* ~~Motivation: summary of user needs and motivating questions~~
  + *~~There is a clear focus, and a developed explanation of the problem, and a reasonable response is proposed.~~*
* ~~Data: summary of data, data types, and data preprocessing~~
  + *~~Data sources match the problem statement and are appropriate.~~*
  + *~~Check out various Visualization and ML papers to see how they describe their data~~*
  + *~~These descriptions should be very explicit so someone could read your paper and properly reproduce your results~~*
* ~~Data Analysis: summary of interesting results~~
  + *~~There is extensive exploration of the data.~~*
* Task Analysis: summary of task table
  + *Clearly describes domain tasks, processes, goals and abstract tasks for domain problems*
* ~~Model Description: summary of modeling options, performance evaluations, model performances, and best performing selected model (750-1000 words minimum)~~
  + *~~Treat this like a mini-ML paper. Check out papers from ACL, NIPS, etc to get an idea of how they describe their models, justify their modeling and evaluation decisions and tabulate results~~*
  + *~~These descriptions should be very explicit so someone could read your paper and properly reproduce your results.~~*
* Design Process: sketches and design choices to justify final visualization
  + *Evidence of iterative improvement.*
  + *Logical discussion of design choices grounded in theory from course.*
  + *Discusses feedback from usability testing*
* Final Visualization: final visualization, design justifications, packages utilized for coding, and UI walk-through, a link to the running demo on your own server
* Conclusion: short summary of work completed and areas for improvement/future-work.
  + *Meaningfully wraps up project and has good future directions.*

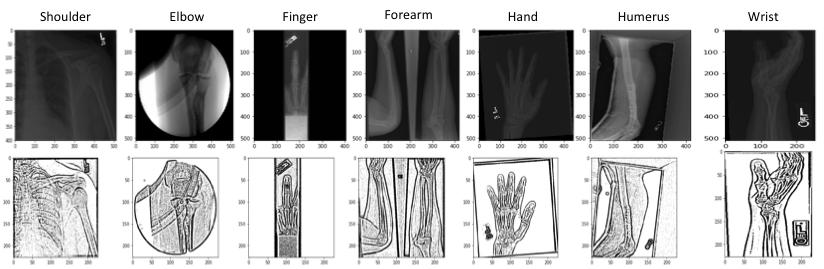
**1. Motivation:**

Large high scale musculoskeletal radiographs have played a pivotal part in the intersection of machine learning and the field of medicine. Existing research leverages radiographs as input data for various machine-learning frameworks in order to help medical practitioners detect diseases at early stages (ex. fractures, cancer, etc.). With radiographs playing a pivotal role in detecting abnormalities, it is important that the framework is sound and mimics the thought process of true radiologists. Because of the latter, we look to implement a model framework composed of multiple convolutional neural networks (CNNs), which will first diagnose the body parts and then feed another model constructed specifically for the given body part, ultimately returning the predicted label for an image. Not only do we look to constructing this model framework, but also we look to integrate it into a visual analytic system, enabling radiologist or data scientist to use.

**2. Data:**

The dataset utilized for detecting abnormalities is the MURA dataset released by Stanford researchers and is compiled from 14,863 studies, containing 40,561 images. The original dataset is originally partitioned such that each the training set contains ~90% of the data and the test set contains ~10%. Each set of data is categorized using a folder structure composed of body part, patient number, study, and image(s). In order to access all images, the paths are required; however, this is found in separate csv file, where each record represents an image. The record also includes, the abnormality label, but does not include a label for the body part. In order to extract that information, we used the path, and assigned a number for each body part. The labels range from 0 to 6, which correspond to following in the specified order: shoulder, humerus, finger, elbow, wrist, forearm, and hand.

For preprocessing, we needed to determine a set structure for the images, since the images varied in terms of channels and size. Initially, we were going to operate with 3 channels since some of the images came in as RBG and we wanted to maintain as much granular information as possible, but upon further inspection, we realized that the RBG photos had the same value per channel. Due to this, we converted everything to gray scale. Next, the CNN models used determined the width and height. Although images do not need to be a fixed size for convolutions, in order to feed a dense layer, the convolutions must result in a fixed length vector. Given our final CNN models leveraged the Alex Net and Dense Net architecture, two transformations were applied. For the Alex Net model, images were scaled to 227 x 277, and for the Dense Net model, they were scaled to 224 x 224. In all each image was represented by 51,529/50,176 quantitative variables. Next, in order to extract more information, we applied a form of object segmentation called adaptive thresholding. Adaptive thresholding is able to intensify key features of radiographs [1], which will heavily aid in detecting abnormalities. We used a mean adaptive method with a binary threshold type, assigning a max value of 255 for pixels/variables that exceeded the calculated threshold. The pixel neighborhood used to calculate the threshold was 11 with a constant of 2. Below are examples of our preprocessing steps:



Finally, the last step required is normalization with respect to the IMAGENET dataset, which has a mean of .456 and standard deviation of .225. This is required because we will be leveraging pre-trained models and this prevents vanishing gradients in the CNN.

**3.** **Task Analysis:**

In terms of how to integrate this framework with an interactive visualization analytic system, we propose utilizing a tree-like structure to depict the performance of the model. The initial branches will represent the body part classification task with point marks representing images. The true label of the body part is encoded using the hue color channel. The visualization will employ motion in order to depict which body part the image is classified as, eventually flowing into a stacked bar charts. The stacked bar chart can then be zoomed on to explore the misclassification distribution further. Dependent on which body part the images are classified as, the appropriate model will classify the images as abnormal or normal, with a summary of the results present in a confusion matrix.

