Springboard Data Science Career Track

Capstone Project 2
Identifying Abnormalities in Bone X-Rays
Final Report
September 2020
by
Adeyemi Adejuwon

Introduction

In many areas of the world, radiologists interpret radiographs visually to determine whether there are defects in bone structures.

The goal of this capstone project is to develop supervised machine learning models to estimate the probability of an X-ray image showing a musculoskeletal abnormality, or not. Determining how these models would be used in an actual healthcare environment is out of the scope of this project.

Objective

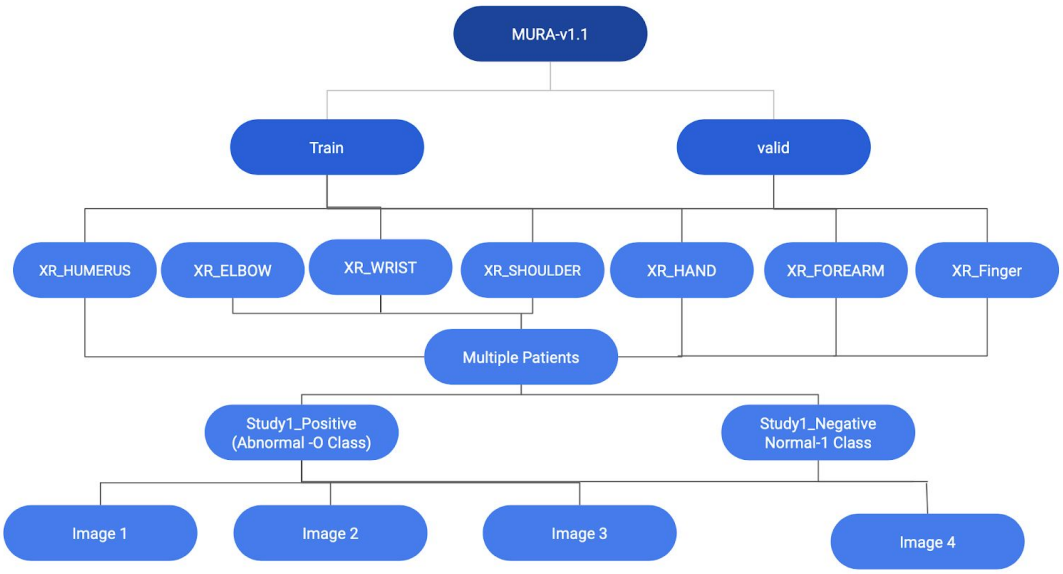
In this capstone project, binary classification models were developed to estimate the probability that an X-ray image is showing a musculoskeletal abnormality or not. The machine learning algorithms used in this project are: Logistic Regression, KNN, Random Forest Classifiers and SVM.

Dataset

I used the dataset made available by Stanford University known as MURA (for **MU**sculoskeletal **RA**diographs)[1]. The dataset is organized in multiple folders divided into training and validation groups. These folders are further subdivided into separate folders of 7 body parts located in the upper extremities: elbow, humerus, wrist, hand, finger,shoulder and forearm. Each upper extremity folder is further subdivided into patient information. The patient information is labeled as patient ID. The patient information is further subdivided into two study folders, negative and positive. The negative X-rays feature bones without abnormalities, while the positive X-rays show some abnormality, e.g., fractures. The study folders contain the X-ray images of the upper extremities. This folder structure of the MURA dataset is displayed in Fig.1.

[1] https://stanfordmlgroup.github.io/competitions/mura/

Fig 1. Folder Arrangement of X-ray dataset



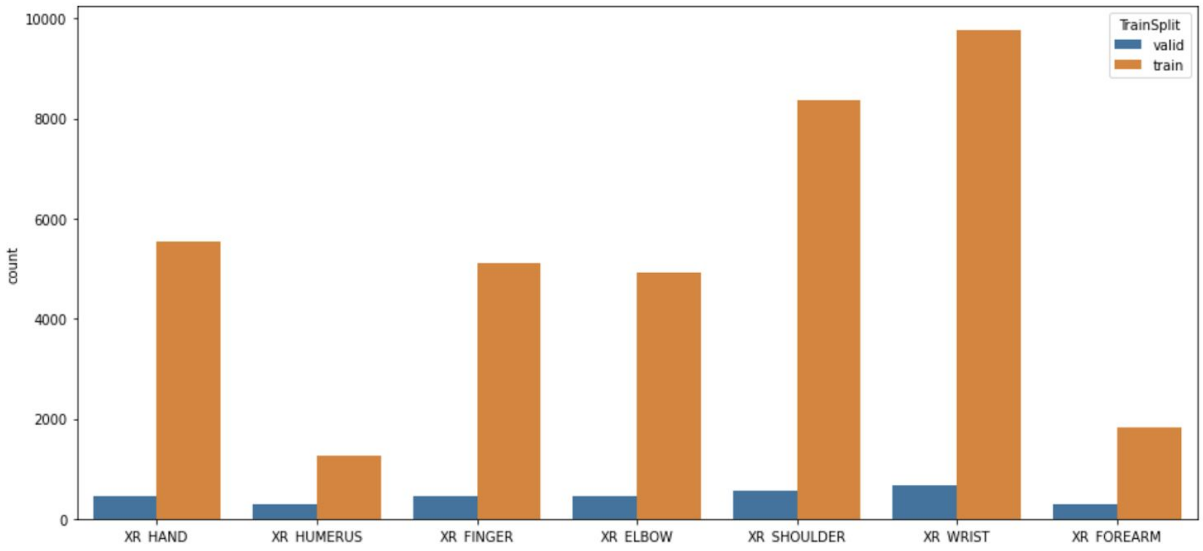
Data Cleaning & Wrangling

Before beginning analysis of the data, it is essential to first prepare them before further analysis can be conducted. This upfront work will give more confidence in interrogating the data and would allow better conclusions to be made as regards the dataset. The libraries used for the data cleaning and wrangling of the data sets are:

- Numpy for scientific computing of the numerical arrays
- Pandas for data analysis and manipulation,
- Matplotlib for visualization

The dataset comprises radiographs from 12,251 patients with a total of 40,895 X-ray images. The distribution of the body parts in the supplied dataset is displayed in Fig 2, divided into images in the training and validation sets.

Fig 2: X-ray Images of Upper Extremities for both Training and validation



In total there are 8,280 X-ray images in the training dataset, classified as zero (0) class for the normal X-rays, and 5,177 classified as one (1) for the abnormal X-rays. Thus, 62% images are normal and 38% abnormal. In the validation dataset there are a total of 661 images for the normal class, and a total of 538 images for the abnormal class. The ratio here is 55% normal and

45% abnormal. This can be seen in Fig. 3 and 4 below. A summary of the data distribution of the upper extremities can be seen in Table 1.

Fig 3: Normal vs Abnormal X-ray Images for Training data

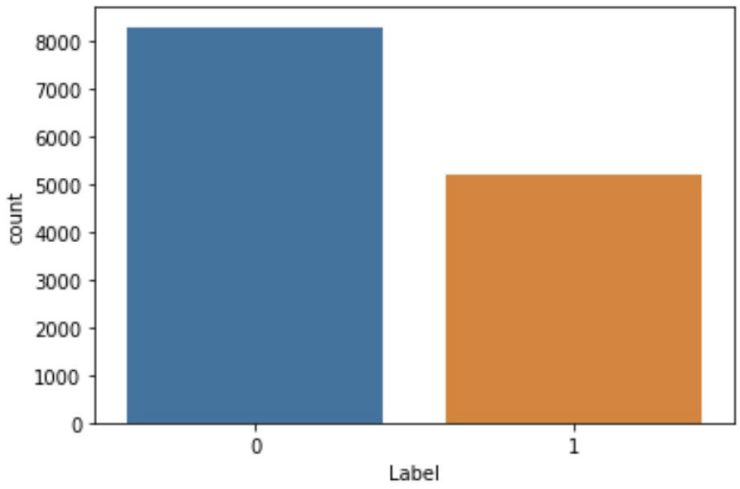


Fig 4: Normal vs Abnormal X-ray Images for Validation Data

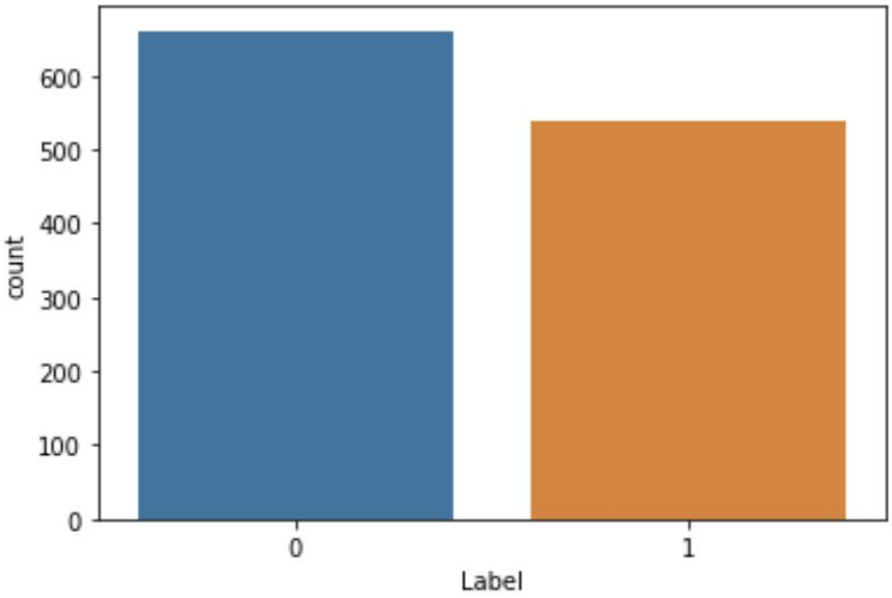
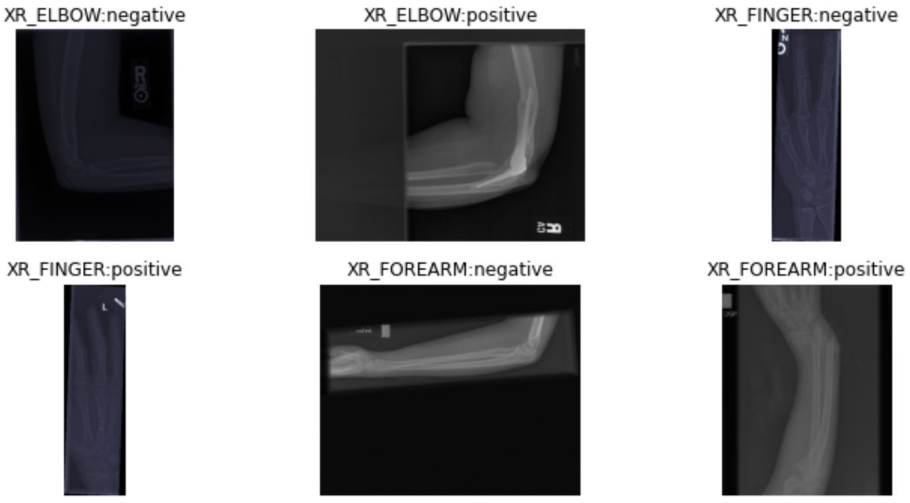


Table 1: Distribution of Upper extremities for both training and validation data

Body Part	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 No. of Studies	8280	5177	661	538	14656

Samples of the clinical image supplied by the Stanford group are displayed in Fig 5. The clinical images supplied by the Stanford group vary in resolution and in aspect ratios.

Fig.5: Samples of X-ray images



As can be seen from the sample dataset in Fig.5, the X-Ray images are of different orientations and dimensions. In order to feed these images into our future machine learning model, each image had to be normalized and reshaped. The images needed to be of appropriate sizes so that not too much information is lost in the reduction of the size, when inputting them into our future model. This means we would have to preprocess the images before it could be used in our model. This reduced image size would also be small enough to be computationally efficient when modeling the images.

This portion of the data wrangling is called image preprocessing. An uniform image size of 224 x 224 was chosen for the image preprocessing, because this was the size used by the Stanford group also to do their preprocessing work[1].

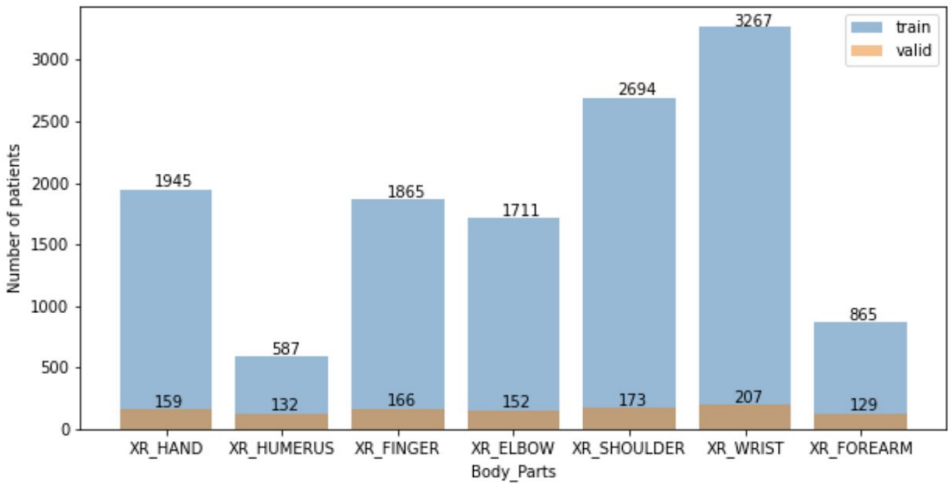
Storytelling and Inferential Statistics

Following the data preprocessing, the next step is to interrogate the dataset and extract features of the dataset. The questions and the answers I found helped me better understand the dataset.

The first question asked is:

What is the distribution of X-rays versus the number of patients in the dataset?

Fig 6: Number of patients versus body parts for both training and validation data



From the bar chart in Fig. 6, it can be observed that most X-rays were obtained from patients with wrist incidents. These had the most training and validation datasets. The X-rays with the least observations for the training dataset was that of the humerus, while the X-rays with the smallest dataset was that of the forearm.

The second question asked is:

What is the distribution of the abnormal (1 class) and normal (0 class) X-rays in our dataset?

The answer to this question is displayed in Fig. 7 and 8

Fig 7: Number of patients versus body parts for both Normal and Abnormal Data for Training data

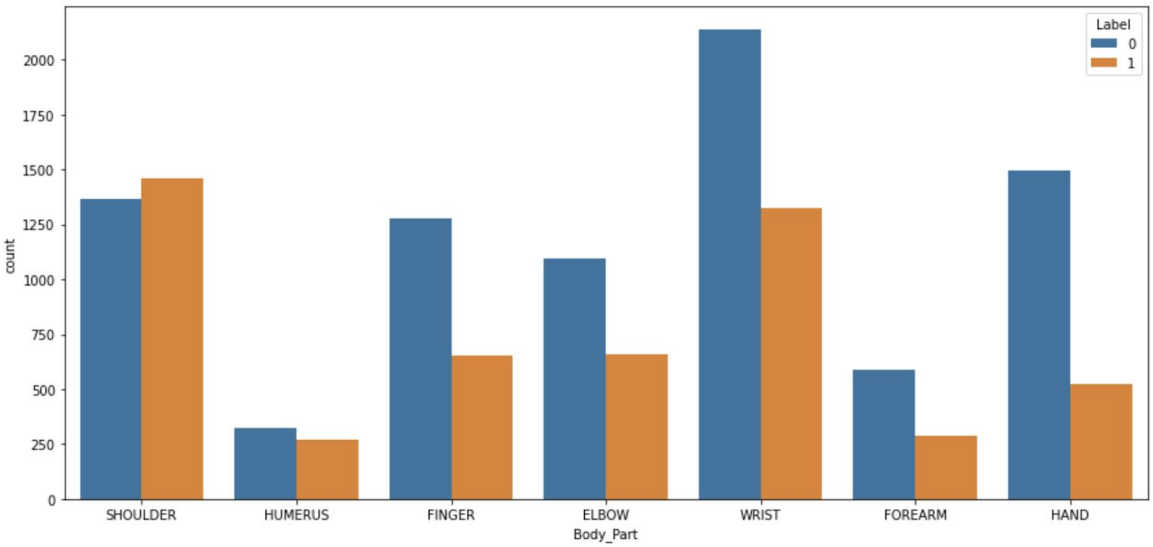
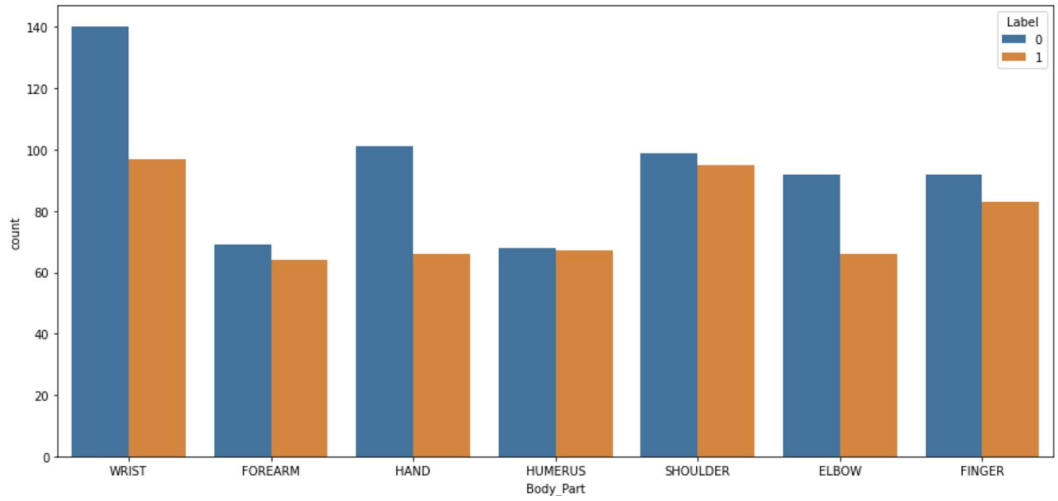


Fig 8: Number of patients versus body parts for both Normal and Abnormal Data for Test data



From the bar chart in Fig. 7 and 8, it can be observed that most X-rays were obtained from patients with wrist incidents for the normal and the abnormal dataset. The training and the validation dataset followed the same distribution for the body parts i.e. The most X-ray observation was for the wrist, and the least X-ray observation was for humerus

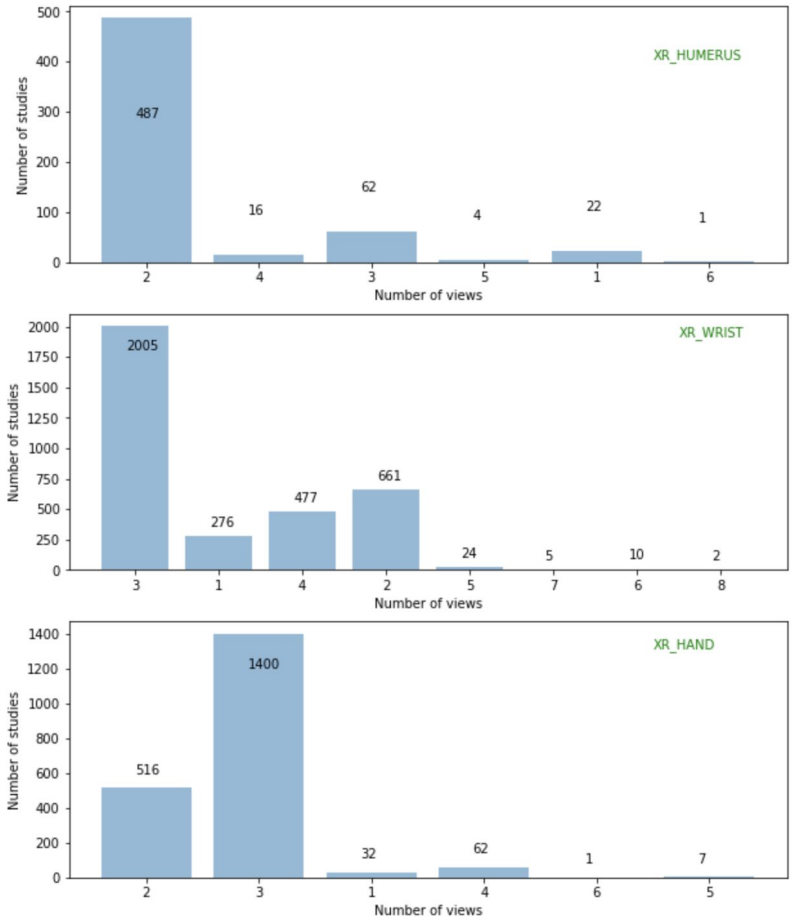
The third question asked:

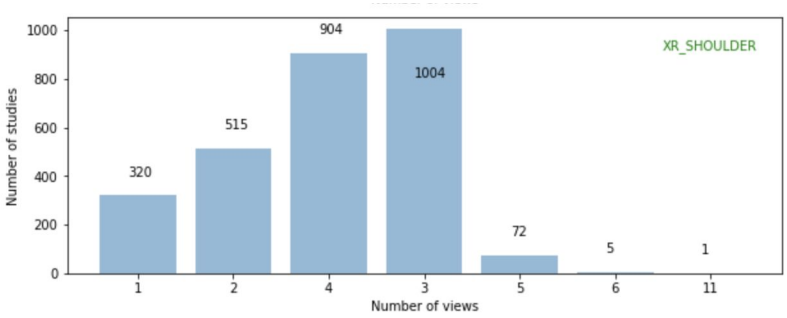
Which upper extremity part had most X-ray views?

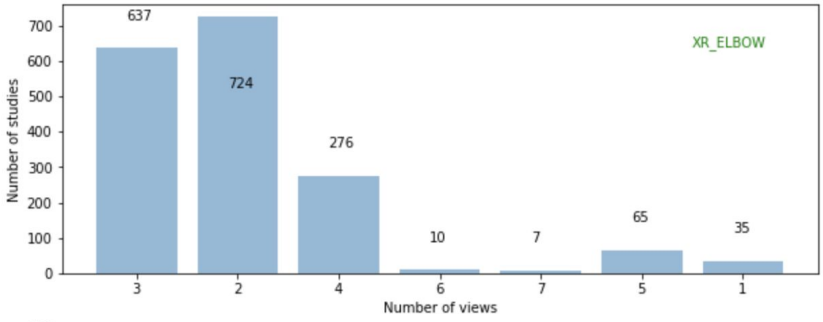
This question is being asked because some body parts might require more views by the radiologist to make a better diagnosis of the patient.

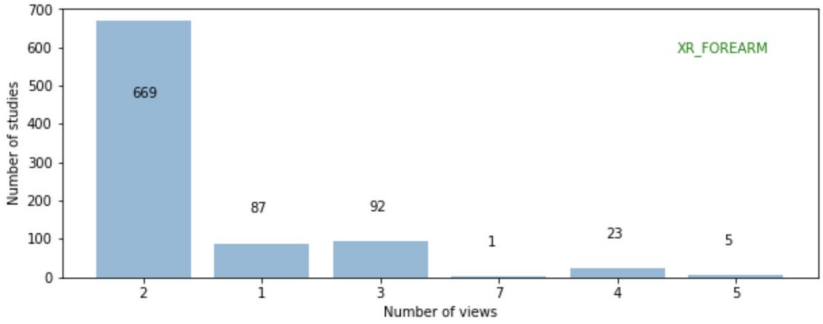
Fig. 9 shows that multiple views are expressed for each upper extremity part. Most studies have mostly 2,3 and 4 images.The maximum number of images is the XR_Shoulder. A patient can be seen to have 11 images for one study type for the same shoulder.

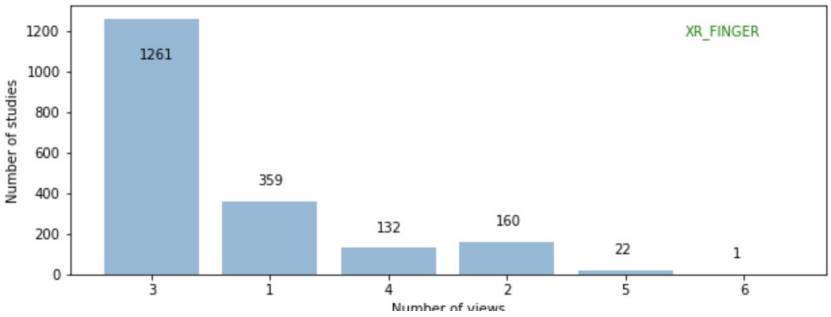
Fig. 9: X-ray Studies for each Upper Extremity Body Type




XR_SHOULDER
Number of studies
Number of views
320
515
904
1004
72
5
1
1
2
4
3
5
6
11


XR_ELBOW
Number of studies
Number of views
637
724
276
10
7
65
35
3
2
4
6
7
5
1

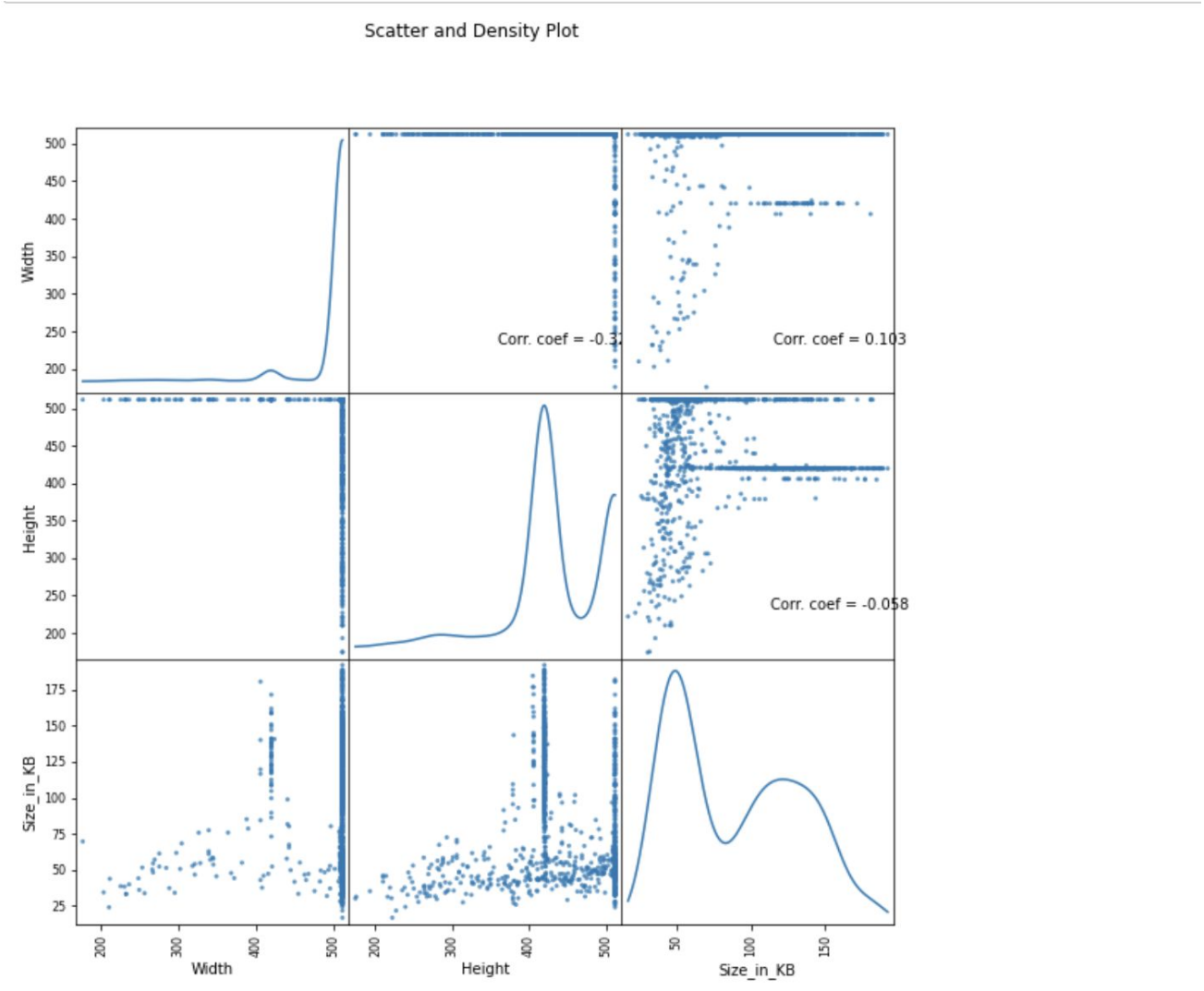

XR_FOREARM
Number of studies
Number of views
669
87
92
1
23
5
2
1
3
7
4
5


XR_FINGER
Number of studies
1261
359
132
160
22
1
3
1
4
2
5
6

Applications of Inferential Statistics

From Fig 9, it can be observed that each study can have one or multiple images. These images vary in width, height and size, with resolution of images varying from 200 pixels to 512 pixels. A variation of height and width distribution of the images from the humerus is displayed in the scatter and density plot in Fig. 10

Fig. 10. Scatter and Density plot for X-ray Images of the Humerus



The scatter and density plot shows that there is little correlation between the height, width and size of images of X-Rays of humerus. Correlation coefficients of -0.31 were obtained for the relationship between the height and the width of images,correlation coefficients of .103 was obtained between the width and size of document, and a correlation of -0.05 was obtained for the height and size of the document. The scatter and density plots also show the size of the image folders to vary from 20 to 250 KB

