DETECTING IRREGULARITIES IN MUSCULO-SKELETAL RADIOGRAPHS USING CONVOLUTIONAL NEURAL NETWORK

Adit Vinay Deshmukh
Dept. Computer Science and Information Systems
BITS Pilani K K Birla Goa Campus

Aashish Agrawal
Dept. Electrical Electronics & Instrumentation Engineering
BITS Pilani K K Birla Goa Campus

Neena Goveas
Department of Computer Science and Information Systems
BITS Pilani K K Birla Goa Campus
Goa, INDIA

Abstract— ***We study the performance of a neural network in classifying Musculoskeletal Images (X-Rays of bones) into two categories, normal and abnormal. We used the MURA dataset released by Stanford ML group which is one of the largest publicly available datasets for Radiographs. Our model is based on Convolutional Neural Networks, performs better than the baseline model in classifying finger and humerus abnormalities. The model can match the performance of radiologist and achieves an accuracy of greater than 80%.***

I. INTRODUCTION

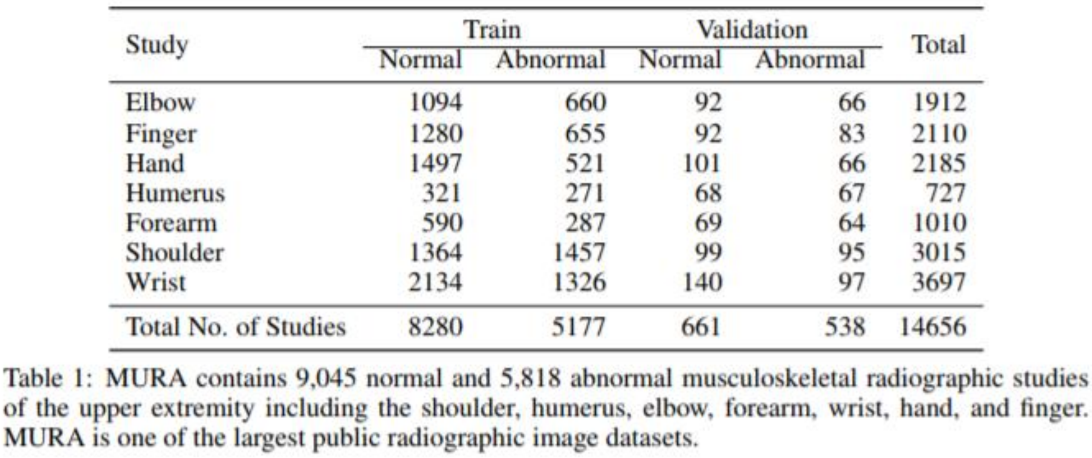
More than 1.7 billion people around the world suffer from Musculoskeletal conditions (BMU, 2017). These conditions cause severe and long-term pain and disability. These conditions also require more than 30 million emergency room visits annually and increasing. The most important tool in discovering and treating musculoskeletal conditions is radiographic studies. Determining whether an X-ray is normal or abnormal is a critical task. If the X-ray is found to be normal, the patient need not undergo further tests.

The MURA dataset released by the Stanford ML group is one of the largest X-ray datasets. Using the dataset to develop medical imaging technologies which can detect abnormalities in bone X-rays has the potential to speed up diagnosis and increase the efficiency of doctors, especially in places which have a shortage of radiologists. It is important to develop models which can perform at the level of radiologists and have a high accuracy so as to be useful for real world applications. The model trained by the Stanford ML group achieved performance close to that of the best performing radiologist.

At the same time, it is important to keep in mind the limitations of the technologies applied. This paper tackles the task of classifying the images of the MURA dataset into normal and abnormal. The model proposed takes in one or more images of a patient and classifies them into one of the two categories. The model gives results comparable to those achieved by the Stanford ML group, even bettering their results in some of the categories.

II. DATASET

Musculoskeletal Radiographs (MURA) is an extensively big dataset of X-Ray images of human bones available to people. It consists of musculoskeletal radiographs of about 14,863 studies from roughly 12,000 patients, containing of 40,561 images taken from different views in total. Each of the images collected can be classified into one of the seven radiographs considered in the study, which are: elbow, shoulder, humerus, finger, hand,



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

Table 1: 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.

The breakdown of images into training and test set of the respective classes is displayed in the table above

forearm and wrist. All of the studies in consideration contains images observed from one or many views, so studies ca have one or multiple images in them. All the studies were labelled manually by board-certified radiologists from the Stanford Hospital at the time of diagnostic radiology environment between the years 2001 and 2012.

The task of labelling the dataset was done during the interpretation on DICOM images presented on at least 3-megapixel PACS medical grade display with max and min luminance of 400 cd/m2 and 1 cd/m2 respectively, with pixel size of 0.2 and native resolution of 1500 x 2000 pixels. The images vary in their aspect ratios and resolution.

The dataset was split into training (roughly 11,100 patients, 13,400 studies and 36,800 images) and validation (roughly 780 patients, 1,200 studies and 3,200 images. No overlap is there in patients between the two sets.

A few of the images from the dataset are displayed below:

III. MODEL ARCHITECTURE AND DATA PREPROCESSING

DenseNet169

Our model is a 169-layer Dense Convolutional Network (DenseNet) (Huang et al., 2016) trained on the MURA dataset. DenseNets benefits by improving the transmission of gradients and information through the network, because they have every layer connected to every other layer in a feed forward fashion, making it easier to trace optimization of very deep network. The last layer of the DenseNet is replaced with a fully connected layer having only one output unit, after which a sigmoid nonlinearity is applied. Pre-trained weights from the ImageNet (Deng et al., 2009) are taken to initialize the weights of the network. Adam optimization is used to train the model end-to-end with default parameters (β1 = 0.9 and β2 = 0.999) (Kingma and Ba., 2014). The model was trained on mini-batches of size 32, with an initial learning rate of 0.001 and exponential learning rate scheduler was used to penalize it by a factor of .01 after every 7 epochs.

Before inputting the images through the DenseNet architecture, each image was downsized to 224 x 224 and were randomly cropped after which they were normalized and randomly horizontally flipped. For the validation data, the images were resized to 256 x 256 and then a center crop of 224 x 224 was applied to it before normalization.

Ensemble

The ensemble model uses a Resnet101 alongside the DenseNet169. This improves our overall Kappa statistic score and accuracy score. The ensemble model is found to perform either similar to DenseNet or in some cases worse than DenseNet architecture.

IV. RESULTS

Table 1: COHEN'S KAPPA

	Baseline Model	Our Model	Ensemble Model
Elbow	0.710	0.656	0.705
Finger	0.389	0.584	0.558
Forearm	0.737	0.651	0.681
Hand	0.851	0.570	0.572
Humerus	0.600	0.748	0.748
Shoulder	0.729	0.681	0.649
Wrist	0.931	0.712	0.712
Overall	0.705	0.664	0.675

We used Cohen's kappa statistic, which expresses the agreement of each model with the gold standard, on our model and compared it with the Kappa statistic of the model by authors of the original paper. The values in green shows that our model had better statistic than the baseline model. The value in red shows that our model performed worst on the Hand dataset as compared to the baseline model. The values in blue show that our models performance is similar to that of the Baseline Model.

Table 2: ACCURACY SCORE

	Our Model	Ensemble Model
Elbow	0.835	0.861
Finger	0.794	0.783
Forearm	0.827	0.842
Hand	0.808	0.808
Humerus	0.874	0.874
Shoulder	0.840	0.825
Wrist	0.865	0.865
Overall	0.836	0.836

The above table shows the accuracy score for our model against every image class available in dataset.

The performance matrices used, Cohen's Kappa and accuracy scores, are reported on the validation dataset provided by the Stanford ML group along with the training dataset, and further testing on the test dataset is still underway. We are yet to receive the scores on the test set.

V. CONCLUSION

Our study shows that neural network-based methods can be used on classification of abnormalities in radiographs and can achieve performance similar to that of radiologists. While implemented on a large scale, this can help radiologists and doctors provide care faster and more efficiently, especially in areas where there are fewer than required radiologists.

VI. REFERENCES

[1] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. arXiv:1712.06957

[2] Pranav Rajpurkar, Jeremy Irvin , Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.

[3] Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[4] Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 248–255. IEEE, 2009.

[5] Huang, Gao, Liu, Zhuang, Weinberger, Kilian Q, and van der Maaten, Laurens. Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

[7] 2017. http://www.boneandjointburden.org/2014-report.