Ensemble Based Neural Network for the Classification of MURA Dataset

Mithun Ghosh
Department of
Systems & Industrial Engineering
University of Arizona
Tucson, AZ 85719
mithunghosh@email.arizona.edu

Abstract

Musculoskeletal Radiographs (MURA) dataset, proposed by Stamford Machine Learning (ML) group, contains 40,561 images of bone X-rays from 14,863 studies. The X-ray images belong to seven body areas of upper extremity- Wrist, Elbow, Finger, Humerus, Forearm, Hand, and Shoulder. The data are classified manually by radiologists into two classes- normal or abnormal. These data samples are labeled using majority vote by six board-certified Stanford radiologists. The majority votes of these radiologists' labels are considered as gold standard. The presence of such rich,complex and diverse labeled dataset inspires to build an accurate but simpler model for bone anomaly detection. The model proposed by Stamford ML group is a 169 layer deep computationally complex Neural Network (NN), that requires a Graphical Processing Unit (GPU) for implementation. This leads to the necessity of smaller neural network based model that are executable on general purpose computers. Moreover, the 169 layer deep model works well on par with the gold standard except for the humerus radiographs, despite the presence of humerus data labeled with high accuracy. Therefore, in this work we propose an ensemble of smaller neural networks and convolution neural network for highly accurate classification of MURA study images of humerus. We use Adaboost algorithm to train this model. The performance of this model is evaluated using training error, validation error, and Cohen's kappa coefficients. The model is available at https://github.com/mythgotham007/Mura_Humerus_CNN-NN/import .

1 Introduction

The current era of big data has accelerated the development of machine learning based models. It is becoming increasingly crucial that these models can handle large dataset without losing any significant information. Out of the several application fields of these models, the medical sector is one of the most demanding sectors. It requires models that can classify large amount of data with high accuracy. Among the several machine learning approaches, neural network is in its prime to handle large dataset (Deng et al. [2009]) with significant accuracy. This makes it a suitable choice for medical applications.

There are several branches in medical field where classification models can play an important role. The radiographic study is one such branch, where the correct classification of studied data is of utmost importance. According to WHO (World Health Organization), musculoskeletal conditions are becoming a major burden on individuals. Woolf and Pfleger [2003] discussed the burden of the major musculoskeletal conditions. About 1.7 billion people are affected by this conditions worldwide. Based on the classification outcome, the patient may need to go further diagnosis and treatment. But most of the time it is not easy even for the naked eye to detect any abnormality in the X-ray images

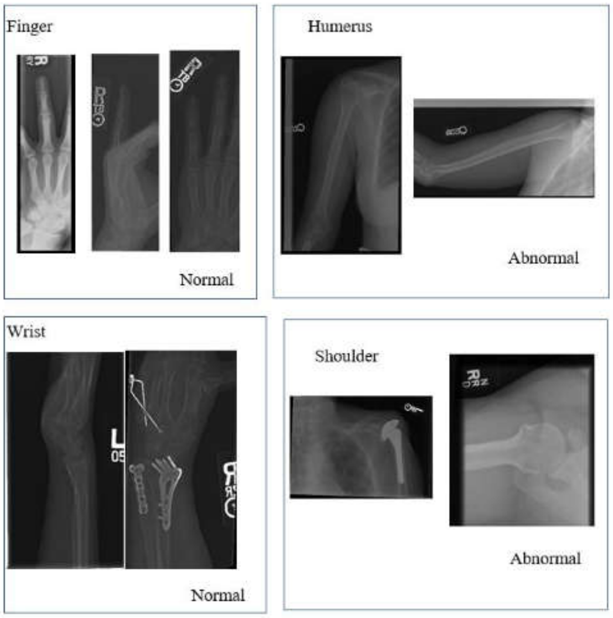

Figure 1: MURA dataset contains 14863 images of the radiography of musculoskeletal studies of the upper extremity. In each of the study multiple views are manually labeled by radiologists. Right side of the above Figure explains some normally labeled images of Elbow and Wrist, respectively wherein left side describes some abnormal images from the Humerus and Shoulder, respectively.

which we can notice in some of the images from Fig. 1. Mistreatment due to the misclassification is thus critical in any situation for that patient.

MURA is a large dataset of musculoskeletal radiographs recently proposed by Stamford Machine Learning (ML) group. This dataset can help train and create accurate models for classification in radiographic studies. It contains 14,863 studies with 40,561 images. Each study in the data contains multiple views of the targeted body part, which are labeled by radiographers as normal or abnormal. These are high resolution images that make the size of the whole dataset around 3.3 GB. The large size of this dataset and the crucial accuracy requirements of the radiographic study sector requires neural network based approaches.

The Stamford ML group proposed a 169 layer deep neural network model of dense layers to reliably classify the MURA dataset. This model classifies most of the X-ray images of different body parts with high degree of agreement with the radiologist classifiers (gold standard). However, it shows low agreement with gold standard for finger and humerus dataset. In case of the finger dataset, Rajpurkar et al. [2018] shows that the existing finger dataset labels had low value of Cohen's kappa coefficient. This causes the model to perform poorly. However, for humerus data, despite presence of highly accurate labels the model performs poorly. Moreover, the computational complexity of this deep neural network model requires a GPU to execute. In order to address these two issues, this work presents an ensemble model of 12 layered neural networks. This model can run on a general purpose computer without the necessity of a GPU. But to get the full advantage of the model, we require some high performing computer(HPC) due to the large size of the image. We also build an autoencoder with Convolution Neural Network(CNN). The testing and training has been carried out on humerus dataset only.

The following part of this paper is organized as- section 2 describes the details of MURA dataset, section 3 describes the proposed ensemble model, section 4 contains the model interpretation, section 5 compares the performance of the proposed model with Stamford ML group's model, section 6 describes the related works and section 7 discusses the overall findings.

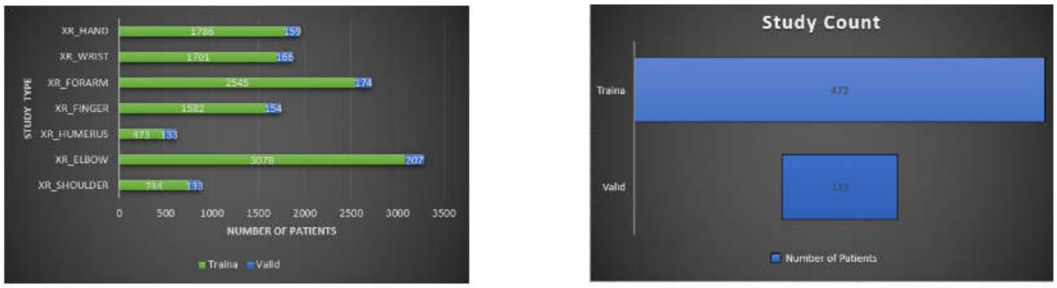

Figure 2: Left: Statistics of the data in each of the seven categories of the studies. In train set XR_WRIST has maximum number of patients, followed by XR_FINGER, XR_HUMERUS, XR_SHOULDER, XR_HAND, XR_ELBOW and XR_FOREARM. X_FOREARM with 606 patients has got the least number. Similar pattern can be seen in valid set, XR_WRIST has the maximum, followed by XR_FINGER, XR_SHOULDER, XR_HUMEROUS, XR_HAND, XR_ELBOW, XR_FOREARM. Here XR_FINGER defines radiographs of Finger upper extremity. Right: Data Statistics for Humerus Data

2 MURA

The task in MURA dataset is to find the binary class of $\{0, 1\}$. Each study contains one or more views of images and the expected output is then denoted as 0 or 1. We denote 0 as normal and 1 as abnormal. A brief summary of the study data is given below:
In the official MURA dataset website (`https://stanfordmlgroup.github.io/competitions/mura/`) we can see the performance of various authors model in different categories. Some of these model perform well for different categories. But For upper extremity categories: Humerus and Finger, almost all of the models perform worst. This may be due to the fact that, the images in these categories are not so clear. Also, the number of samples are not sufficiently high in these categories. We can look that in right side of Fig. 1, these images are not so clear to be easily classified by the models. Thus, for this reason, we concentrate our model to the Humerus cases. In left side of Fig. 2, we can also see that number of studied patients are very few for both training and test dataset in Humerus study.

2.1 Data Collection

Data is collected from MURA dataset official website. A large number of patients are studied with multiple view of their radiographs. Total number of radiographs image is 14,863 that was collected from 12,713 patients with a total of 40,561 multi-view X-ray images. These images belong to one of the seven standard upper extremity radiographic study types as described previously.
At the moment of medical radiographic analysis in the diagnostic radiology atmosphere between 2001 and 2012, each study was manually marked as normal or abnormal by Stanford Hospital board-certified radiologists. The labeling was conducted during the assessment of DICOM(Digital Imaging and Communications in Medicine) images on a medical grade display of at least 3 megapixels PACS(Picture Archiving and Communication System) with max luminance of 400 cd/m^2 and min luminance 1 cd/m^2 with pixel size of 0.2 and native resolution of 1500 x 2000 pixels. Clinical images vary in pixel density and aspect proportions. The data set is divided into training sets (11,184 patients, 13,457 studies, 36,808 images), validation sets (783 patients, 1,199 studies, 3,197 images), and test sets (206 patients, 207 studies, 556 images). There is no crossover between any sets in patients.

2.2 Test Set Collection

Extra labels of data are obtained from board-certified Stanford radiologists on the test set, composed of 207 musculoskeletal experiments, to assess models to get a resilient assessment of radiologist performance. The radiologists individually examined and categorized each study in the test set as a DICOM file as normal or abnormal in the clinical reading room environment using the PACS. Radiologists have an average of 8.83 years of professional experience varying from 2 to 25 years.

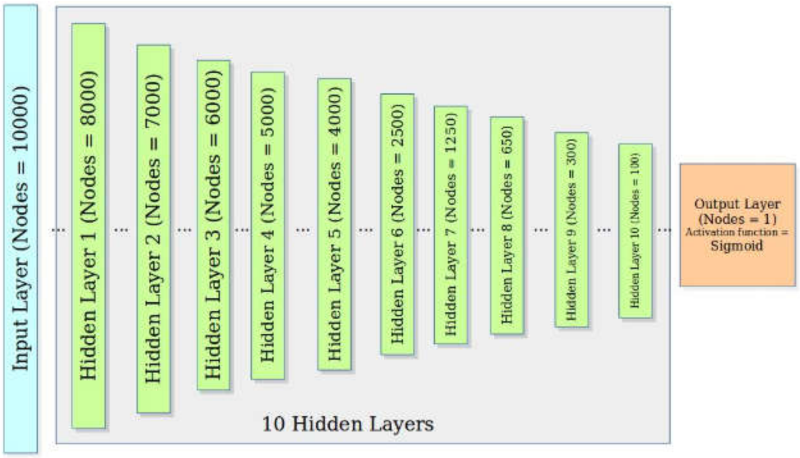
Input Layer (Nodes = 10000)
Hidden Layer 1 (Nodes = 8000)
Hidden Layer 2 (Nodes = 7000)
Hidden Layer 3 (Nodes = 6000)
Hidden Layer 4 (Nodes = 5000)
Hidden Layer 5 (Nodes = 4000)
Hidden Layer 6 (Nodes = 2500)
Hidden Layer 7 (Nodes = 1250)
Hidden Layer 8 (Nodes = 650)
Hidden Layer 9 (Nodes = 300)
Hidden Layer 10 (Nodes = 100)
Output Layer
(Nodes = 1)
Activation function =
Sigmoid
10 Hidden Layers

Figure 3: Structure of the small neural network used as weak classifier in the ensemble model.

The radiologists had no direct exposure to any clinical data. Classifications have been entered into a standardized data entry program.

2.3 Abnormality Analysis

In order to investigate the types of abnormalities present in the dataset, we 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 abnormalities including lesions and subluxations.

3 Model

The proposed ensemble model is made up of smaller deep neural networks. These neural networks are trained on the train data from humerus X-ray images of MURA using the Adaboost algorithm (Freund and Schapire [1999]). Section 3.1 explains the structures and training of each of these smaller neural networks. Section 3.2 explains the overall ensemble training and prediction methods.

3.1 Network Architecture and Training

The structural details of each smaller deep neural networks is shown in Fig. 3. Each neural network has 10000 nodes in input layers and one node in output layer. There are 10 hidden layers between input and output layer. As the network grows deeper, number of nodes into each new layer reduces. These neural networks are modeled using models available in Keras library.
Now, The structural details of each of CNN is shown in Fig. 4. Each of the CNN consists of kernel window size of 4x4 with stride one and Maxpooling of 2x2. This auto-encoder network compress the size of the image to from 100x100 to 9x9 with minimum reconstruction error. After that, the layers are flatten and fully connected to dense layer with 500 nodes and rectified linear activation function and then another dense layer with 20 nodes and finally the two output layer.

In order to optimize the network, each neural network uses the ADADELTA optimizer (Zeiler [2012]), a novel per-dimension learning rate method for gradient descent. The method does not require any manual tuning of a learning rate and appears robust to noisy gradient information, different model architecture choices, various data modalities and selection of hyperparameters. We considered minimizing the binary cross entropy loss as there are only two output classes- Normal and Abnormal. For image X of study type T in the training set, the loss is:

$$L(X, y) = w_{T,1} y \log p(Y = 1|X) - w_{T,0}(1 - y) \log p(Y = 0|X). \tag{1}$$

Here, y and $p(Y = i|X)$ is the label and probability that the network assigns to the label i, respectively. Also, $w_{T,1} = \frac{|N_T|}{|A_T|+|N_T|}$ and $w_{T,0} = \frac{|A_T|}{|A_T|+|N_T|}$, where N_T and A_T defines normal and abnormal images of the particular study T in the training set, respectively.

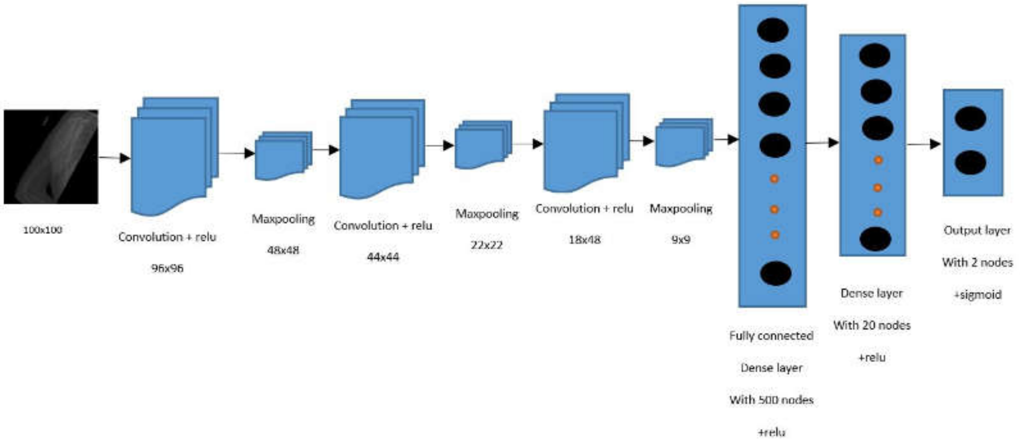

Figure 4: Structure of the small convolution neural network used as weak classifier in the ensemble model.

The hidden layer nodes used Rectified Linear Units (ReLU) function as activation function. This activation function is easier to calculate and enables faster convergence during training. The output layer uses Sigmoid function for activation. The threshold for the output node is set at 0.5. The ReLU and Sigmoid functions are shown in equation 2 and 3 respectively.

$$f(x) = \max(0, x) \tag{2}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

Among the several training parameters of the neural networks, we kept the training epoch number as a variable to study the effect of epochs on the classifiers performance on training and validation dataset. There are overall 1272 training images on humerus. The batch size is kept 30 during training. We vary the epoch sizes from 10 to 40.

3.2 Ensemble Training and Prediction

In the previous sections, the smaller neural networks are described. Each of these neural networks perform slightly better than a random classifier on the humerus training data. Therefore, we use the Sampling based Adaboost algorithm to train these neural networks with emphasis on the wrong classifications made by previously trained classifier. The ensemble model contains 5 small neural networks, which implies that 5 Adaboost iterations are used to train the model. For the CNN, due to lack of resource we just fitted it with five classifiers and ten epoches.

The overall flow of the ensemble training is shown in Algorithm 1. Initially, all the training images and their labels are imported from directories. The training images contain different amount of pixels. We resize all the images into 100×100 pixels, and flatten it to 1×10000 pixels shape. After that, we normalize these pixel values between $[0, 1]$. Therefore we launch the iterative training following Adaboost algorithm, with the initial training data distribution (D) equal for each sample. In each iteration, a small NN or CNN is trained on training data sampled as per the updated distribution by the previous iteration. Upon completion of training, the weighted error ϵ_t is calculated (step 5). Based on ϵ_t, the weight of the trained neural network, α_t is calculated (step 6). Then, the distribution D is updated as per the equation described in step 7 of the algorithm. Here, Z_t is a normalization term. Finally, the training dataset is sampled as per the updated D for training purposes in the next iteration.

Algorithm 1 Training steps for neural network ensemble model using Adaboost Algorithm

```
**Input:** $S := \{x_i, y_i\}_{i=1}^{N}$, *Learning rounds* $T$, and *hypothesis class* $H$
**Initialize:** *Distribution* $D_1(i) = \frac{1}{n}$, $S := normalize(S)$
**for** $t = 1, 2, ...., T$ **do**
    $h_t = argmin_{h \in H} \hat{err}(h, S, D_t)$
    $\epsilon_t = \sum D_t(i)[h(x_i) \neq y_i]$
    $\alpha_t = \frac{1}{2} log(\frac{1-\epsilon_t}{\epsilon_t})$
    $D_{t+1}(i) = \frac{D_t(i)}{Z_t} exp(-\alpha_t y_i h_t(x_i))$
**end for**
```

The prediction using this ensemble neural network model is conducted as per the following equation.

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{4}$$

4 Model Interpretation

As mentioned before, we aim to focus on the most vulnerable upper extremity in MURA dataset namely, Humerus. We calculated both training error and testing error based on train and test data. The error calculation formula:

$$Error = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})}{n} \tag{5}$$

where y_i defines the ith true observed value , $\hat{y_i}$ defines corresponding predicted value form our model and n defines length of the observation.

The kappa coefficient (κ) of Cohen is a measurable statistic of inter-rater agreement for qualitative (categorical) items. It is generally regarded to be a much more robust measure than a simple percentage agreement calculation, as the probability of the arrangement occurring by chance is taken into consideration by κ. The kappa of Cohen measures the agreement between two raters, each classifying N items into categories that are mutually exclusive. Galton [1892] is ascribed to the first mention of a kappa-like statistics. κ is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

where p_o is the relative observed agreement among raters and p_e is the hypothetical chance agreement probability, using the observed data to calculate the probabilities of each observer seeing each category randomly. If the raters are in full agreement, $\kappa = 1$. If the raters do not agree other than what is expected by chance (as given by p_e), $\kappa = 0$, which implies that there is no effective agreement between the two raters or the agreement is worse than random. In the following section, we describe our model performance using these statistics.

5 Radiologist vs Model Performance

In order to evaluate the performance of our proposed model, we use training error, validation error, and Cohen's kappa coefficient (κ) for both training and validation data. Table 1 presents the κ values of classifications made by the Radiologist with best performance, model by Stamford ML group, and model proposed in this work. The results contain model performance for different training epochs. All the mentioned κ values in this table is calculated by comparing with the gold standard.

Table 1 results show that, for classification of training data, the model performs somewhat reliable. With increasing training epochs, the κ value increases, indicating increasing reliability of the model.

For classification on validation data, the κ value is quite lower compared to the radiologist and Stamford ML group model. However, with increasing training epochs, κ of our model for validation data classification increases. This indicates, with larger training epochs, our model can be made reliable enough to compete with the other two models. However, increasing the number of epochs

	Radiologist 3	Stamford ML Model	Proposed Model(NN) (Epoch = 20)	Proposed Model(NN) (Epoch = 30)	Proposed Model(NN) (Epoch = 40)	Proposed Model(CNN) (Epoch=10)
Validation	0.933	0.600	0.025	0.050	0.119	0.123
Train	-	-	0.319	0.441	0.459	0.51

Table 1: We compare the performance of our proposed model(NN) with best radiologist performance and Stamford ML group model performance, on the Humerus dataset of MURA. The comparison in done in terms of Cohen's kappa (κ)performance.

increases training time significantly, which is why we could not present results for higher epochs. We suggest using High Performance Computing (HPC) servers to train this model for higher epochs.

Along with κ values, we also measured the training and validation error of our model. For training epoch 20, the model yields training and validation error of 34% and 48% respectively. For training epoch 30, training and validation errors are 27% and 47%. Finally, for 40 epochs, training error is 27% and validation error is 44%. The study of training and validation errors show that increasing training epochs result in decreasing training and validation errors. This again brings us to the conclusion of training the ensemble model for more epochs to obtain better performance. Moreover, we may increase the number of smaller neural networks to get more reliable classification performance. For the CNN with five classifiers, we get 25% training and 0.43% testing errors.

As, we can see from the table that CNN performs better with comparatively lower epoch. For the CNN, with the increase of the number of classifiers we believe that we will get a better κ value and also training and testing error will be reduced. As mentioned above, due to the lack of resource, we could not perform much larger number of classifiers. We fixed epoch size to 10 in the CNN model.

6 Related Work

Our model is robust in a sense that we can apply in many large available dataset in many categories such as: speech recognition (Hannun et al. [2014]), question answering (Rajpurkar et al. [2016]), skin cancer (Esteva et al. [2017]), heart arrhythmias (Rajpurkar et al. [2017]) etc. Nowadays, because of large development in technology, we get large amount of data in many areas. Large data means more information and our model should be able to handle in these areas also. Its not so easy to find openly available radiographs dataset although there are some growing effort to make repositories of medical datasets available openly. Previously collected datasets are smaller than MURA in size. Table 2 gives publicly available dataset in radiographs images.

There are truly few publicly accessible dataset of musculoskeletal radiograph. The Stanford Artificial Intelligence Program in Medicine and Imaging hosts a dataset comprising illustrated skeletal age (AIMI) pediatric hand radiographs. The Digital Hand Atlas consists of left hand x-rays of children of different ages labeled with bone age radiology readings (Gertych et al. [2007]). The Osteoarthritis Initiative is hosting the 0.E.1 dataset containing knee radiographs labeled with the K&L grade of osteoarthritis (OAI). Each of these datasets contain less than 15,000 images. Some of the seminal work have been done in MURA dataset. From MURA dataset official website, we can see the rank and performance of these models. Xuan et al. [2018] did a family of embedding functions that can be used as an ensemble to give improved results.

7 Discussion

Early stage detection of anomaly in radiographs is crucial for the patient. We can notice from the results of Rajpurkar et al. [2018] that even experienced radiologists may sometime misclassify some anomalies. Moreover, human classification is costly, more time consuming and requires more effort. These reasons have made machine learning based classifier model a reliable alternative. Although there are several established models that perform relatively good on MURA dataset, for some upper extremities they are not reliable enough. To add on top of that, these models require compute intensive complex Neural Network based models that are difficult to train. We attempt to address these issues in our model by building an ensemble based neural network model, that can perform well on these

Dataset	Study Type	Label	Images
MURA	Musculoskeletal (Upper Extremity)	Abnormality	40561
Pediatric Bone Age (AIMI)	Musculoskeletal (Hand)	Bone Age	14,236
O.E.1 (OAI)	Musculoskeletal (Knee)	K&L Grade	8,892
Digital Hand Atlas (Gertych et al., 2007)	Musculoskeletal (Left Hand)	Bone Age	1,390
ChestX-ray14 (Want et al., 2017)	Chest	Multiple Pathologies	112,120
Openl (Demner-Fushman et al., 2015	Chest	Multiple Pathologies	7,470
MC (Jaeger et al., 2014)	Chest	Abnormality	138
Shenzen (Jaeger et al., 2014)	Chest	Tuberculosis	662
JSRT (Shiraishi et al., 2000	Chest	Pulmonary Nodule	247
DDSM (Health et al., 2000)	Mammogram	Breast Cancer	10, 239

Table 2: Overview of publicly available medical radiographic image datasets.

images. The obtained results suggest that this model works with some degree of reliability. We further notice an increasing trend in model reliability with the increasing number of training epochs. Based on these results we strongly believe that, with the increase in computational resources, our model can be one of the most reliable candidates in MURA dataset classification.

References

J. Deng, W. Dong, R. Socher, L. Li, and and. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–, January 2017. URL http://dx.doi.org/10.1038/nature21056.

Yoav Freund and Robert E. Schapire. A short introduction to boosting. In *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1401–1406. Morgan Kaufmann, 1999.

Francis Galton. *Finger Prints*. Macmillan and Co., 1892.

Arkadiusz Gertych, Aifeng Zhang, James Sayre, Sylwia Pospiech-Kurkowska, and H.K. Huang. Bone age assessment of children using a digital hand atlas. *Computerized Medical Imaging and Graphics*, 31(4):322 – 331, 2007. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2007.02.012. URL http://www.sciencedirect.com/science/article/pii/S0895611107000274. Computer-aided Diagnosis (CAD) and Image-guided Decision Support.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014. URL http://arxiv.org/abs/1412.5567.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016. URL http://arxiv.org/abs/1606.05250.

Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *CoRR*, abs/1707.01836, 2017. URL `http://arxiv.org/abs/1707.01836`.

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, and Andrew Y. Ng. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *CoRR*, abs/1712.06957, 2018. URL `http://arxiv.org/1712.06957`.

A.D. Woolf and B. Pfleger. Burden of major musculoskeletal conditions. *Bulletin of the World Health Organization*, 81:646–656, 2003.

Hong Xuan, Richard Souvenir, and Robert Pless. Deep randomized ensembles for metric learning. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, Proceedings, Part XVI*, pages 751–762, 2018. doi: 10.1007/978-3-030-01270-0_44. URL `https://doi.org/10.1007/978-3-030-01270-0_44`.

Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012. URL `http://arxiv.org/abs/1212.5701`.