Abnormality Detection in Musculoskeletal Radiographs using Transfer Learning Models

Jasmine Chhikara

Assistant Professor, Dept. of Electronics and Communication Engineering, Maharaja Surajmal Institute of Technology, Delhi, India

---***---

Abstract - *Millions of people get affected with musculoskeletal diseases every year and in a highly populated country like India there is low ratio of qualified doctors to the total population who can significantly treat joint and muscle related issues. Finding best treatment of health issues becomes problematic especially in remote and rural areas. The automated prediction system can help in diagnosis of the disease quicker and more accurately without leaving radiologist exhausted. In pursuit of this objective, we have utilized transfer learning models of deep learning architecture convolutional neural networks including DenseNet, Inception and Inception Resnet models which are trained using cyclic learning rate scheduler for predictive model using less number of epochs. The neural networks are evaluated on cohen kappa (95%), sensitivity, specificity and area under ROC curve with the highest kappa score of 0.698 and AUROC of 0.899. It achieved comparable results to baseline model in less training epochs and even performed better in radiographic studies.*

***Key Words*: Convolutional Neural Networks (CNN), Musculoskeletal Radiographs, Cohen Kappa, Cyclic Learning Rate**

1. INTRODUCTION

Musculoskeletal condition is a state in which pain and injury are caused in muscles and bones due to sudden exertions. The parts that are affected the most by this condition are joints, ligaments, muscles, nerves, tendons, and structures that support limbs, neck and back. Analysis of data from WHO's study on global ageing and adult health point to the high prevalence of arthritis in low and middle income settings, particularly among those in a lower socioeconomic position (S. L. Brennan-Olsen et al. [1]). The Global Burden of Disease (GBD) reported that nearly a third of people across the globe live with a painful and disabling musculoskeletal condition (James et al. [4]). It significantly limits physical mobility and ability to accumulate wealth over time.

The ground motive for taking this project is to help people living in rural areas, overcome the pain and trouble caused by the musculoskeletal condition. The models are constructed to determine whether an X-Ray is normal or abnormal. The objective of the experiments is to obtain results which are accurate, economic and mobile. We believe that anyone from any corner of the earth can use these models to detect abnormalities caused by musculoskeletal conditions.

2. DATASET

Musculoskeletal Radiographs (MURA) is a collection of 14,656 studies from 11,967 patients' bone X-rays from Stanford University (Rajpurkar et al. [7]) which are publically available. It has 40,005 multi-view radiographic images of elbow, finger, fore- arm, hand, humerus, shoulder, and wrist studies. Each study contains one or more views (images) indicating whether the study is normal or abnormal, respectively.

The images are in .png format and vary in resolution and aspect ratios. The dataset is divided into training and validation sets having no overlap in patients between the sets.

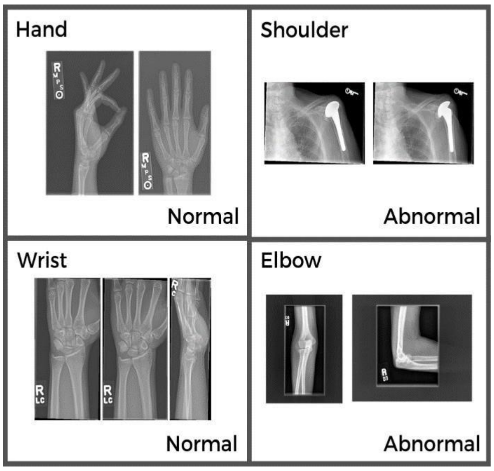

Fig-1: These examples show a normal hand study (top left), an abnormal shoulder study (top right), a normal wrist study (bottom left), and an abnormal elbow study (bottom right).

3. METHODOLOGY

The model takes as input one or more views for a study of an upper extremity and the overall probability for the study is computed by taking the mean of the probabilities output by the network for each image. The model makes the binary prediction of abnormal if the probability of abnormality for the study is greater than 0.5 as done by Rajpurkar et al. [7].

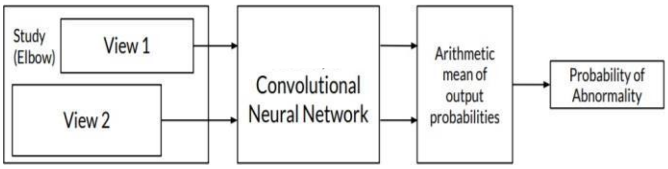
Study (Elbow)
View 1
View 2
Convolutional Neural Network
Arithmetic mean of output probabilities
Probability of Abnormality

Fig-2: The model takes as input one or more views for a study. The per-view probabilities are then averaged to output the probability of abnormality for the study.

3.1 Learning Rate Scheduler

A triangular window cyclic learning rate (Smith, 2017) is chosen to change the learning rate during training with a half cycle of 4 epochs. The learning rate scheduler requires an upper limit and a lower limit on the learning rate which is calculated by plotting a graph of loss versus learning rate. The lower limit is chosen where the loss just starts to reduce and the upper limit is chosen where the loss starts to increase again.

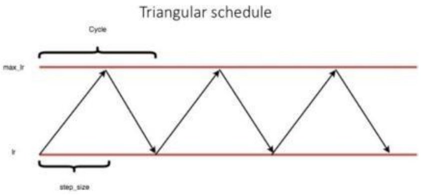

Fig-3: Triangular window cyclic learning rate policy

3.2 Training

Each image is a grayscale image which is rescaled to 320x320 size and Imagenet normalized. Images are given random transformation like rotation, horizontal flipping or vertical flipping before feeding into the model.

The abnormality probability for images in a study is predicted using convolutional neural networks followed by a sigmoid function applied to final output. The neural network is initialized with pre-trained weights on Imagenet (Deng et al. [2]).

For each image in the training set, a weighted binary cross-entropy loss is used as done by Rajpurkar et al. [7] The network is trained using Adam β1 = 0.9 and β2 = 0.999 (Kingma & Ba et al. [5]). Batch size of 16 is used for training.

4. EXPERIMENTS

We conducted three experiments as described below with different model topologies with the methodology described in the previous section.

4.1 Experiment 1: DenseNet Model

This experiment consisted of random rotation of 30 degrees, horizontal & vertical flips and learning rate ranged from 5x10^{-6} - 7.5x10^{-5}. The network for training used Dense Convolutional Network architecture (Huang et al. [3]) with 169 layers model as the base model and added average pooling layer and output layer on top of it.

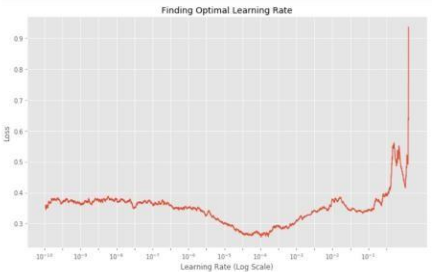

Fig-4: Finding the learning rate range for Experiment 1

4.2 Experiment 2: Inception v3 Model

This experiment consisted of random rotation of 37 degrees, horizontal flips and learning rate ranged from 3.16x10 to 4x10. The network for training used Inception Network architecture (Szegedy et al. [10]) with 159 layers model as the base model and added average pooling layer and output layer on top of it.

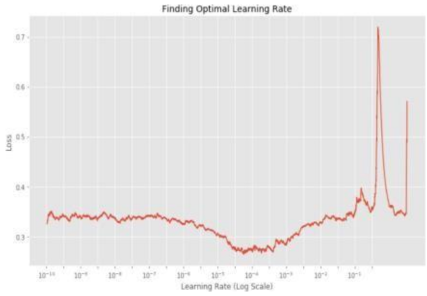

Fig-5: Finding the learning rate range for Experiment 2

4.3 Experiment 3: Inception-ResNet Model

This experiment consisted of random rotation of 37 degrees, vertical flips and learning rate ranged from 1x10^{-5} - 1x10^{-4}. The network for training used a combination of Inception and Residual Network architecture (Szegedy et al. [10]) with 164 layers model as the base model and added aver- age pooling layer and output layer on top of it.

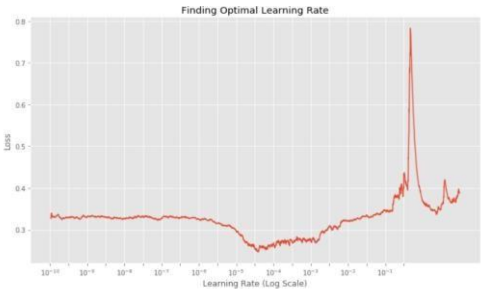

Fig-6: Finding the learning rate range for Experiment 3

5. RESULTS

The model performances are assessed on Cohen Kappa (McHugh et al. [6]), 95% Confidence Interval, Sensitivity, Specificity and AUROC statistics.

The model which achieved the highest performance on the test dataset is the Dense-Net169-layer architecture with the Kappa statistic of 0.698 and 95% CI of [0.657, 0.739]. Receiving Operating Characteristics are plotted and Area Under ROC curve is 0.899. The model has a Sensitivity of 0.776 and a Specificity of 0.913.

Similarly, Inception v3 and Inception-ResNet model's Kappa statistic with CI is 0.697 [0.656, 0.738] & 0.687 [0.646, 0.729] and AUROC is 0.897 & 0.888 respectively.

In all three experiments, the model performed comparably to the model presented in Rajpurkar's work but took few epochs to achieve that accuracy (8-14 epochs). In all experiments, the model starts to overfit the training dataset and achieves a Kappa statistic of more than 0.80. Here, we will refer to Rajpurkar's neural network which was DenseNet169 as RDEN169.

Table-1: DenseNet model performance

	Cohen Kappa (95% CI)	Sensitivity	Specificity	AUROC
Hand	0.593	0.636	0.930	0.835
Wrist	0.776	0.793	0.964	0.932
Humerus	0.777	0.880	0.897	0.924
Shoulder	0.566	0.778	0.787	0.856
Elbow	0.747	0.787	0.945	0.906
Finger	0.677	0.771	0.902	0.888
Forearm	0.742	0.781	0.956	0.926
Overall	**0.698**	**0.776**	**0.913**	**0.899**

Table-2: Inception model performance

	Cohen Kappa (95% CI)	Sensitivity	Specificity	AUROC
Hand	0.645	0.666	0.950	0.856
Wrist	0.775	0.773	0.978	0.936
Humerus	0.762	0.850	0.911	0.912
Shoulder	0.628	0.800	0.828	0.885
Elbow	0.696	0.772	0.913	0.881
Finger	0.665	0.759	0.902	0.882
Forearm	0.681	0.718	0.956	0.890
Overall	**0.697**	**0.765**	**0.922**	**0.897**

Table-3: Inception-ResNet model performance

	Cohen Kappa (95% CI)	Sensitivity	Specificity	AUROC
Hand	0.552	0.606	0.920	0.816
Wrist	0.749	0.773	0.957	0.929
Humerus	0.762	0.835	0.926	0.916
Shoulder	0.618	0.821	0.797	0.851
Elbow	0.760	0.787	0.956	0.907
Finger	0.667	0.807	0.858	0.885
Forearm	0.681	0.718	0.956	0.869
Overall	**0.687**	**0.769**	**0.910**	**0.888**

Table-4: Comparison with baseline RDEN169 on kappa

	DenseNet	Inception-v3	Inception-ResNet	RDEN169
Hand	0.593	0.645	0.552	0.851
Wrist	0.776	0.775	0.749	0.931
Humerus	**0.777**	**0.762**	**0.762**	0.600
Shoulder	0.566	0.628	0.618	0.729
Elbow	**0.747**	0.696	**0.760**	0.710
Finger	**0.677**	**0.665**	**0.667**	0.389
Forearm	**0.742**	0.681	0.681	0.737
Overall	0.698	0.697	0.687	0.705

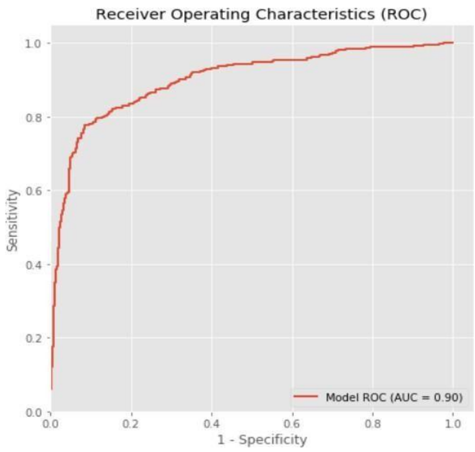

Fig- 7: ROC curve for DenseNet model

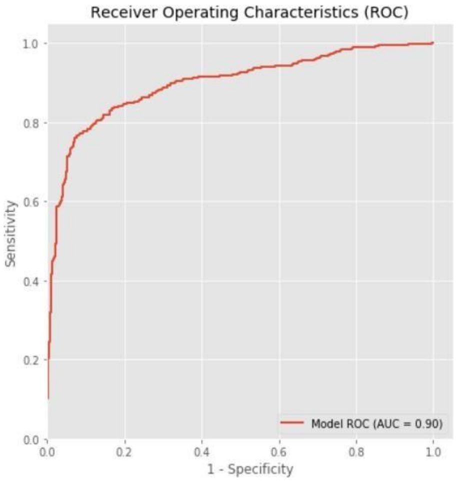

Fig-8: ROC curve for Inception model

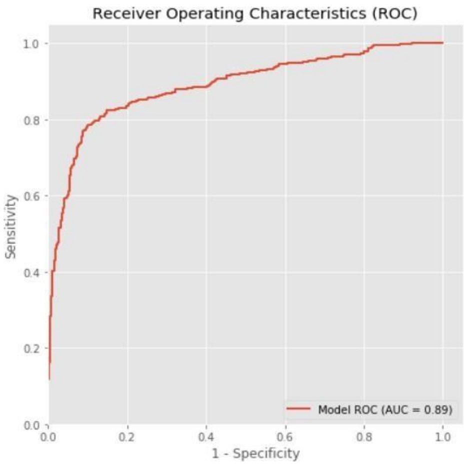

Fig-9: ROC curve for Inception-ResNet model

REFERENCES

[1] Brennan-Olsen, S. L., Cook, S., Leech, M. T., Bowe, S. J., et al.: Prevalence of arthritis according to age, sex and socioeconomic status in six low and middle income countries: analysis of data from the World Health Organization study on global AGEing and adult health (SAGE) Wave 1. BMC Musculoskeletal Disorders 18(1), (2017).

[2] Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, Li: Imagenet: A large-scale hier- archical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE, Miami, Florida (2009).

[3] Huang, G., Liu, Z., Weinberger, K., and van der Maaten, L.: Densely connected convolu- tional networks. In: 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700-4708. IEEE, Honolulu, Hawaii (2017).

[4] James, S. L., Abate, D., Abate, K. H., et al.: Global, regional, and national incidence, prev- alence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet 392(10159), 1789-1858 (2018).

[5] Kingma, D., and Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations. ICLR, San Diego, California (2015).

[6] McHugh, M. L.: Interrater reliability: the kappa statistic. Biochemia Medica 22(3), 276–282 (2012).

[7] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., et al.: MURA: Large dataset for abnormality detection in musculoskeletal radiographs. In: 1st Conference on Medical Imaging with Deep Learning. MIDL, Amsterdam (2018).

[8] Smith, L. N.: Cyclical Learning Rates for Training Neural Networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision, pp. 464-472. IEEE, Santa Rosa, California (2017).

[9] stanfordmlgroup. Stanford Machine Learning Group, https://stanford-mlgroup.github.io/competitions/mura, last accessed 2019/08/10.

[10] Szegedy, C., Vanhoucke, V., Loffe, S., Shlens, J., and Wojna, Z.: Rethinking the inception architecture for computer vision. In: 29th IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2816. IEEE, Las Vegas, Nevada (2016).

[11] Szegedy, C., Vanhoucke, V., Loffe, S., and Alemi, A.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: 31st AAAI Conference on Artificial In- telligence, pp. 4278-4284. The AAAI Press, San Francisco, California (2017).