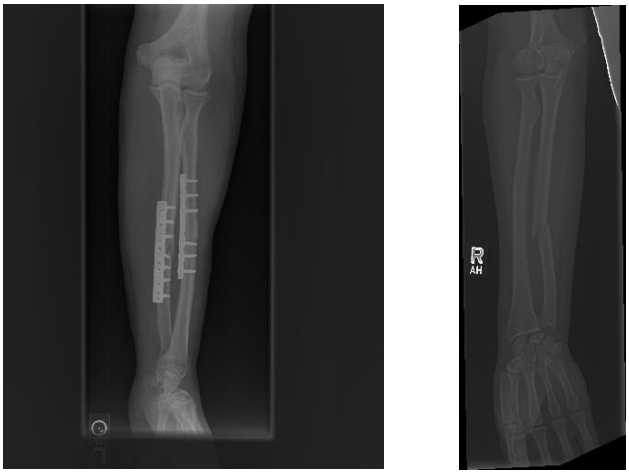
Abnormality Detection in Musculoskeletal Radiographs

Submitted By:
Asif Sohail Mohammed - AXM190041
Pragya Nagpal - PXN190012

Dataset Description

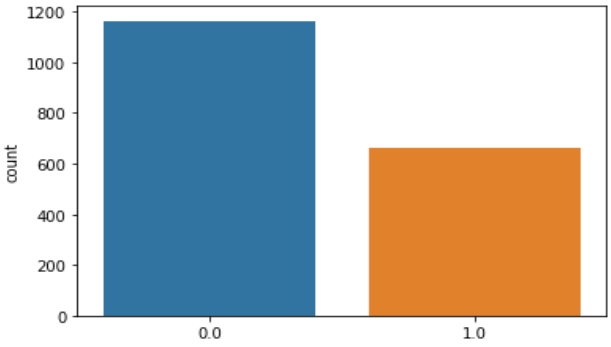
Our project is based on the MURA dataset which is a large dataset of musculoskeletal radiographs containing 40,561 images from 14,863 studies, where each study is manually labelled by radiologists as either normal or abnormal. The dataset has images for abnormalities in elbow, forearm, hand, humerus, shoulder, finger and wrist. However, for not working with a very large number of images we have only considered the X-ray for the forearm images of 1010 patients with a total of 2126 images given as train and test separately. We aim at classifying the forearm X-ray images as normal (0) and abnormal (1).



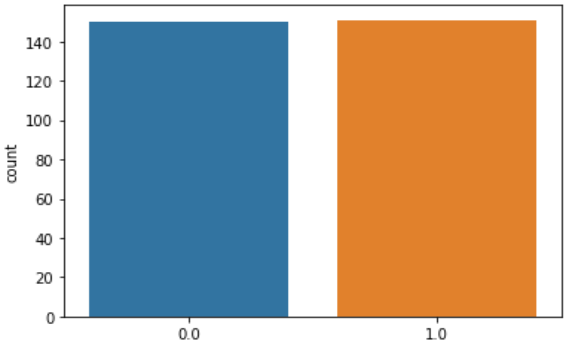
Abnormal X-ray Images

Normal X-ray Image

We have the training dataset of 1825 images of 877 patients and testing dataset of 301 images of 133 patients.



Train Dataset Image count



Test Dataset Image count

Preprocessing

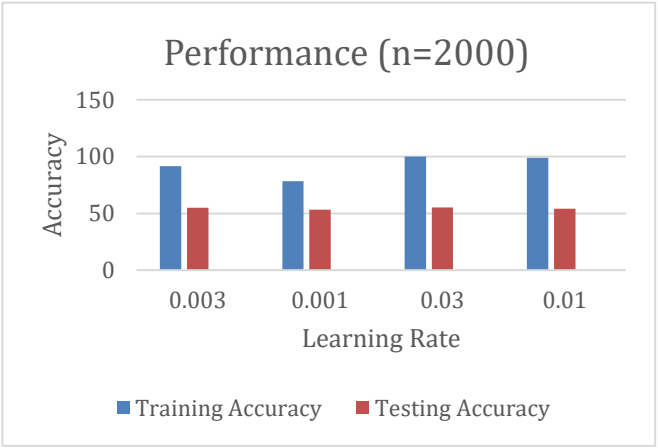
We did some preprocessing to reduce the complexity and enhance the image of the applied algorithm. We used a filter to sharpen the images.