<insert image to show how it would be encorporated>

Also, incorporated with the visualization are hyper parameter tuning charts, which enable the user to modify any of the abnormality detection models. A given set of hyper parameters can be considered λ such that each model

**4. Model Descriptions:**

Image classification is a common task in the world of machine learning with the optimal approach revolving around CNNs. The CNN exploits the pixel dependency in images by utilizing filters as sliding windows and pooling the values in order to reduce the dimensionality of the problem, while maintaining the crux of the information. Although powerful, there are numerous ways to construct a CNN in regards to its architecture. Because of this, it is fairly easier to construct a poor performing model than it is to construct a successful one. In order to avoid this issue, we looked at several state-of-the-art CNNs that have proven success on the complex ImageNet dataset. The only modification necessary was in the last dense layer where the output was changed from 1,000 to 7 in order to accommodate the number of layers.

The model framework implemented is a form of an ensemble where body parts are classified first and then the abnormality. The general premise behind this approach is that each body part is generating noise for another body part, making it difficult for one model to detect all of the abnormalities. By dividing the model, up we trade off more data for specificity. In theory, the sub-models can be considered radiologists who are specialists.

**4.1 Body Part Classification**

The first model was chosen with computational time taken into account. Currently, there are numerous state-of-the-art architectures for CNNs; however, to train each model with +30,000 images is time consuming so we trained the body part classifier on a few considering average loss and run time. Architectures considered were Dense Net, VGG net, and Alex Net. The former produced the longest run time with each epoch taking several hours, which is the reason why it was not considered pragmatic. The performance between Alex Net and VGG Net was negligible; however, Alex Net performed exceptionally better in terms of run time per epoch.

Prior to training, the hyper parameters needed to be tuned and an optimization algorithm needed to be chosen. With no prior research in terms of classifying the MURA dataset into body parts, random search was employed. Random search has been proven to perform better than grid search and can achieve within 5% of the local minima using 64 trials for two hyper parameters [4]. The two hyper parameters considered were learning rate and weight decay with each being drawn geometrically from a set. Drawing geometrically from a set A and B means drawing uniformly in the log domain between log(A) and log(B), exponentiating to get a number between A and B. The set for the learning rate was .0001 and .01, while the weight decay was chosen from 3.1e-7 and 3.1e-5. Next, the two optimization algorithms considered were Adam and Stochastic Gradient descent. The latter was discarded after 16 trials, as all of the results were significantly less promising than all of the Adam results. Below are the top 5 trials:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial** | **Optimization Algorithm** | **Learning Rate** | **Weight Decay** | **Average Loss** | **Accuracy** |
| 3 | Adam | 0.0013 | 1.02E-06 | 1.06 | 66.97% |
| 46 | Adam | 0.0019 | 8.22E-07 | 1.10 | 65.66% |
| 63 | Adam | 0.0012 | 2.74E-05 | 1.07 | 65.56% |
| 22 | Adam | 0.0009 | 4.69E-07 | 1.16 | 62.79% |
| 60 | Adam | 0.0014 | 2.13E-05 | 1.18 | 61.72% |

Based on the results, the top performing hyper parameters were used. The model was trained for 30 epochs with each epoch iterating over the entire training set with a batch size of 1,024. Since this a multi-task problem, a cross entropy loss function was used. The final results for both the training set and test set is below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Shoulder | Humerus | Finger | Elbow | Wrist | Forearm | Hand |
| Shoulder | **99.29%** | 0.18% | 0.00% | 0.36% | 0.00% | 0.18% | 0.00% |
| Humerus | 5.21% | **80.90%** | 1.39% | 5.21% | 0.69% | 6.25% | 0.35% |
| Finger | 0.22% | 0.43% | **93.28%** | 0.22% | 1.30% | 0.43% | 4.12% |
| Elbow | 1.94% | 0.43% | 0.22% | **94.19%** | 1.94% | 1.29% | 0.00% |
| Wrist | 0.46% | 0.15% | 1.37% | 0.30% | **95.29%** | 1.37% | 1.06% |
| Forearm | 0.00% | 5.65% | 1.99% | 9.30% | 6.31% | **76.74%** | 0.00% |
| Hand | 0.22% | 0.00% | 0.43% | 0.00% | 1.52% | 0.43% | **97.39%** |
| **Total** | **92.80%** | | | | | | |

**4.2 Abnormality Detection:**

The abnormality detection model implemented is a 169 hidden layer network, as existing research has proven it’s the optimal architecture for the task [3]. Other architectures were initially attempted, but the models such as Alex Net and VGG Net did not learn, consistently revolving around ~50%.

The approach for hyper parameter tuning follows the same as the previous model for body part classification. In this case, we did not consider stochastic gradient descent, as the leading model employs Adam using the same architecture [3] . Another difference from the other model is that all models were maintained such that none were discarded. There were 16 trials for 7 body parts, resulting in 112 models. We maintained all models so that a user could see how the different hyper parameter sets influence the test information. Below are the top performing abnormality detection models per body part:

**Citations:**

[1] Nomir and Abdel-Mottaleb, 2005 O. Nomir, M. Abdel-Mottaleb

A system for human identification from X-ray dental radiographs?

Pattern Recognit., 38 (8) (2005), pp. 1295-1305

[2] 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;

1st Conference on Medical Imaging with Deep Learning (MIDL 2018)

[3] “HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition” Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, Yizhou Yu; The IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2740-2748

[4] <http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf>