Abnormality Detection in X-Rays of Upper Extremities Using Deep Learning

Gargee Jagtap (wra2jv), Ali Rivera (wat6sv), Anne Louise Seekford (bng3be)

Introduction

Of all diagnostic imaging techniques, radiographs (X-rays) are the most commonly used and widely available. Even when doctors recommend more advanced tests, such as CT or MRI scans, they still order an X-ray as the first round of diagnosis (*X-rays, American Academy of Orthopedic Surgeons*). The AAMC (Association of American Medical Colleges) predicts there will be a shortage of 17,000 - 42,000 radiologists in the next decade. In the United States, each year on average the number of imaging studies increases around 5%, while the number of radiology residency positions only increases by 2% (*Makary & Takacs*). This shortage of radiologists can cause increased turnaround times for results to be generated, and one study found that radiology imaging delays were an independent predictor of the length of a patient's hospital stay. This delay in critical information can negatively impact patient outcomes significantly. Furthermore, these imagining delays increase the burden on not only patients, but healthcare systems, insurance providers, etc. This problem of the radiologist shortage along with the problem of limited access to skilled radiologists is what motivated this project.

Our dataset contains information on musculoskeletal conditions, which are the most common cause of severe, long-term pain and disability. They are characterized by pain that is usually persistent and limitations in mobility and dexterity, thus reducing people's ability to work and function on a daily basis. These conditions range from short-lived to long-term, with common ones including rheumatoid arthritis, osteoarthritis, tendinitis, and carpal tunnel (*Musculoskeletal health, World Health Organization)*. More than 1.7 billion people worldwide are affected by musculoskeletal conditions, and they are the cause of over 30 million emergency department visits annually. MURA (musculoskeletal radiographs) is one of the largest public radiographic image datasets of bone X-rays. Our goal for this project is to see if our algorithm can determine whether an X-ray study is abnormal, and diagnose at the level of experts. We want to see if our model can perform as well as the radiologists on the task, in hopes of the findings leading to advances in medical imaging technologies.

Dataset

The data that will be used in this project comes from the Stanford Machine Learning Group, and was found on the public data and research-sharing site, Papers with Code (MetaAI). The MURA dataset is a comprehensive set of 40,561 (multi-view) musculoskeletal radiographs, i.e. x-rays. In total, the data is a collection of 14,863 studies coming from 12,173 patients. Each x-ray image belongs to one of the seven standard upper extremity radiographic study types as seen in figure 1: either an elbow, finger, forearm, hand, humorous, shoulder, or wrist (Rajpurkar, P., Irvin, J., et al). Each bone image is labeled as negative or positive, which corresponds to *Normal* or *Abnormal*.

When downloaded, the MURA dataset is separated into train and validation files, as the Stanford ML Group pre-split the data. As is, the MURA training data consists of over 26,398 musculoskeletal radiograph images; broken down into images of 3,801 finger patients, 3,964 hand, 1,743 forearm, 6,728 wrist, 5,516 shoulder, 1,180 humerus, and 3,466 elbow patient multi-view x-ray images. The validation data contains over 3,000 images, also separated by bone type. There exists an obvious discrepancy in this data, as we were not provided the test set. Our group is responsible for either, a) reformatting and re-splitting the data into train, valid, and test, or b) finding an additional dataset for testing purposes.

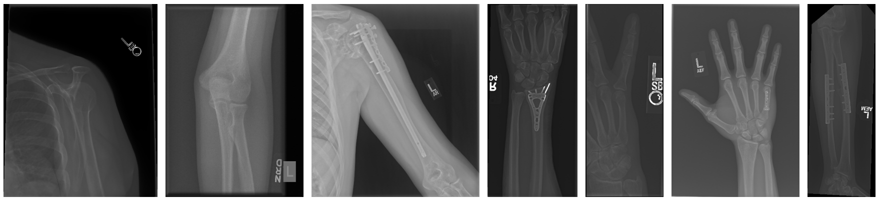


Figure 1: X-ray images from each of the seven classes (shoulder, elbow, humerus, wrist, finger, hand, forearm)

