Springboard Data Science Career Track

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

Introduction **3**

Objective **3**

Dataset **3**

Fig 1. Folder Arrangement of X-ray dataset 4

Data Cleaning & Wrangling **5**

Fig 2: X-ray Images of Upper Extremities for both Training and validation 5

Fig 3: Normal vs Abnormal X-ray Images for Training data 6

Fig 4: Normal vs Abnormal X-ray Images for Validation Data 6

Table 1: Distribution of Upper extremities for both training and validation data 6

Fig.5: Samples of X-ray images 7

Storytelling and Inferential Statistics **8**

Fig 6: Number of patients versus body parts for both training and validation data 8

Fig 7: Number of patients versus body parts for both Normal and Abnormal Data for Training data 9

Fig 8: Number of patients versus body parts for both Normal and Abnormal Data for Test data 10

Fig. 9: X-ray Studies for each Upper Extremity Body Type 11

Statistics **13**

Fig. 10. Scatter and Density plot for X-ray Images of the Humerus 13

Fig. 11. Heat map and Histogram representation of Pearson correlation Coefficients for the Humerus 15

Baseline Modeling **16**

Logistic regression 16

Extended Modeling **16**

K-Nearest Neighbors 16

Random Forest 16

Support Vector Machine 16

Table 2: Machine Learning Models compared using Classification Report 17

Discussion of Machine Learning Models **18**

Conclusion **18**

Future work **18**

Introduction

In many areas of the world, radiologists interpret radiographs visually to determine whether there are abnormal defects in bone structures.Sometimes this interpretation can be prone to error, and other times the interpretation of these data can be limited by the presence of trained radiologists.

The goal of the project is to develop a model that would help a clinical assistant working with a radiologist, to be able to take a previously unseen X-ray image and link it directly to previous cases. By doing this, a clinical assistant can offer suggestions on possible treatment. In addition to this advantage, the model can also be used in parts of the world where access to skilled radiologists is limited.

It should however be stated that this model is not to be used as a replacement for a qualified radiologist. It should only serve as a check to verify that the radiologist has identified abnormalities in the hand bone.

Objective

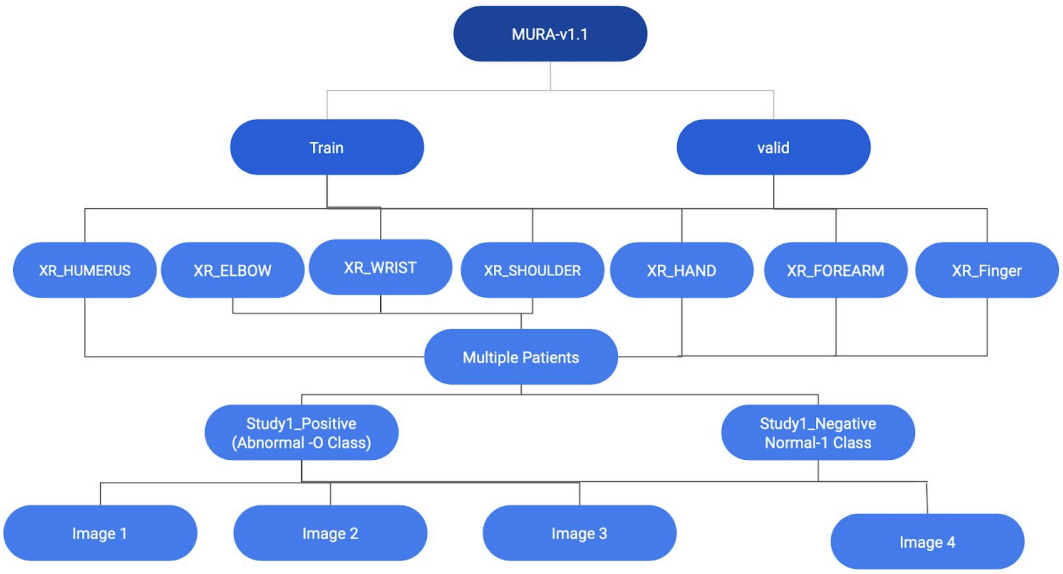
The goal of my project is to utilize machine learning techniques to automatically detect fractures in X-ray images of upper extremities body parts. In this capstone report , a binary classification model will be developed to distinguish between normal and abnormal X-ray images. The machine learning models used in this capstone report are Logistic Regression , KNN, Random Forest and SVM algorithms.

