	0.83	4458
1	0.74	0.75	0.75	2904
accuracy			0.80	7362
macro avg	0.79	0.79	0.79	7362
weighted avg	0.80	0.80	0.80	7362

Time taken to predict the model 0.017354965209960938.

This report shows that the weighted average obtained after model evaluation and prediction is 80%, macro average is shown to be 79% whereas weighted average is obtained as 80% on testing dataset. Table 5: given below shows the comparison of our results with literature using the same proposed technique. The table shows with a comparison that our study achieves comparable results using a single model on complete dataset with fine-tuned hyper-parameters and multiple preprocessing steps.

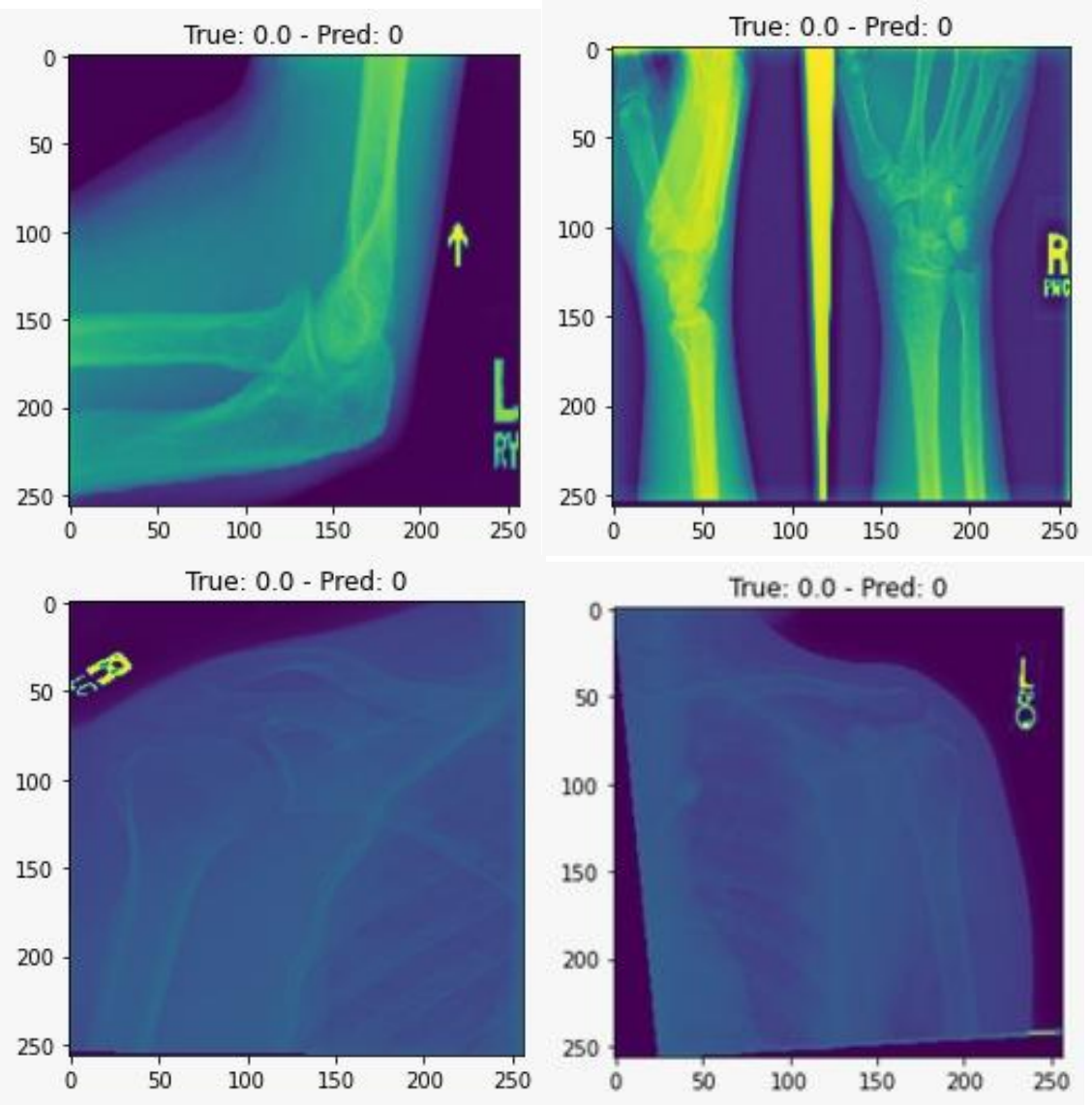
Table 5: The comparison of our results with literature