After downloading all of the data, we started our cleaning process by unzipping all of the folders. Within each body part folder, we unzipped the patient folder, and then the study folder to extract all of the images. We labeled each image to contain the patient number, study number, the image number, as well as a positive or negative label which corresponds to whether the classification was normal or abnormal. The format for the image name appears as such in the dataframe: patient10791_study1_negative_image1.png. We then created a dataframe for each bodypart where each image was a row, and there were three columns: one column for the image name, one column for the label that represented the binary classification (0 for negative and 1 for positive), and the final column had a label for the body part. After creating these 7 dataframes for the training set and 7 dataframes for the validation set, we concatenated them respectively to get our final dataframes. The training data frame now has 36,978 rows and the validation data frame has 3,197 rows. As mentioned above, when it came to the test set of this data, we decided to do a train/test split using a ratio of 80:20. Our final training dataset now has 29,582 rows and the test dataset has 7,396 rows.

Related Work

This data was used by Stanford University's Computer Science, Medicine, and Radiology departments to develop a Convolutional Neural Network that performed at least as well as a radiologist on all body parts. The model developed from this research is a DenseNet with 169-layer Convolutional Neural Network. The DenseNet architecture is a fully connected network with all preceding layers feeding into each layer. The final fully connected layer is a single output classification with a sigmoid nonlinearity. The model uses pretrained weights from the ImageNet dataset, an Adam optimizer with default parameters and an initial learning rate or 0.0001 that decays by a factor of 10 when the validation loss plateaus after an epoch, and mini-batches of size 8. The validation was evaluated using weighted binary cross entropy loss. The final model was an ensemble of the 5 models with the lowest validation loss.

Model performance was compared to the majority vote radiologist performance with Cohen's kappa statistics for each image type as well as overall. The model performed as well as the best radiologist performance for the finger and wrist images, but worse on the other image types and overall. When compared on an ROC curve, the overall model performance with varying classification thresholds is seen to be worse than each of the radiologist's performance. The AUROC is 0.929, with 0.815 sensitivity and 0.887 specificity.

There are other studies using medical imaging that have been successful in using CNNs to classify different pathologies. One study using several deep CNNs to classify multiclass pathology related to eight different disease classes in chest X-rays compared performance amongst models as well as between classes in each model.

This dataset is popular on Kaggle for training CNN models. One example by user @Alkoby uses a two stage pipeline to improve model performance, where x-rays first pass through a model that classifies them according to body part, and based on that classification the image moves onto a model for that specific body part. For example, an image could enter the body part classification model and be determined to be an x-ray of an elbow. The image would then enter the model to classify elbow x-rays and normal or abnormal. We plan to use this system design to improve the accuracy for the body parts that had low performance in the original publication.

Technical Approach

Our model pipeline consists of two main stages, first, a body part image classification, followed by a binary normal/abnormal bone classification.

Body Part Classification

The image classification consists of seven classes, one for each body part: elbow, finger, forearm, hand, humerus, shoulder, wrist. To begin the image classification, we used a pre-trained DenseNet121 model with 121 filly connected layers and a final linear layer with an output size of 7 for each of the 7 body parts in the data. Within this model, a softmax activation function with an Adam optimizer was used, followed by a CrossEntropy loss function. Upon initial model creation, we do not employ early stopping, however, may implement if the model seems to overfit. The learning rate was set to 0.001, and the model ran for 15 epochs.

Normal/Abnormal Classification

The second stage of the model pipeline is the normal/abnormal classification. Following the body part classification, the images from the seven classes will be used as input. This classification model consists of two classes for each image as seen in figure 2: normal (0) and abnormal (1). To begin the bone status (normal/abnormal) classification, we also used a pre-trained DenseNet121 model with 121 filly connected layers and a final linear layer with an output size of 2 for the *normal* or *abnormal* classification. Similar to the body part classification, this model uses a softmax activation function with an Adam optimizer, followed by a CrossEntropy loss function. Again, upon initial model creation,

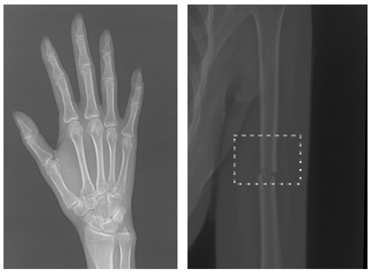


Figure 2: X-ray image from 0: normal (left) and 1: abnormal (right) class

we do not employ early stopping, however, may implement if the model seems to overfit. The learning rate was set to 0.001, and the model ran for 15 epochs.

