Abnormality Detection in Bone X-Rays using Deep Learning

Mahesh Latnekar
Indiana University
Bloomington, Indiana
mrlatnek@iu.edu

Mayank Kumar Raunak
Indiana University
Bloomington, Indiana
mraunak@iu.edu

Parth Naik
Indiana University
Bloomington, Indiana
naikpa@iu.edu

ABSTRACT

Bone X-Rays are widely used for detecting and localizing abnormalities in various body parts such as shoulder, humerus, elbow, finger, forearm, wrist and hand. The automatic detection of abnormalities with accuracy better or equivalent to human radiologists' performance could be a tremendous improvement in the diagnosis process. Automatic detection of abnormal musculoskeletal conditions can help radiologists select and prioritize the affected people. This is particularly important in underdeveloped and developing countries which have a scarcity of experienced radiologists. Computer-aided detection can provide reliable and informed results for a large volume of cases in very less time.

We got the idea of this project from an ongoing Bone X-Ray Deep Learning competition on the topic, 'Abnormality Detection in Musculoskeletal Radiographs'. For our work, we collected data from https://stanfordmlgroup.github.io/competitions/mura/. MURA is a large data set of musculoskeletal radiographs containing 40,561 images from 14,863 studies. Each study was labeled by radiologists manually as abnormal or normal. We train a VGG16 model, VGG19 model and an InceptionV3 model to detect abnormalities. Our best model (VGG16) achieved AUROC of 0.91 with sensitivity 0.88.

The model performance is compared with that of the radiologists on the Cohen's Kappa statistic, which expresses the agreement of our model and of each radiologist with the gold standard.

Keywords

Abnormality Detection, X-Rays, Deep Learning, Convolutional Neural Networks, VGG16, VGG19, InceptionV3

1. INTRODUCTION

Determining whether a radiographic study is normal or abnormal is a critical radiological task: a study interpreted as normal rules out abnormality and can eliminate the need for patients to undergo further diagnostic procedures or interventions. The musculoskeletal abnormality detection task is particularly critical as more than 1.7 billion people are affected by musculoskeletal conditions worldwide (BMU, 2017). These conditions are the most common cause of severe, long-term pain and disability (Woolf & Pflieger, 2003), with 30 million emergency department visits annually and increasing. Our dataset, MURA, contains 9,045 normal and 5,818 abnormal musculoskeletal radiographic studies of the upper extremity including the shoulder, humerus, elbow, forearm, wrist, hand, and finger. MURA is one of the largest public radiographic image datasets.

We find that the model performance mentioned in the MURA paper (Rajpurkar & Irvin et al., 17) was comparable to the best radiologist's performance in detecting abnormalities on finger and wrist studies. Therefore we went ahead with elbow studies and achieved a higher kappa score than the model used in the paper and comparable to two of the three radiologists using our modified VGG16 architecture.

2. DATASET

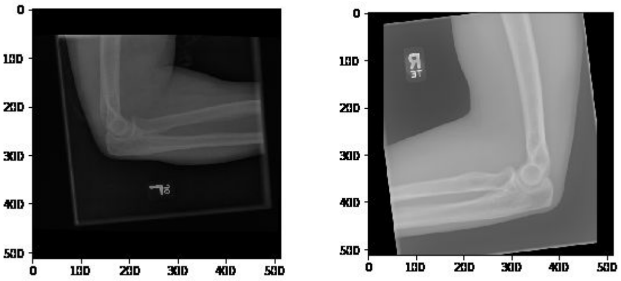
We have selected the x-ray images of elbows from the MURA data set for our analysis and predictions.

Train set : There were around 4931 images with corresponding labels of abnormality (1) and normality (0), manually labeled by the radiologists

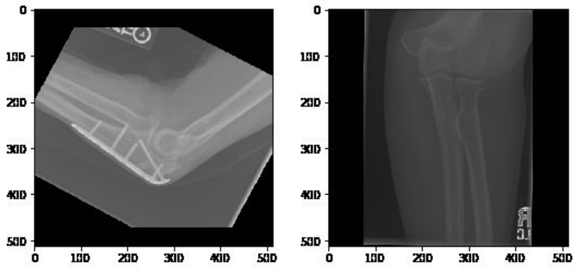
Test-set-collection: There were 400 images with their corresponding labels.

Abnormalities Types: Authors of Paper (Rajpurkar & Irvin et al., 17) reviewed the radiologist reports to manually label 100 abnormal studies with the abnormality finding: 53 studies were labeled with fractures, 48 with hardware, 35 with degenerative joint diseases, and 29 with other miscellaneous abnormalities, including lesions and subluxations.

Examples of normal elbow x-ray images



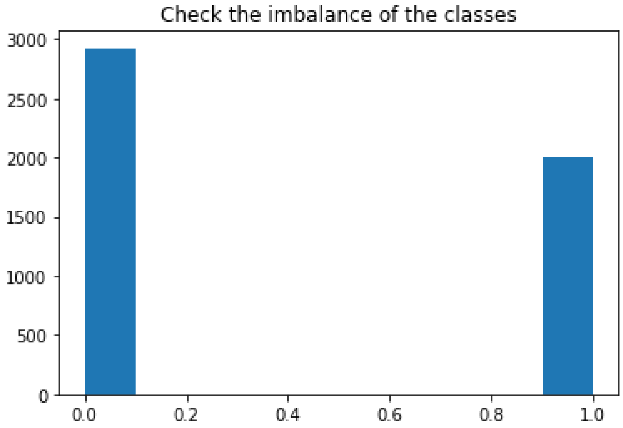
Examples of abnormal elbow x-ray images



The dataset is slightly imbalanced containing 2925 normal elbow x-rays and 2006 abnormal elbow x-rays. We hence calculate the class weights of the normal and abnormal classes to prevent bias in our models.

Class 0 (normal) weight = 0.685

Class 1 (abnormal) weight = 1.458



3. RELATED WORK

Extensive research has been conducted on using Deep Learning models for medical applications. Much of this research is focused on processing medical images. Convolutional neural networks have proved to be state of the art algorithms in medical image processing in many applications. In [14], Lee et. al. (2015) proved that Deep ConvNet architectures outperform non-neural approaches like AdaBoost in finger joint detection. Transfer Learning has proved to be an effective way to enhance the performance of Deep ConvNets due to the scarcity of data. In [10], Lakhani et. al. implemented an ensemble of AlexNet and GoogLeNet pre-trained on ImageNet for automated classification of pulmonary tuberculosis to get an AUC of 0.99. Ausawalaithong et al.[8] (August 2018) used 121-layer DenseNet pre-trained on ImageNet for automatic lung cancer prediction. Due to limitations of clinical approaches, ranging from operator variability of radiologists to the lack of practitioners in underdeveloped and developing countries, Deep Learning models have been used on X-Ray images(radiographs) for automated assessment. In [7], Rajpurkar et. al (2017) build a CNN model called CheXNet to train on Chest X-ray 14, the largest publicly available chest X-ray dataset. Fos-Guarinos B. et. al. used Chest X-ray images from Indiana University's Open-I collection for automated chest X-ray screening. Many researchers have trained Convolutional Neural Networks on bone X-ray images. In [13], Spampinato et. al. (2017) used deep learning for investigating the generalization capabilities of deep-learning models trained on general imagery. They performed a skeletal bone age assessment on pre-trained networks like GoogLeNet, OxfordNet and a model they built trained from scratch on X-ray scan dataset called BoNet. Lee et. al used CaffeNet, pretrained on ImageNet for Bone Age Estimation. In [12], Kaloi and He (2018) utilized CNNs on hand radiographs for child gender determination. Yune et. al. (2018) used ImageNet Inception network for gender identification from hand and wrist radiographs of children. In particular, VGG networks have found applications in classification and detection in radiographs. VggNet was the best performing architecture in the research conducted by Xue et. al.[11] (2018), for detecting gender in chest radiographs. VGG-19 architecture pretrained on ImageNet was implemented by Bradshaw et. al.[15] for classification of benign and malignant bone lesions.

4. MODELS

4.1 Data Pipeline

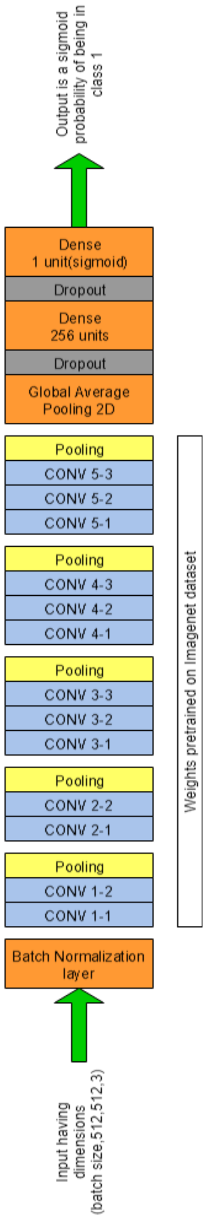
For the purpose of training and validation, we created data pipelines using Tensorflow. The data pipelines pre-fetched the images from the disk, decoded the .png images into numpy arrays, performed random horizontal flipping on the images and random rotation by 0 to 0.5 radians in clockwise or anticlockwise directions to make our models robust towards horizontal flips and the slight random rotations encountered while taking x-rays.

4.2 VGG16

The first model we tried out was VGG16 due to its robustness to noise(Grm & Struc et al. 17). The model takes as input one or more views for a study of an upper extremity. On each view, our VGG16 convolutional neural network predicts the probability of abnormality. We compute the overall probability of abnormality for the study by taking the arithmetic mean of the abnormality probabilities output by the network for each image. The model makes the binary prediction using the sigmoid function. If the output of sigmoid is less than 0.5, then it predicts abnormal otherwise it predicts normal.

4.2.1 VGG16 modifications

The VGG16 model was modified by us in terms of the architecture. We used the convolutional base of VGG16 pre-trained on the ImageNet dataset. We added a batch normalization layer as the first layer before the convolutional base to reduce internal covariate shifts and the dependence of the gradients on the scale of their initial values (https://gist.github.com/shagunsodhani/4441216a298df0fe6ab0). Furthermore after the convolutional base instead of flattening the convolutional output we used global average pooling in the 2D space which helped us generalize and reduced the dimensionality because as opposed to ImageNet which has 1000 classes our problem has only 2 classes and hence the next 2 dense layers need to reduce the dimensionality to a great extent if we directly flatten the layers instead of using global average pooling. We also added dropout of 30% between the first and second layers after the convolutional base and between the second dense layer and the final output sigmoid layer.



The above figure shows the modified VGG16 architecture with the inputs and outputs (green), the orange and the grey blocks are our modifications to the architecture.

4.3 INCEPTION V3

InceptionV3 model was used as it has certain advantages over naïve inception models, first, it reduced representational bottleneck. The intuition was that neural networks perform better when convolutions didn't alter the dimensions of the input drastically. Reducing the dimensions too much may cause loss of information, known as a "representational bottleneck"

Using smart factorization methods, convolutions can be made more efficient in terms of computational complexity. It is also incorporated with Label smoothing—A type of regularizing component added to loss function that prevents the network from becoming too confident about a class.

The architectural modifications were similar to that of VGG16.

Figure showing one block of the naive Inception arc

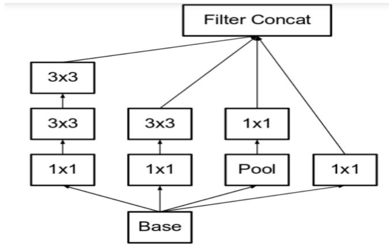


Figure showing one block of the InceptionV3 architecture with factorization

5. TRAINING THE MODELS

5.1 Loss function

Our problem is a binary classification problem, therefore, the output layer of the models has one unit which gives the probability that an image belongs to class 1 (abnormal). Therefore the suitable loss function is the negative log loss function. However we have a slight imbalance in our classes, hence we weight the classes with suitable weights to avoid our models getting biased. As a result, our loss function is the weighted negative log loss

$$L(X,y) \; = \; -w_1 * y_{true} log(y_{predicted}) \; -w_0 * (1-y_{true}) log(1-y_{predicted})$$

In the above equation y_{true} is the true label of the image, $y_{predicted}$ is the predicted label of the image and w_0, w_1 are the class weights of the normal (0) and abnormal (1) classes.

5.2 Optimization techniques

The network was trained end-to-end using Adam with default parameters β1 = 0.9 and β2 = 0.999 . We trained the model using mini-batches of size 10. We used an initial learning rate of 0.0001 that is decayed by a factor of 10 each time the validation loss plateaus after an epoch.