-1	-1	-1
-1	9	-1
-1	-1	-1

Algorithms used

Logistic Regression

Since we are working on a classification problem we started with a simple logistic regression algorithm by writing the code from scratch.
We evaluated our approach for 2000 iterations with different learning rates (0.003, 0.01, 0.03, 0.01).



The best accuracy was obtained for logistic regression at the learning rate= 0.03.

Results Generated

Train accuracy = 99.94
Test accuracy = 55.14
Confusion Matrix:

115	35
100	51

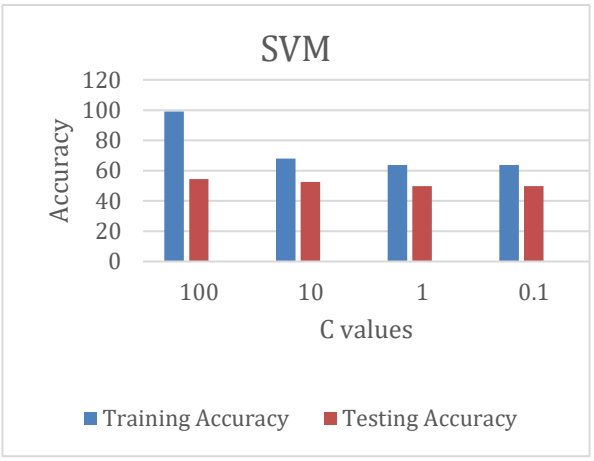
Precision = 0.5930232558139535
Recall = 0.33774834437086093
F1 score = 0.430379746835443

True Positive Rate (TPR) = 0.5348837209302325
False Positive Rate (FPR) = 0.4069767441860465

In logistic Regression, we observed a significant difference in the training and test accuracy and it overfits leading to high variance. Henceforth we used SVM.

Support Vector Machines

Implemented the support vector machine algorithm by using the Sklearn library. Model was tuned on different C values (0.1, 1, 10, 100). The figure alongside shows how the training and test accuracy vary with different C values. It performs the best when C=100.



Results Generated

Training Accuracy = 0.9912328767123287
Testing Accuracy = 0.574750830564784

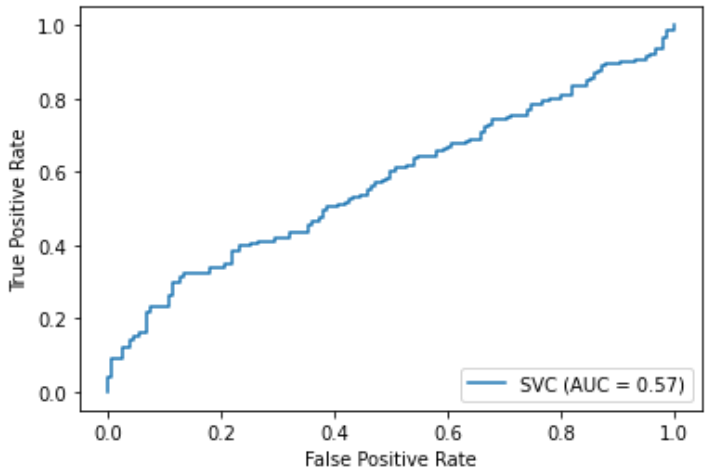
Confusion Matrix:

124	26
102	49

Precision = 0.6533333333333333
Recall = 0.32450331125827814
F1 score = 0.4336283185840708

True Positive Rate (TPR) = 0.5486725663716814
False Positive Rate (FPR) = 0.3466666666666667

ROC Curve for SVM



Bagging

Due to observed high variance for SVM, bagging was done using the Sklearn library. The base estimator was SVM and the number of estimators was 10 and 20.