Dataset

The dataset of bone X-rays, is a public dataset provided by Stanford[1]. The dataset is called MURA(**MU**sculoskeletal **RA**diograph). The MURA dataset supplied by Stanford is in multiple folders divided into training and validation dataset. These dataset are further subdivided into separate folders of 7 upper extremities body parts. These upper extremities are 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 with no defect, while the positive X-rays have defects 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



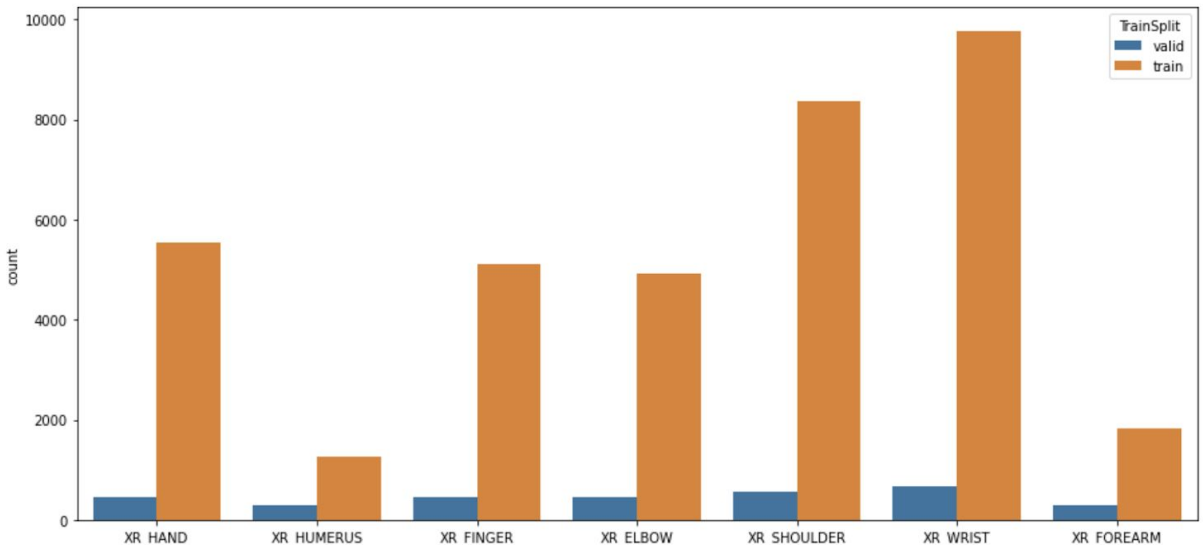
Data Cleaning & Wrangling

Before beginning analysis of the data, it is essential to explore the type of data. 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 upper extremities in the supplied dataset is displayed in Fig 2.

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



In total there is a total of 8,280 X-ray images, classified as zero (0) class for the normal X-rays, and 5177 abnormal X-rays, classified as the one (1) class for the training dataset i.e. 62 % normal data and 38 % abnormal data for the training dataset, while for the validation dataset there is a total of 661 X-rays for the normal dataset (0 class), and a total of 538 (1 class) for the abnormal dataset. The ratio here is 55% normal data and 45% abnormal data. 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

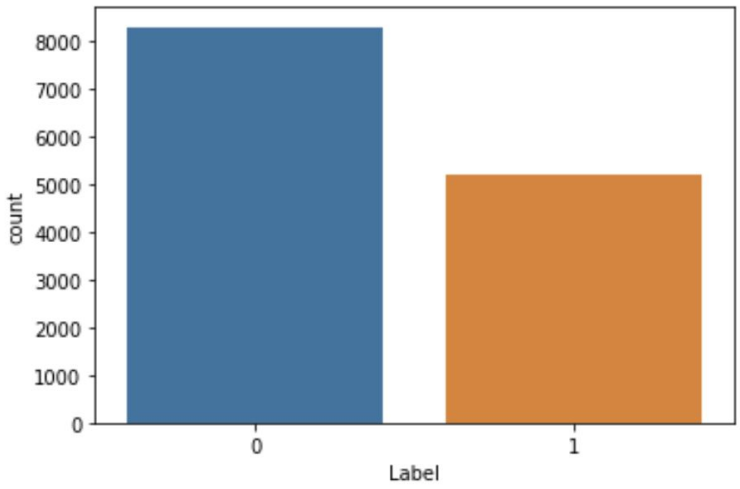


Fig 4: Normal vs Abnormal X-ray Images for Validation Data

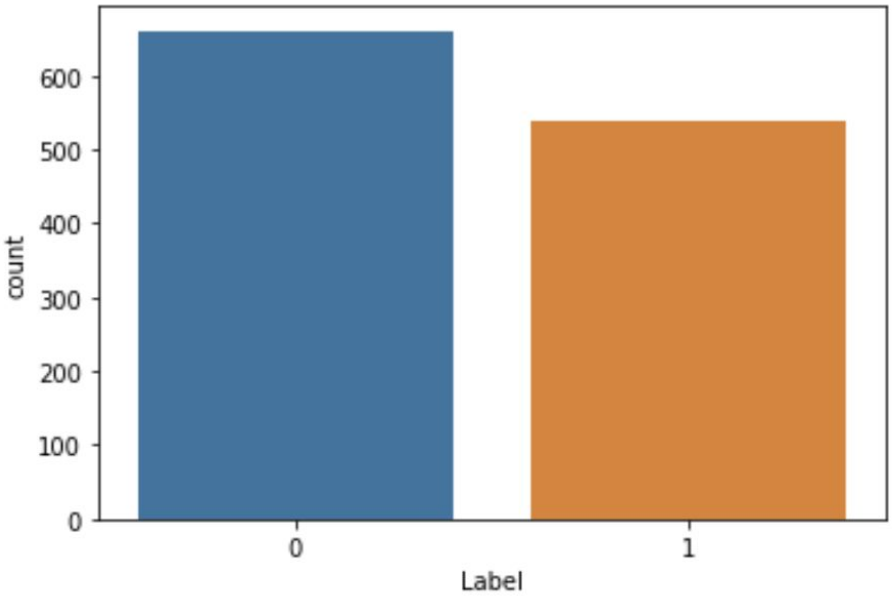
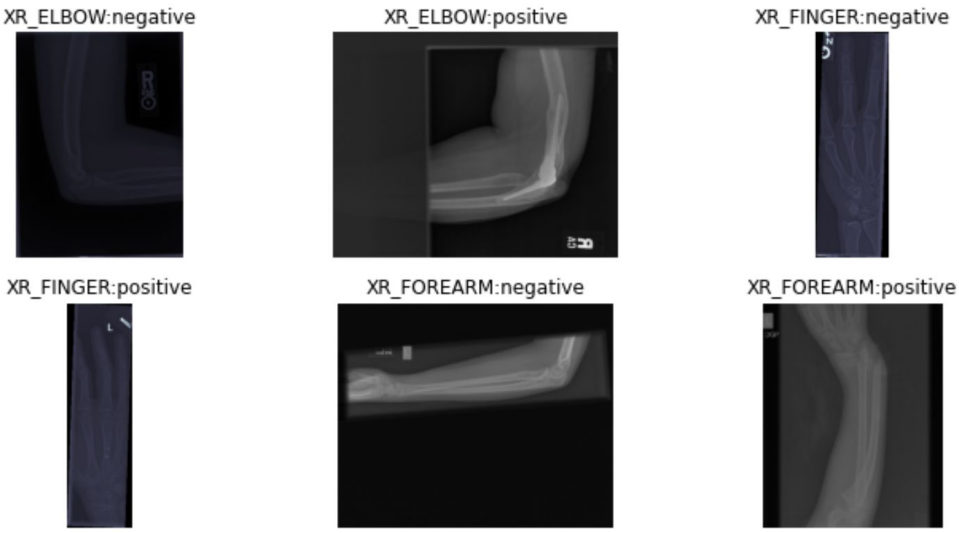


Table 1: Distribution of Upper extremities for both training and validation data

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 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 need to be of appropriate sizes so that not too much information is lost in the reduction of the size, when inputting it 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 modeling is called image preprocessing. A pixel size of 224 x 224 was chosen for the image preprocessing, because this was the size used in the Stanford article[1].

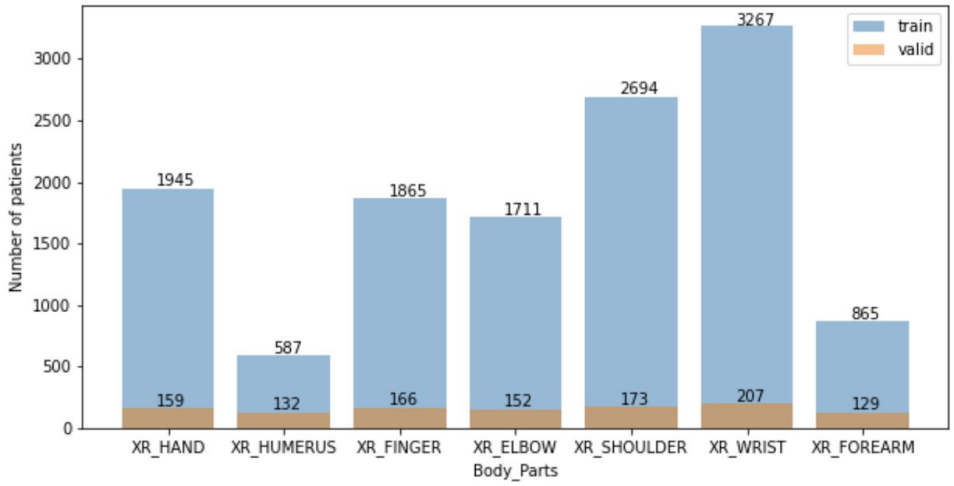
Storytelling and Inferential Statistics

Following the data preprocessing of the data, the next step is to interrogate the dataset, extract features of the dataset and also ask a series of questions of the dataset. These questions will help identify the contributing factors affecting the X-rays for the upper extremities.

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 conducted on patients with wrist incidents. These had the most training and validation datas. The X-rays with the least observation for the training dataset was that of the humerus, while the X-rays with the least validation dataset was that of the forearm. This result shows that most of the data collected were from regions where individuals had more wrist damage.

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

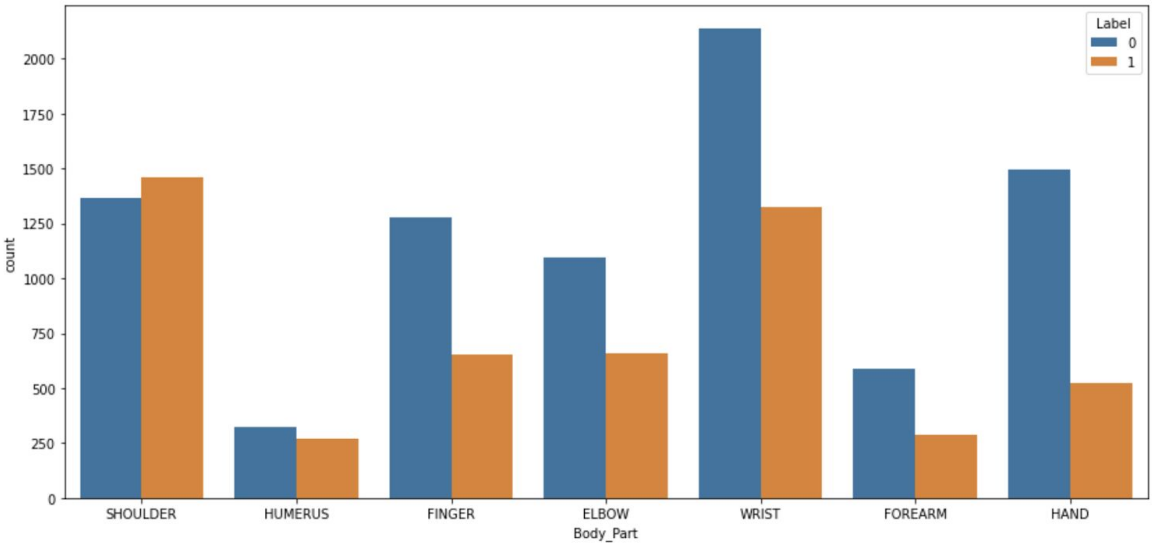
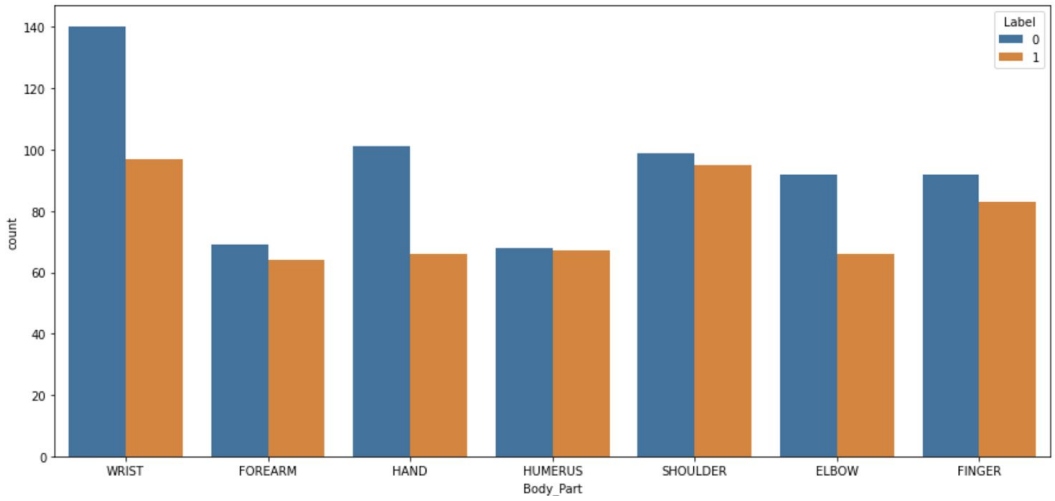


Fig 8: Number of patients versus body parts for both Normal and Abnormal Data for Test data



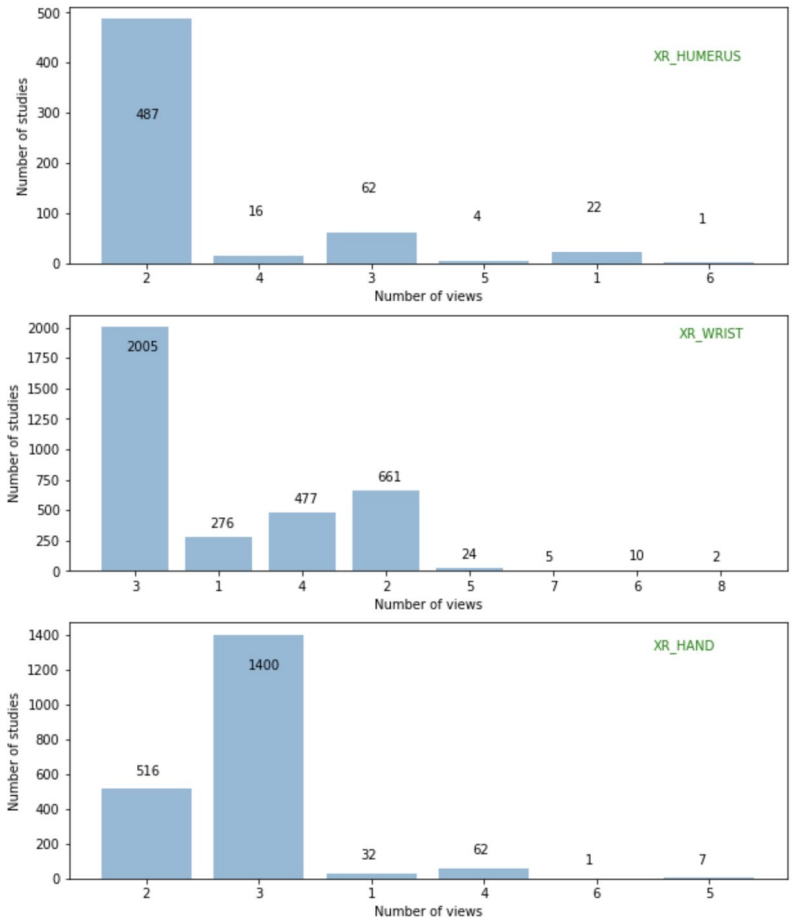
From the bar chart in Fig. 7 and 8, it can be observed that most X-rays were conducted on patients with wrist incidents for the normal and the abnormal dataset. The training and the validation dataset followed the same distribution for the body parts i.e. The most X-ray observation was for the wrist,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

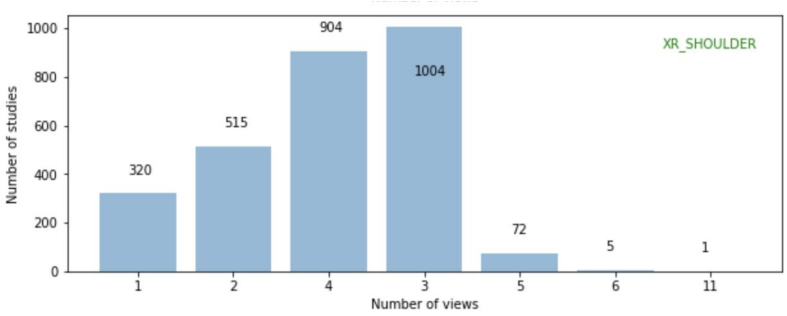
The third question asked:

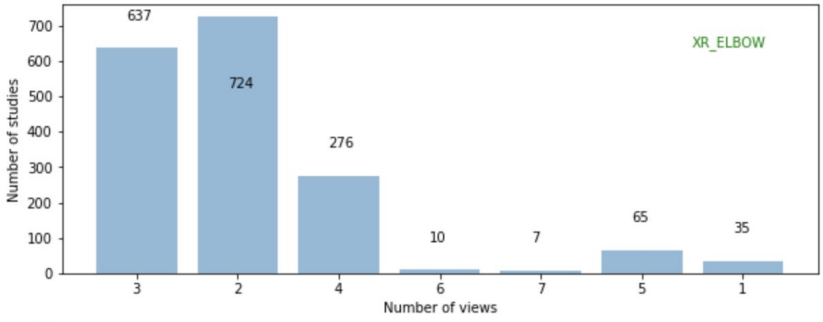
Which upper extremity part had most X-ray views?

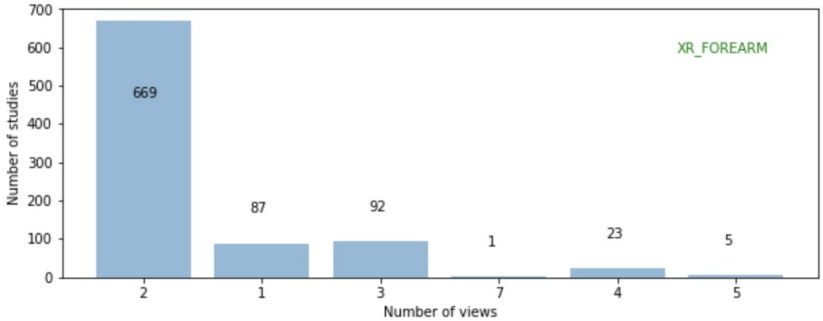
This question is being asked because some body parts might require more views by the radiologist to make a better diagnosis of the patient. The idea that multiple views are expressed for each upper extremity part. Fig. 9 shows that most studies have mostly 2,3 and 4 images.The maximum number of images is the XR_Shoulder. A patient had 13 images for one study type for the same shoulder.