4.2.1 Fine Tuning

Transfer Learning*:* We used the pre-trained weights for our models trained on the ImageNet dataset. To finetune these networks we first freezed the convolutional base of the models and trained only our dense layers using a considerable initial learning rate of 1e-04, we freezed the convolutional base to prevent destroying the features of the pre-trained weights due to large initial gradients.

We don't unfreeze the bottom (initial) blocks of the convolutional base as the lower layers are supposed to have learnt some very generic features.

When we observed no further increase in validation accuracy and decrease in the validation loss, we unfreeze the top 3 blocks of the convolutional base and then optimize the network using a smaller initial learning rate of 1e-05 to prevent modifying the pre-features by much but at the same time modify them to be somewhat specific to our problem.

Used Early Stopping: As training data was small, deep convolutional networks tend to overfit. Hence, stopping the training as soon as performance on a validation set starts to degrade helps to combat overfitting.

Used decayed learning rate: We are using mini-batch which contains different noises, during the initial phase when the learning rate is still large, the movement towards minima is faster and when the loss function reaches near minima, smaller learning rates/steps allows to converge or oscillates around the tighter region around local minima, keeping the probability of overshooting low.

$$\alpha = \alpha_0/(1 + \textit{decay rate} * \textit{epoch} - \textit{num})$$

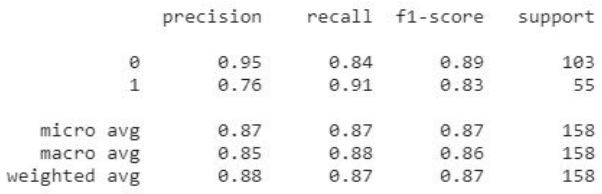
Global Average pooling: It was another technique applied to improve the performance of our models. One advantage of global average pooling over the fully connected layers is that it is more like the convolution structure by enforcing correspondences between feature maps and categories. Thus, the feature maps can be easily interpreted as categories confidence maps. Another advantage is that there is no parameter to optimize in the global average pooling thus overfitting is avoided at this layer. Furthermore, global average pooling sums out the spatial information, thus it is more robust to spatial translations of the input.

6. RESULTS

In order to measure the model performance, we present a classification report along with the kappa score which is used in the MURA paper (Rajpurkar & Irvin et al., 17) to compare the performance of the human radiologists and models.

6.1 VGG16

The VGG16 model is our best model. Apart from achieving a good classification report, it obtained a kappa score of 0.7201 which is better than the paper baseline mode (Rajpurkar & Irvin et al., 17), also the kappa score was as good as the kappa score of two of the human radiologists.

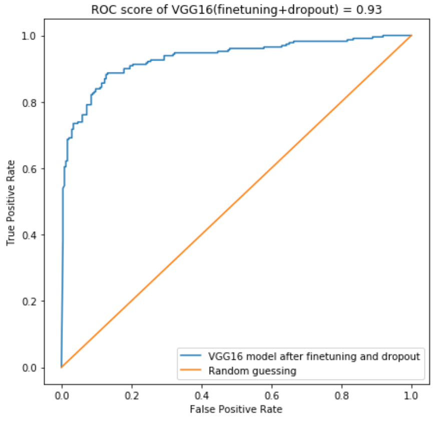


```
              precision    recall  f1-score   support

           0       0.95      0.84      0.89       103
           1       0.76      0.91      0.83        55

   micro avg       0.87      0.87      0.87       158
   macro avg       0.85      0.88      0.86       158
weighted avg       0.88      0.87      0.87       158
```

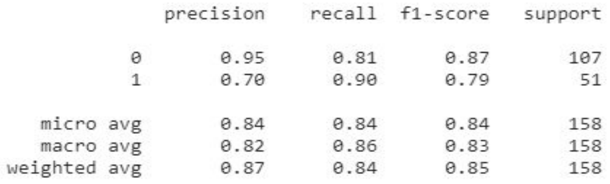
Classification report of our VGG16 model



The ROC curve and the AUROC score (0.93) achieved by our VGG16 model

6.2 VGG19

VGG19 was our second best model achieving a kappa score of 0.66. Its obtained kappa score was less than all of the radiologists and the paper baseline model.

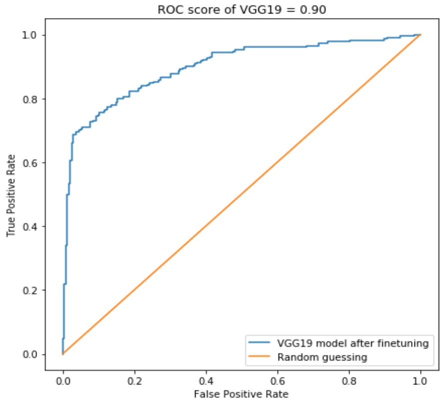


```
              precision    recall  f1-score   support

           0       0.95      0.81      0.87       107
           1       0.70      0.90      0.79        51

   micro avg       0.84      0.84      0.84       158
   macro avg       0.82      0.86      0.83       158
weighted avg       0.87      0.84      0.85       158
```

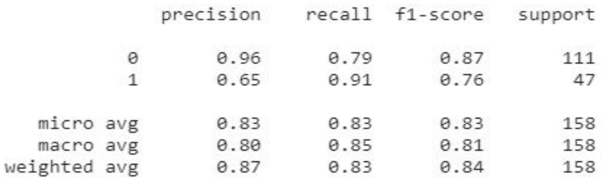
Classification report for our VGG19 model



The ROC curve and the AUROC score (0.90) achieved by our VGG19 model

6.3 InceptionV3

The inception model was our worst model obtaining a kappa score of 0.63, which is less than all of the radiologists and the paper baseline model.

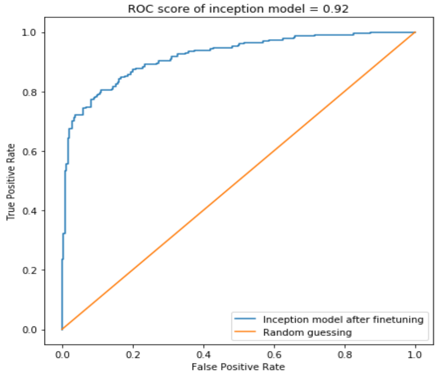


```
              precision    recall  f1-score   support

           0       0.96      0.79      0.87       111
           1       0.65      0.91      0.76        47

   micro avg       0.83      0.83      0.83       158
   macro avg       0.80      0.85      0.81       158
weighted avg       0.87      0.83      0.84       158
```

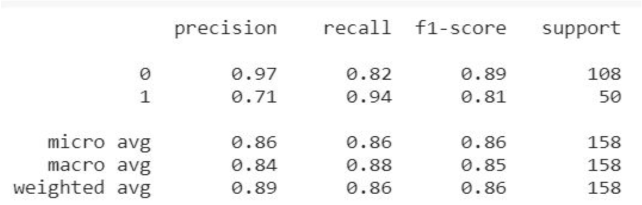
Classification report of our Inception model



The ROC curve and the AUROC score (0.92) achieved by our Inception model

6.4 Ensemble Modelling

Implemented a weighted ensemble model by combining all three models(VGG16, VGG19, InceptionV3) obtaining a kappa score of 0.7036 which is close to the score (0.710) obtained on the model in the paper(Rajpurkar & Irvin et al., 17) and one of the Radiologists.



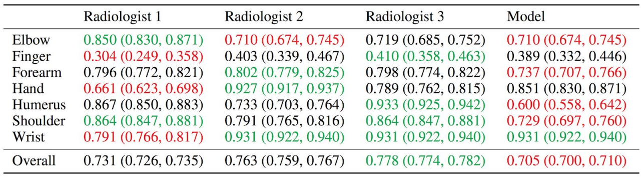
```
              precision    recall  f1-score   support

           0       0.97      0.82      0.89       108
           1       0.71      0.94      0.81        50

   micro avg       0.86      0.86      0.86       158
   macro avg       0.84      0.88      0.85       158
weighted avg       0.89      0.86      0.86       158
```

7. CONCLUSION

We tried out 3 different convolutional models (VGG16, VGG19, InceptionV3) using transfer learning approach. Of the three models we achieved the best results using VGG16 model which gave us a kappa score of 0.7201. This score was higher than the MURA paper baseline model. The model score was also as good as two of the three human radiologists.



	Radiologist 1	Radiologist 2	Radiologist 3	Model
Elbow	0.850 (0.830, 0.871)	0.710 (0.674, 0.745)	0.719 (0.685, 0.752)	0.710 (0.674, 0.745)
Finger	0.304 (0.249, 0.358)	0.403 (0.339, 0.467)	0.410 (0.358, 0.463)	0.389 (0.332, 0.446)
Forearm	0.796 (0.772, 0.821)	0.802 (0.779, 0.825)	0.798 (0.774, 0.822)	0.737 (0.707, 0.766)
Hand	0.661 (0.623, 0.698)	0.927 (0.917, 0.937)	0.789 (0.762, 0.815)	0.851 (0.830, 0.871)
Humerus	0.867 (0.850, 0.883)	0.733 (0.703, 0.764)	0.933 (0.925, 0.942)	0.600 (0.558, 0.642)
Shoulder	0.864 (0.847, 0.881)	0.791 (0.765, 0.816)	0.864 (0.847, 0.881)	0.729 (0.697, 0.760)
Wrist	0.791 (0.766, 0.817)	0.931 (0.922, 0.940)	0.931 (0.922, 0.940)	0.931 (0.922, 0.940)
Overall	0.731 (0.726, 0.735)	0.763 (0.759, 0.767)	0.778 (0.774, 0.782)	0.705 (0.700, 0.710)

The baseline model results in the MURA paper (Rajpurkar & Irvin et al., 17)

We also tried ensembling the three models but the kappa score of the ensemble was 0.7 as the VGG19 and the InceptionV3 models did not have kappa scores comparable to our VGG16 model.

7. ACKNOWLEDGMENTS

Our thanks to Dr. Minje Kim who motivated us and taught us the necessary skills to take on an complete our deep learning project.

8. REFERENCES

[1] Rajpurkar & Irvin et al. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. https://arxiv.org/pdf/1712.06957.pdf

[2] https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202

[3] Christian Szegedy Google Inc., Vincent Vanhoucke,Sergey Ioffe, Jonathon Shlens,Zbigniew Wojna: Rethinking the Inception Architecture for Computer Vision https://arxiv.org/pdf/1512.00567.pdf

[4] Inception-v3 for flower classification Xiaoling Xia ; Cui Xu ; Bing Nan https://ieeexplore.ieee.org/document/7984661

[5] https://towardsdatascience.com/learning-rate-schedules-and-adaptive-learning-rate-methods-for-deep-learning-2c8f433990d1

[6] https://cv-tricks.com/cnn/understand-resnet-alexnet-vgg-inception/

[7] Rajpurkar et. al. (2017). Radiologist-Level Pneumonia detection on chest X-rays with Deep Learning. arXiv:1711.05225

[8] Ausawalaithong, Thirach, Marukatat, Theerawit (2018). Automatic lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approaches. arXiv:1808.10858v1

[9] Jang Hyung Lee, Kwang Gi Kim, "Applying Deep Learning in Medical images: The Case of Bone Age Estimation". (2018); doi: 10.4258/hir.2018.24.1.86

[10] Paras Lakhani, Baskaran Sundaram, (2017). "Deep learning at Chest Radiography".

[11] Kaloi, He, (2018). "Child Gender determination with Convolutional Neural Networks on Hand Radiographs. arXiv:1811.05180v1"

[12] Zhiyun Xue, Sameer Antani, L. Rodney Long, and George R. Thoma "Using deep learning for detecting gender in adult chest radiographs", Proc. SPIE 10579, Medical Imaging 2018: Imaging Informatics for Healthcare, research, and Applications, 105790D (6th March 2018); doi 10.1117/12.2293027

[13] Spampinato C., Palazzo S., Giordano D., Aldinucci M, Leonardi R. "Deep Learning for Automated Skeletal Bone Age Assessment in Bone X-rays"

[14] Sungmin Lee, Minsuk Choi, Hyun Soo Choi, Moon Seok Park, Sungroh Yoon; "Fingernet: Deep Learning based robust finger joint detection from radiographs"; Published in IEEE Biomedical Circuits and Systems Conference 2015 DOI: 10.1109/ BioCAS.2015.7348440

[15] Tyler Bradshaw, Timothy Perk, , Song Chen, Hyung-Jun Im, Steve Cho,Scott Perlman and Robert Jeraj;"Deep learning for classification of benign and malignant bone lesions in [F-18]NaF PET/CT images.";J Nucl Med May 1, 2018 vol. 59no. supplement 1 327