Comparison **Metrics**	**(Rajpurkar et al., 2017)**	**(Solovyova & Solovyov, 2020)**	**(He et al., 2021)**	**(Madan et al., 2021)**	**Our proposed** **Model**
Model/Ensemble	Ensemble of 5 DenseNet-169	Ensemble of 4 DenseNet-169	Calibrated Ensemble of 3 models	DenseNet-169 on Humerus region	Single DenseNet-169 model
Accuracy	0.887	0.863	0.93	0.840	0.80
AUC/AUROC	0.929	0.870	0.93	-------	0.8481

We had generated model predictions where threshold value was 0.5. Any of the images that had threshold value greater than 0.5 was labelled as 1 (abnormal) and those images whose values were less 0.5 were predicted as 0 (normal) label. Then we compared true labels with model predictions to check whether our model was performing accurately or not.

Examples where model correctly predicted true labels for class 0:

Here, we have printed the examples from the dataset which were correctly predicted by our proposed model from class 0 (normal/negative). By correct prediction it means that model predicted labels match the true labels provided by the dataset:

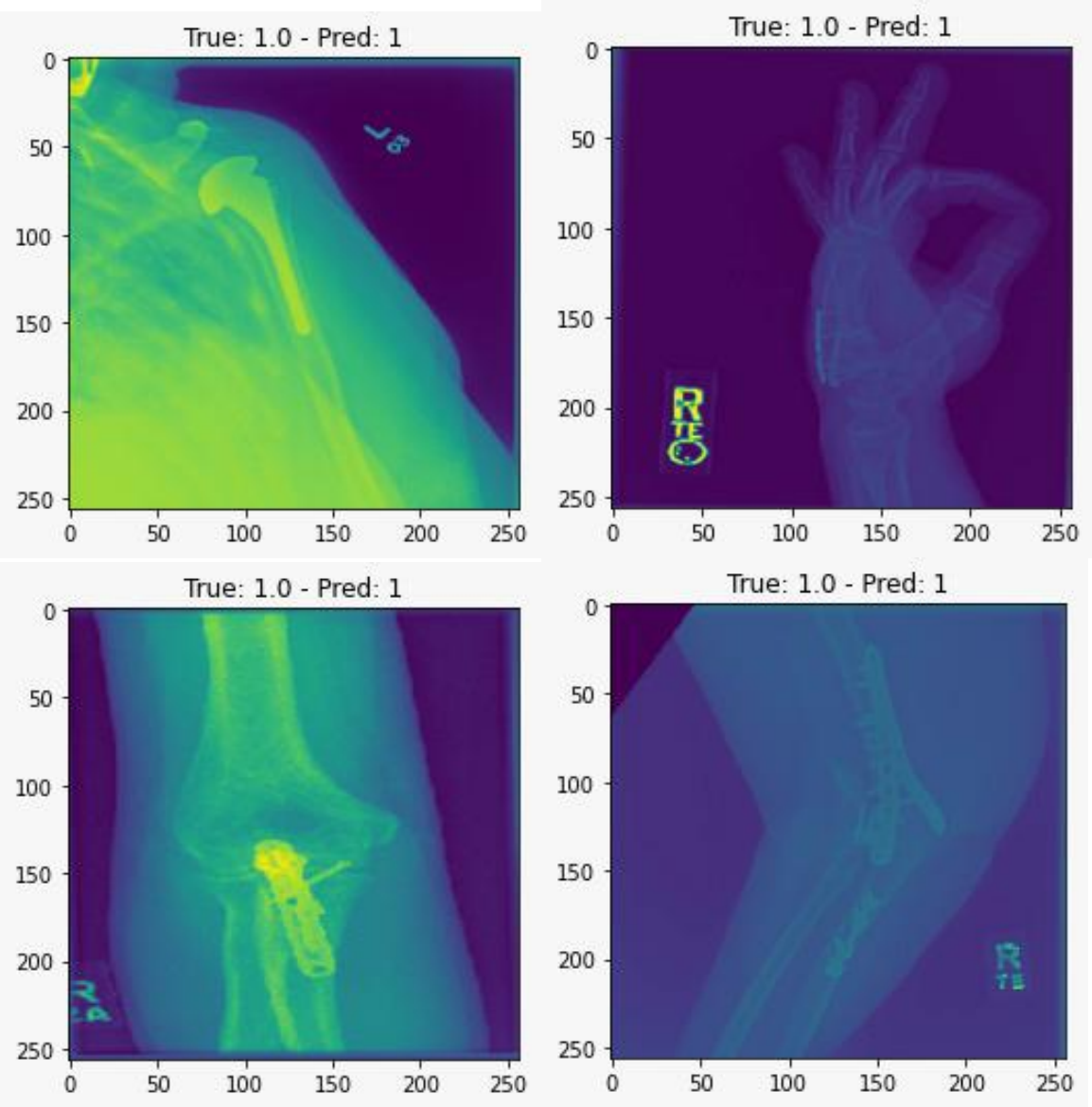


In all these images, true label was = 0

And model also predicted them in class = 0

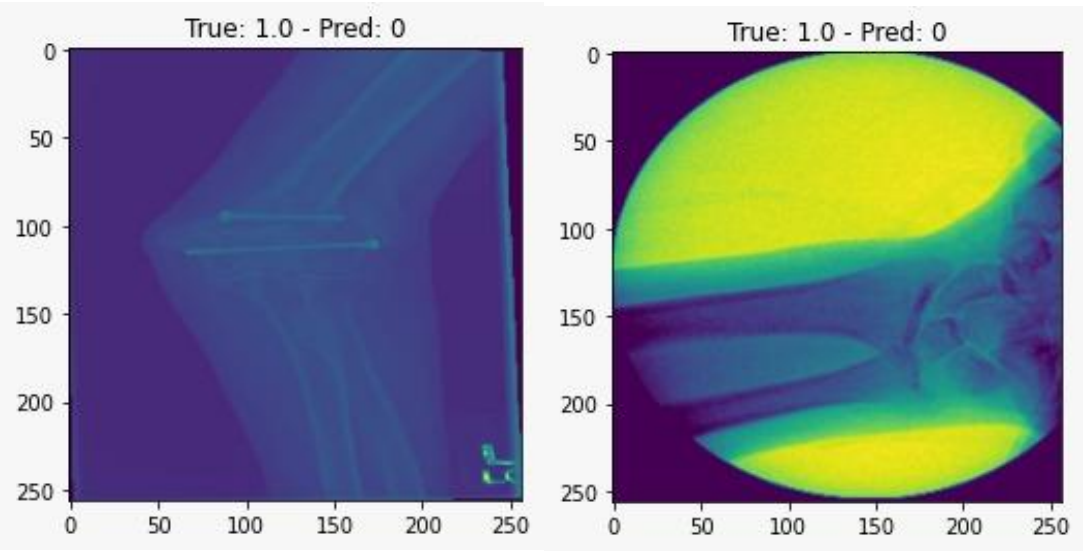
Examples where model correctly predicted true labels for class 1:

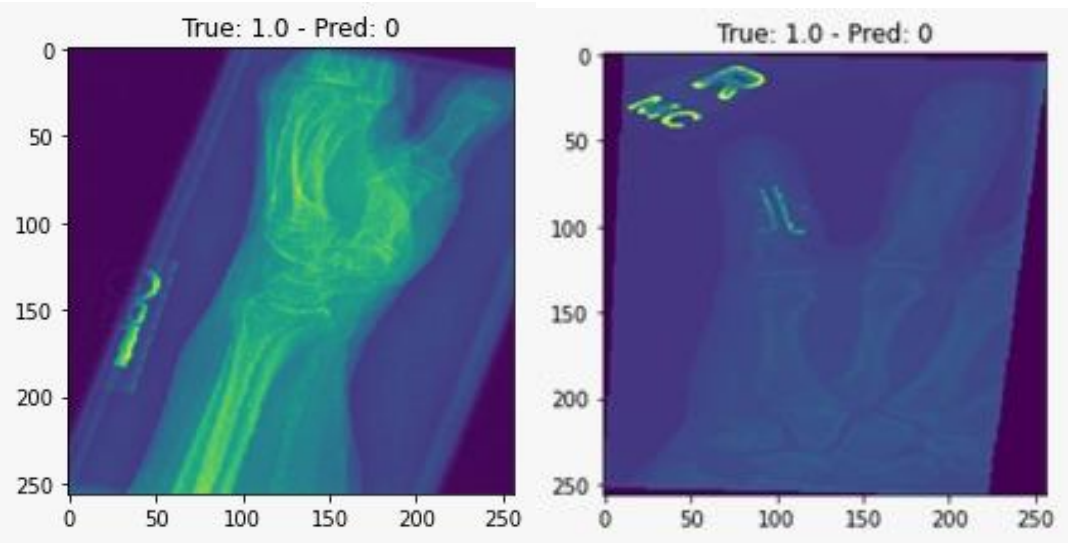
Here, we have printed the examples from the dataset which were correctly predicted by our proposed model from class 1 (abnormal/positive). By correct prediction it means that model predicted labels match the true labels provided by the dataset:



Examples where model predicted false labels:

Following images show the examples where our proposed model did not predict the true labels, hence gave false negative results:





In all these printed examples, images belonged to class (1) abnormal and were predicted falsely to be from class (0) normal.

CHAPTER 5: DISCUSSION

This study emphasizes the classification of abnormal and normal X-rays of bones and joints belonging to upper limb because of the high incidence and prevalence of musculoskeletal traumas like fractures and dislocation. Many studies have used MURA dataset in the previous years. But not all the studies have used all the subsets contained within MURA dataset. There are several studies that used only one or two subsets of this complete dataset(Ghosh et al., 2021; Madan et al., 2021; Uysal et al., 2021) because of the large size of this dataset. Our study is one of those few studies which used complete dataset for experimentation and provided closely lying results. One thing which was common in all of the previously published studies using the complete dataset is that they trained more than one model on provided dataset. And thus, calculating final predictions using ensembled models. However, we differentiate this study not only by working on complete dataset released by Stanford ML group, but also by training a single model and computing comparable results. DenseNet-169 model was selected in this study for training purposes because of its high recognition efficiency according to an analysis provided by (Bianco et al., 2018). To create a fair comparison, we used the same ratio of train-test split as used in literature that is 80:20, in which 80% represents training data and 20% testing data was split. During experimentation on the MURA database, certain limitations were observed. The first limitation was the large size of dataset due to which this dataset was difficult to be used completely and people used only a single or two subsets of this complete dataset. Secondly, it was noticed that less number of images of each subset are present in this data as stated by (Teeyapan, 2020). Additionally, a considerate class imbalance is also present in certain subsets which are found to be underperforming. Despite all these limitations mentioned here, closely lying accuracy was achieved using multiple pre-processing steps, and fine-tuned hyper-parameters with a single DenseNet-169 model.

Our study aimed at using a single classification model and computing results. Hence, this is the first ever study in literature which used single DenseNet-169 model and not ensemble of models. Here we are comparing our results with the ones who used this same approach in literature.

(Rajpurkar et al., 2017) used the same approach and ensembled 5 DenseNet-169 models to compute results on the original test set released by them. The value of AUC/AUROC = 0.929.

(Solovyova & Solovyov, 2020) used the same approach by adding more preprocessing steps and using ensemble of 4 models of DenseNet-169. With this approach they attained an improved kappa score, and AUC value was noted to be AUC = 0.870, Accuracy = 0.863. (He et al., 2021) used calibrated ensemble that contained ResNet, DenseNet and ConvNet which provided AUC = 0.93 and accuracy = 0.93 respectively.

(Madan et al., 2021) used the same DenseNet-169 model but trained it only on Humerus subset of the complete MURA dataset and thus achieving accuracy = 84.03% , kappa score = 0.68. Whereas our proposed methodology using single DenseNet-169 model over complete dataset achieved accuracy = 80% and AUC = 0.8481 which is comparable to the models which used ensembles to generate similar results.

CHAPTER 6: CONCLUSION

In this study we proposed using a single model without many architectural modifications. We used convolutional neural network's DenseNet-169 model on keras by adding more pre-processing techniques and enlarged the size of data by increasing data augmentation steps. Through this approach we achieved better results in terms of accuracy and AUC. Training data was divided into 80:20 for training and validation sets respectively, whereas, testing of model was done on validation set. The results obtained through the proposed technique include 80% testing accuracy. This validates the effectiveness of this method for bone fractures classification. The practical significance of this study is the implementation of AI algorithms to assist radiologists in improving their diagnostic accuracy by reducing the chance of incorrect diagnosis of fractured radiographs. For future work we plan to use ensemble of more than one model to improve accuracy further. Also, if ground truths are made available publicly, then we can localize the fractured site by segmenting that region and combine this classification with segmentation problem.

CHAPTER 7: REFERENCES

Al-Ayyoub, M., Hmeidi, I., & Rababah, H. (2013). Detecting Hand Bone Fractures in X-Ray Images. *J. Multim. Process. Technol.*, *4*(3), 155-168.

Bianco, S., Cadene, R., Celona, L., & Napoletano, P. (2018). Benchmark analysis of representative deep neural network architectures. *IEEE access*, *6*, 64270-64277.

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

de La Torre, J., Puig, D., & Valls, A. (2018). Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, *105*, 144-154.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition,

Fernholm, R., Pukk Härenstam, K., Wachtler, C., Nilsson, G. H., Holzmann, M. J., & Carlsson, A. C. (2019). Diagnostic errors reported in primary healthcare and emergency departments: a retrospective and descriptive cohort study of 4830 reported cases of preventable harm in Sweden. *European Journal of General Practice*, *25*(3), 128-135.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

Ghosh, M., Hassan, S., & Debnath, P. (2021). Ensemble based neural network for the classification of mura dataset. *Journal of Nature*, *4*, 1-5.

Hameed, M. H., Ghafoor, R., Khan, F. R., & Bada, S. B. (2016). Prevalence of musculoskeletal disorders among dentists in teaching hospitals in Karachi. *JPMA: Journal of Pakistan Medical Association*, *66*(10), S-36.

Haroon, H., Mehmood, S., Imtiaz, F., Ali, S. A., & Sarfraz, M. (2018). Musculoskeletal pain and its associated risk factors among medical students of a public sector University in Karachi, Pakistan. *JPMA. The Journal of the Pakistan Medical Association*, *68*(4), 682-688.

He, M., Wang, X., & Zhao, Y. (2021). A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs. *Scientific Reports*, *11*(1), 1-11.

Hooda, M., & Shravankumar Bachu, P. (2020). Artificial Intelligence Technique For Detecting Bone Irregularity Using Fastai. International Conference on Industrial Engineering and Operations Management Dubai, UAE,

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Ishida, T., Niu, G., & Sugiyama, M. (2018). Binary classification from positive-confidence data. *Advances in neural information processing systems, 31*.

Kandel, I., & Castelli, M. (2021). Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset. *Health information science and systems*, *9*(1), 1-22.

Madan, S., Kesharwani, S., Akhil, K. V. S., Balaji, S., Bharath, K., & Kumar, R. (2021). Abnormality Detection in Humerus Bone Radiographs Using DenseNet. 2021 Innovations in Power and Advanced Computing Technologies (i-PACT),

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., & Ball, R. L. (2017). Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*.

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

Teeyapan, K. (2020). Abnormality Detection in Musculoskeletal Radiographs using EfficientNets. 2020 24th International Computer Science and Engineering Conference (ICSEC),

Uysal, F., Hardalaç, F., Peker, O., Tolunay, T., & Tokgöz, N. (2021). Classification of shoulder X-ray images with deep learning ensemble models. *Applied Sciences, 11*(6), 2723.

Walker-Bone, K., Palmer, K. T., Reading, I., Coggon, D., & Cooper, C. (2004). Prevalence and impact of musculoskeletal disorders of the upper limb in the general population. *Arthritis Care & Research*, *51*(4), 642-651.