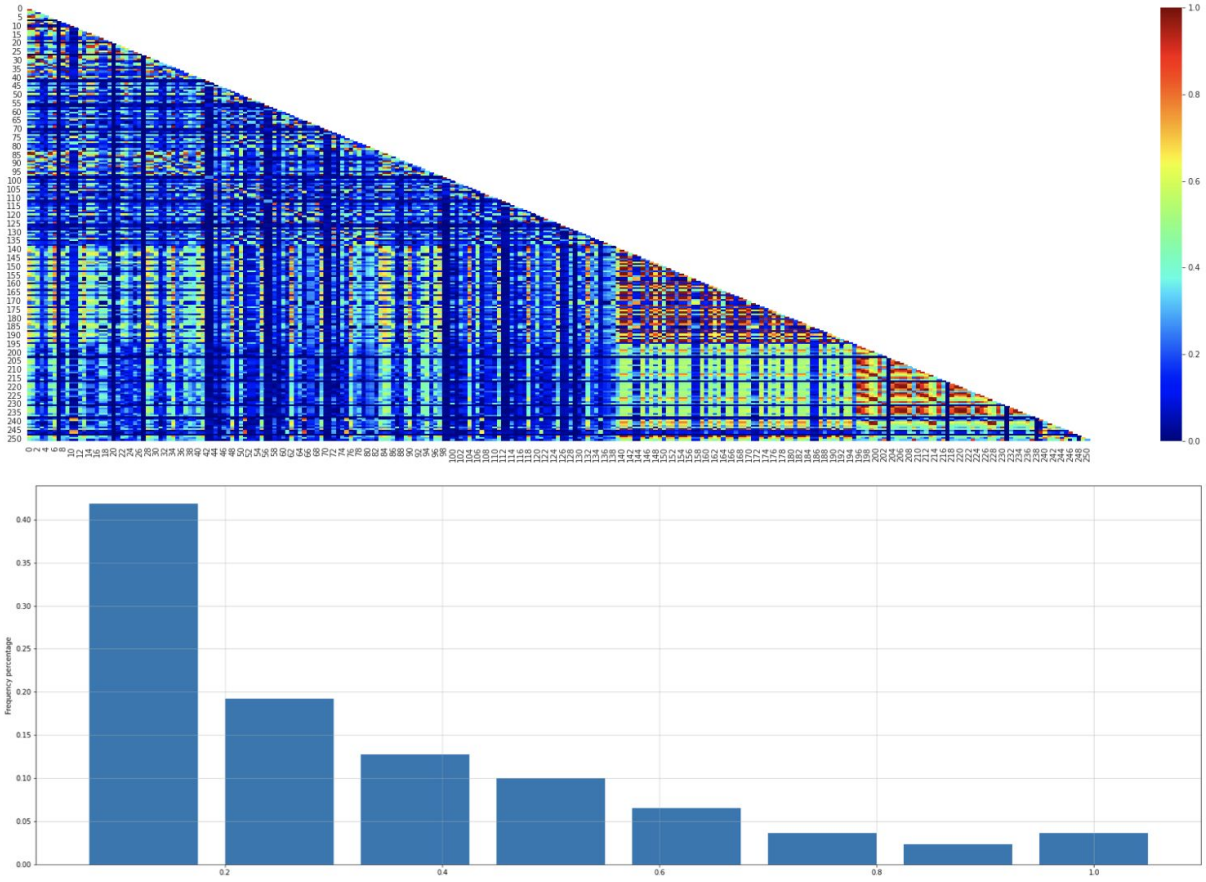
Feature Extraction

Feature extraction was done using the Gray Level Co-occurrence Matrix (GLCM) method described in reference 2. In this method, a feature array was constructed for each X-ray image of the humerus using the texture , FFT, wavelet, GLCM, and GDM methods. The measured features for each image consisted of mean, std, skewness, kurtosis, energy, entropy, max, min, mean deviation, median, range, RMS, uniformity,mean gradient, and std gradient.

The feature extraction scheme resulted in 252 features for each X-ray image in total (14 features from Texture, 14 features from FFT, 56 features from GLCM, 56 features from GLDM, and 112 features from Wavelet).

A Pearson correlation correlation coefficient resulting in a matrix for each pairwise feature combination was computed and displayed in Fig. 11. Analysis of the histograms of the feature show that 85% of the features have correlation coefficients of less than 0.5. The bar chart also shows some pairs of features to be highly correlated. In order to find out the values that are highly correlated , values greater than 0.8 are filtered out, and for each of these pairs of highly correlated values, one value is chosen.

Fig. 11. Heat map and Histogram representation of Pearson correlation Coefficients for the Humerus



Baseline Modeling

1. **Logistic regression**

The performance evaluation of the X-ray images for the logistic regression model after hyperparameter tuning is displayed in Table 2. C values ranging from 0.001 to 100, and solvers of lbfgs, sag, saga, newton-cg were tested on the dataset. The optimal C value calculated on the dataset was 100 while the optimum solver chosen was saga.

The performance measures calculated in Table 2 show that the upper extremity with the best F1 score on the validation dataset for the abnormal class was the **humerus.** The F1 score calculated was 0.58

Extended Modeling

The performance evaluation of the X-ray images for upper extremities body parts is further modeled using different machine learning techniques; K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest Classifiers (RF). Hyper-parameters were tuned for some models.

1. **K-Nearest Neighbors**

For K-nearest neighbors, the best F1-score was for the **humerus**. This F1-score was 0.66. The optimal k -value for the KNN was 9 after testing k-values of 1-10 on the dataset.

2. **Random Forest Classifiers**

For the Random Forest model, the best F1-score was for the **finger**. This F1-score was 0.61. It was found that the optimal value for the Random Forest estimators was 100 after testing estimators ranging from 10 -100 on the dataset.

3. **Support Vector Machine**

For the SVM model, the best F1-score was for the **humerus**. This F1 score was 0.65. It was found that the optimal C value for the SVM model was 10 after hyperparameter tuning.C values ranging from 0.001 to 100, and kernels of rbf and gamma were tested on the dataset. The optimal C value, and kernel calculated on the dataset was 1 and rbf.