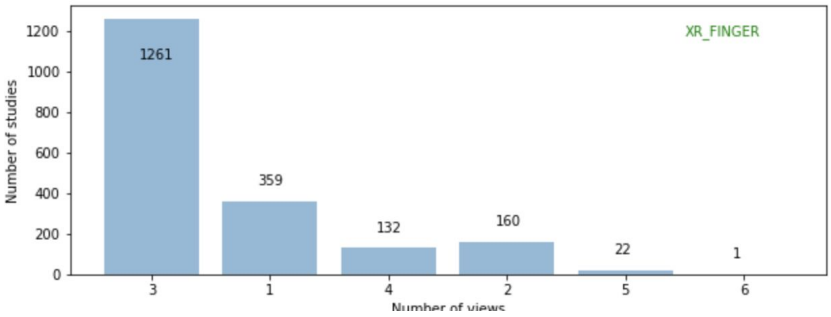
Fig. 9: X-ray Studies for each Upper Extremity Body Type




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

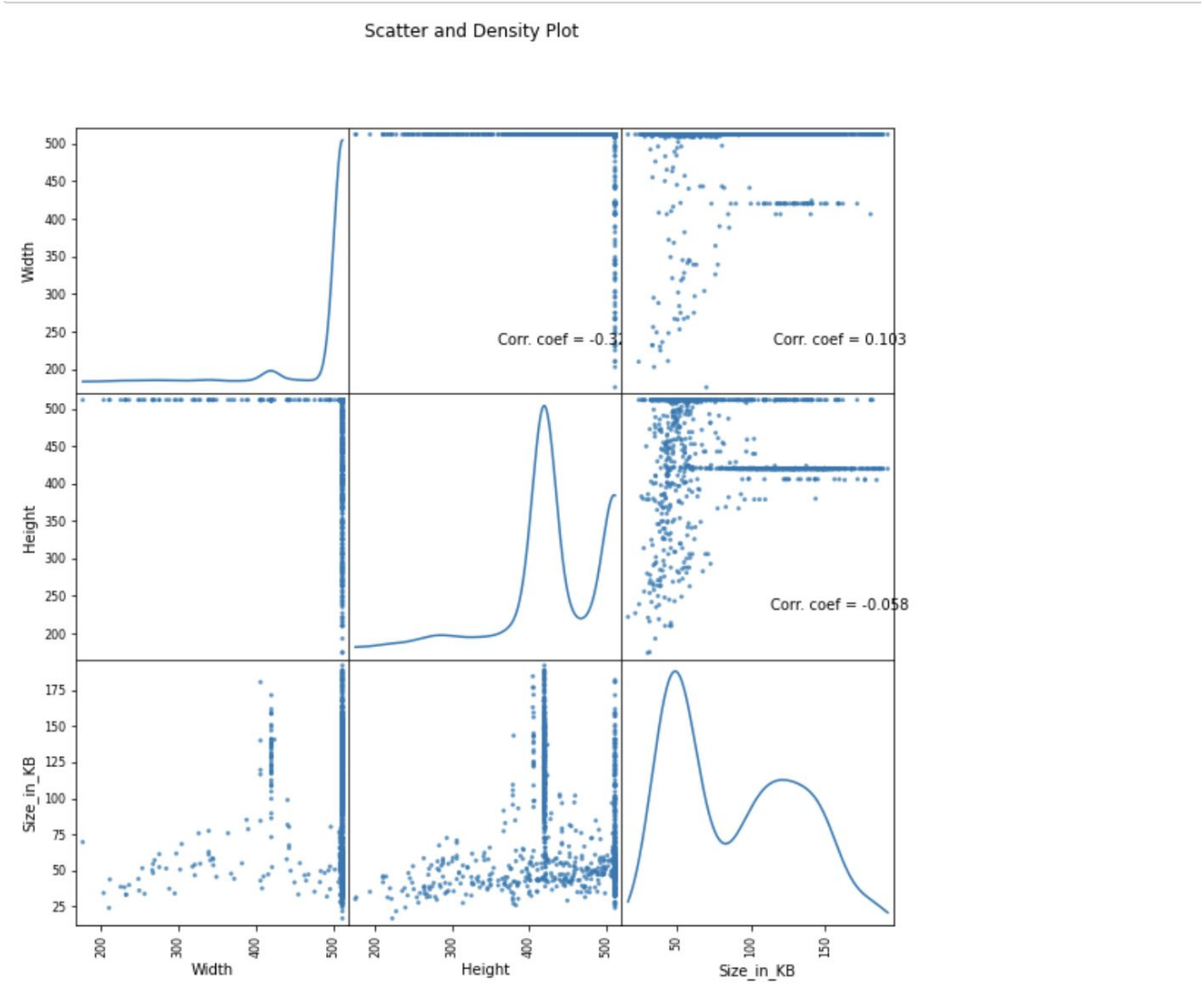

XR_FOREARM
Number of studies
Number of views
669
87
92
1
23
5
2
1
3
7
4
5


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

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

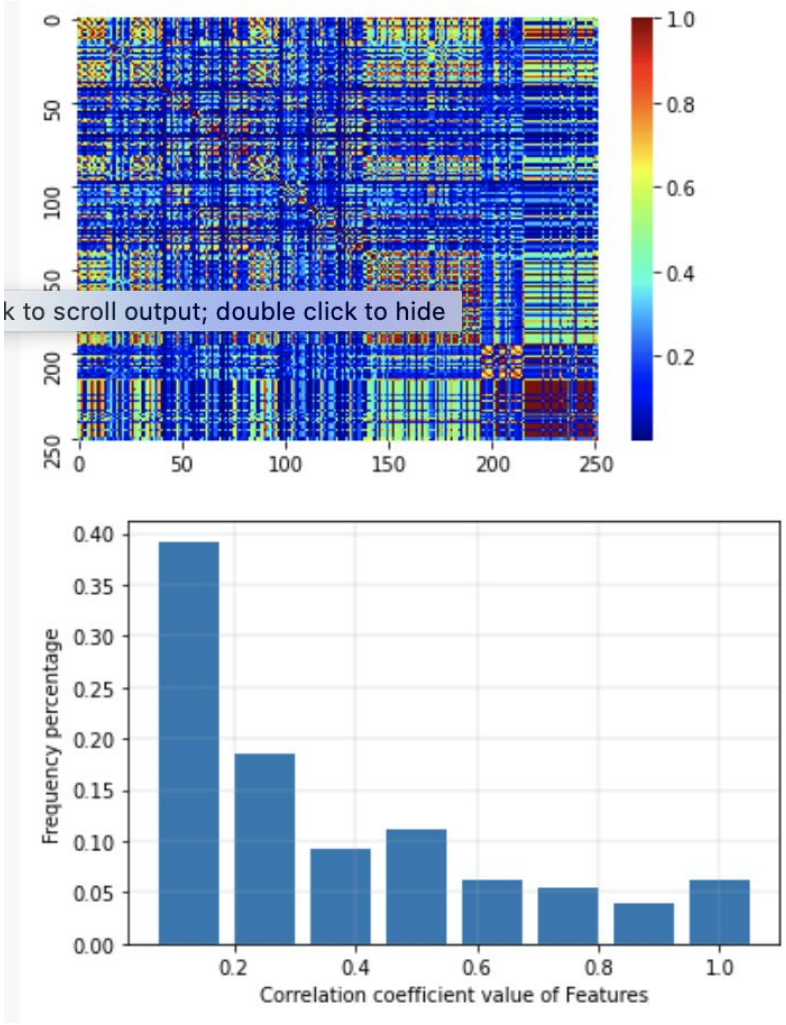
Feature Extraction

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,MeanGradient, and StdGradient.

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 shows that most of the features have correlation coefficients of less than 0.4.

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 one.

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 (RF). The training dataset is tuned with hyperparameters.

1. **K-Nearest Neighbors**

For K-nearest neighbors, the best F1-score was for the **humerus** body part. 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**

For the random forest model, the best F1-score was for the **finger** body part. 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** body part. 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 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 that the model offers poor predictors on elbow, forearm, hand for all the models, while offering good predictors for the humerus and shoulder body parts i.e. F1 scores greater than 0.6.

Based on these results, one can recommend the client to use random forest as the model of choice when trying to predict X-ray images for fingers.

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 four statistical parameters such as recall, accuracy, precision and F1 score.

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

Future work

Develop a website where a radiologist could upload his X-ray images to. This website would use the random forest model to help determine abnormal versus normal images.

Further future work that could be carried out could involve using deep learning algorithms such as CNN to include more pre-processing steps, such as 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 would make feature extraction simpler, and would consequently significantly improve the accuracy of our models across the board.

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.