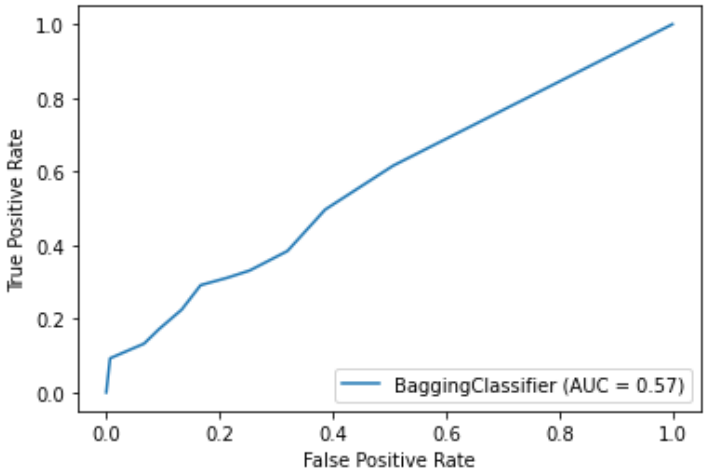
Results Generated

Training Accuracy = 0.943013698630137
Testing Accuracy = 0.5614617940199336
Confusion Matrix:

125	25
107	44

Precision = 0.6376811594202898
Recall = 0.2913907284768212
F1 score = 0.39999999999999997
True Positive Rate (TPR) = 0.5387931034482759
False Positive Rate (FPR) = 0.36231884057971014

ROC Curve for Bagging



Gradient Boosting

As we have observed high variance with previous experiments, we experimented with another form of boosting, gradient boosting with learning rate = [0.1, 0.2, 0.25] and n_estimators = [30, 50, 70] using Sklearn.
We obtained the best accuracy on learning rate=0.2 and n_estimators =50.

Results Generated

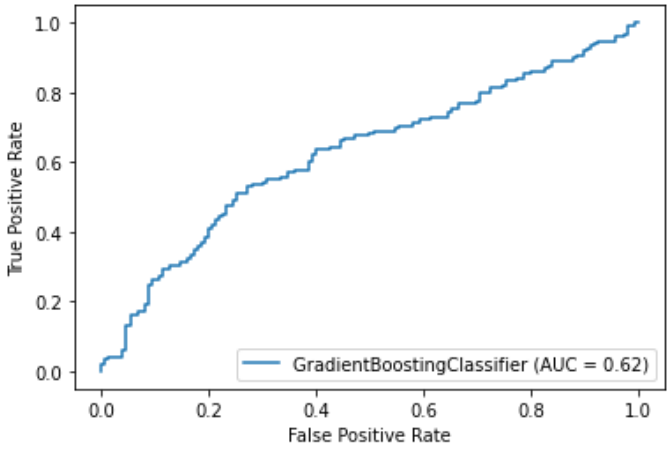
Training Accuracy = 0.9523287671232876
Testing Accuracy = 0.5813953488372093
Confusion Matrix:

126	24
102	49

Precision = 0.6712328767123288
Recall = 0.32450331125827814
F1 score = 0.4375

True Positive Rate(TPR) = 0.5526315789473685
False Positive Rate(FPR) = 0.3287671232876712

ROC Curve for Gradient Boosting



Convolutional Neural Network

We used CNN since it works best with images, we implemented the CNN architecture similar to VGG16 having a convolution layer of 3x3 filter with a stride of 1 and a maxpool layer of 2x2 filter. It follows this pattern of convolutional and maxpool layers consistently and ends with 3 fully connected layers with a relu output with a total of 16 layers.
To increase the number of images to be fed to CNN, we performed Image Augmentation with rotation range 90 and horizontal flip.
We obtained the best accuracy on optimizer nadam with softmax activation.

Results Generated

Training Accuracy: 0.645
Testing Accuracy: 0.608
Confusion Matrix:

128	22
113	38

Precision is 0.538961038961039
Recall = 0.5496688741721855
F1 score = 0.5442622950819672

True Positive Rate (TPR) = 0.5311203319502
False Positive Rate (FPR) = 0.366666666666

Drawbacks in dataset

The labels are dependent on the patient and not the X-ray image. This means that for the same patient, multiple studies generate X-rays from multiple angles. In some X-rays, the abnormality is not visible but it's still labelled as abnormal because it has an abnormality which can be observed in other X-rays. Figure alongside shows the number of patients vs the number of X-ray images each person has.
We think that this problem can be solved by inducing some form of relationship between multiple X-rays of any given patient.