Table 2: Machine Learning Models compared using Classification Report

Model	Logistic Regression				Random Forest			
Body Type		Precision	Recall	F1-Score		Precision	Recall	F1-Score
Forearm	0 1	0.54 0.67	0.94 0.12	0.68 0.21	0 1	0.56 0.74	0.93 0.22	0.70 0.34
Wrist	0 1	0.59 0.43	0.83 0.19	0.69 0.26	0 1	0.68 0.70	0.88 0.40	0.77 0.51
Finger	0 1	0.64 0.76	0.87 0.46	0.74 0.57	0 1	0.70 0.78	0.85 0.59	0.76 0.67
Hand	0 1	0.61 0.60	0.98 0.05	0.75 0.08	0 1	0.63 1.00	1.00 0.09	0.77 0.17
Humerus	0 1	0.58 0.56	0.54 0.60	0.56 0.58	0 1	0.63 0.69	0.76 0.54	0.69 0.61
Shoulder	0 1	0.53 0.50	0.45 0.58	0.49 0.54	0 1	0.67 0.62	0.59 0.69	0.62 0.65
Elbow	0 1	0.60 0.62	0.95 0.12	0.73 0.20	0 1	0.63 0.87	0.98 0.20	0.77 0.32

Model	SVM				KNN			
Body Type		Precision	Recall	F1-Score		Precision	Recall	F1-Score
Forearm	0 1	0.54 0.73	0.96 0.12	0.69 0.21	0 1	0.56 0.59	0.77 0.36	0.65 0.45
Wrist	0 1	0.64 0.57	0.84 0.32	0.72 0.41	0 1	0.68 0.59	0.77 0.48	0.72 0.53
Finger	0 1	0.65 0.81	0.90 0.46	0.75 0.58	0 1	0.60 0.68	0.84 0.39	0.70 0.49
Hand	0 1	0.61 0.50	0.99 0.02	0.75 0.03	0 1	0.63 0.62	0.94 0.15	0.75 0.24
Humerus	0 1	0.65 0.65	0.66 0.64	0.66 0.65	0 1	0.66 0.63	0.60 0.69	0.63 0.66
Shoulder	0 1	0.58 0.54	0.49 0.62	0.53 0.58	0 1	0.62 0.58	0.55 0.65	0.58 0.61
Elbow	0 1	0.60 0.62	0.95 0.12	0.73 0.20	0 1	0.64 0.69	0.91 0.27	0.75 0.39

Discussion of Machine Learning Models

Amongst the four models assessed; the Random Forest Classifiers gives the best F1-score of 0.67 for the abnormal class. The best upper extremity body part predicted was the finger body part. All other models predict the humerus to have the best F1-score for the abnormal class.
The classification report for the abnormal class shows all models to offer poor prediction for the elbow, forearm, hand,wrist while offering better than average predictions for the humerus, finger and shoulder body parts i.e. F1-scores greater than 0.5 were obtained for the humerus, finger and shoulder body parts..

Conclusion

The problem undertaken in this paper is abnormality detection of bone X-ray. The approach implemented to solve this problem involved dividing the problem into three parts: image preprocessing, feature extraction and classification of images using machine learning models. The machine learning algorithms used for the evaluation were KNN, Random Forest, Logistic Regression and SVM. The performance evaluation of the abnormality detection in the MURA dataset was performed by using three statistical parameters such as recall, precision and F1 score.

Random Forest Classifiers for the X-ray images of fingers provide the best performance metrics when predicting the abnormal class for the upper extremity body part.

Recommendations for the Client

Based on the information provided by the models, one can recommend the client to use Random Forest Classifiers as the model of choice when trying to predict X-ray images for shoulder, humerus and fingers. The F1-scores for these body parts are greater than 0.6. The other models offer poorer predictions for the abnormal class. The F1-scores achieved for these body parts by the other models besides the Random Forest Classifier are less than 0.6.

The models associated with wrist, hand,forearm, and elbow currently have low performance and are therefore not recommended.

This is because the F1-scores achieved with these models are less than 0.5 for these body parts. This means besides the Random Forest models, there is a lower than average score obtained when using the machine learning models to predict these body parts

Future work

- Use deep learning algorithms such as Convolutional Neural Nets (CNN) to see if accuracy of model classification can be improved.

- Include more pre-processing steps with the images. These pre-processing steps could include adding masks to the images, and applying different transforms, to make the currently extremely varied data more uniform in contrast, orientation, and scale. This preprocessing step could make feature extraction simpler, and could improve the accuracy of models.

- Develop a digital data repository for radiologists to upload data securely. Results from these X-rays are then interpreted with a Random Forest model to see if X-rays are normal or not.

Works Cited

1. "What Is MURA?" *MURA Dataset: Towards Radiologist-Level Abnormality Detection in Musculoskeletal Radiographs*, stanfordmlgroup.github.io/competitions/mura/.

2. Khuzani, Abolfazl Zargari, et al. "COVID-Classifier: An Automated Machine Learning Model to Assist in the Diagnosis of COVID-19 Infection in Chest x-Ray Images." *MedRxiv : the Preprint Server for Health Sciences*, Cold Spring Harbor Laboratory, 18 May 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7273278/#S6.