Experiments

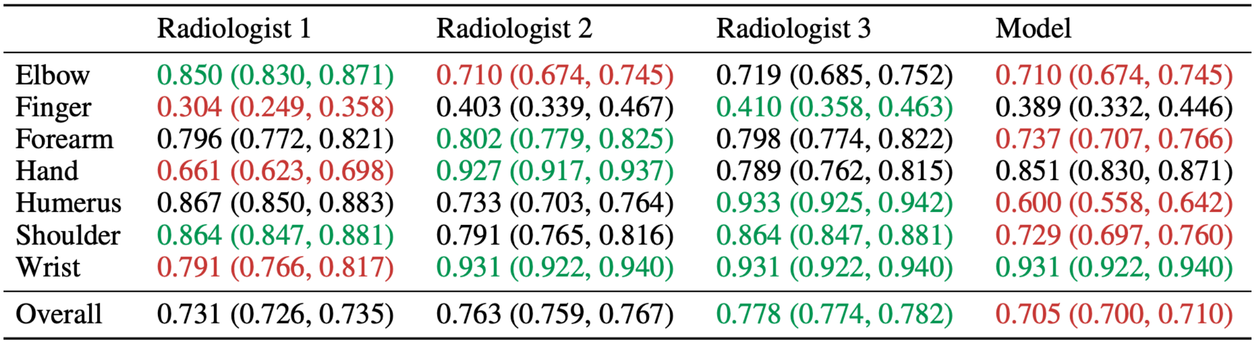
Evaluation

To evaluate model performance, we return model accuracy (in %) for both the body part classification and normal/abnormal classification models. As aforementioned, there are seven classes in the body part classification model, one for each body part, and two classes in the binary normal/abnormal model.

In previous work, the binary classification of *abnormal* finger and wrist images was successful. In our model, we classify all seven body parts, and aim to improve overall performance across all classes in comparison to the aforementioned work. Looking at accuracy with respect to the distinctive upper extremities, we will be able to evaluate the other three classes in comparison to the original paper's finger and wrist accuracies, with a goal of establishing the five worst classifications (from prior work) extremely close.

Previous Results

As aforementioned, in previous work done with this data model results were compared with radiologists using Cohen's Kappa statistic, which expresses the agreement of each radiologist and/or model with the gold standard, defined as the majority vote of a disjoint group of radiologists (Wang, et al). As shown in Figure 3, the best and worst performances are highlighted in green and red, respectively. In previous work, the model performed well with finger, hand, and wrist images. The model's worst performance is shown in the forearm, humerus, and elbow images.



	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)

Figure 3: Results from previous work, compared using Cohen's Kappa Statistic

Results

In stage one of our model, the body part image classification yielded good results, with a 97.8% accuracy rate. Meaning, 97.8% of these images were correctly classified as a bone from one of the seven classes. Again, following this classification, the images within these classes were fed into the binary normal and abnormal bone classifier. The results here, although overall

Body Part	*Elbow*	*Finger*	*Forearm*	*Hand*	*Humerus*	*Shoulder*	*Wrist*
Accuracy	81.5%	76.6%	72.1%	79.9%	73.4%	76.6%	83.1%

Table 1: Stage 2 (Normal/Abnormal Classification) Accuracies

accuracies are not as high as the former body part classifier, were promising as well. As shown in Table 1, our model most accurately classified wrist and elbow images, with 83.1% and 81.5% accuracies respectively. The classes of images with the worst performance in our model were associated with the forearm and humerus.

When comparing our results to the previous, successful work of Wang et. al, it is clear we significantly improved the results of elbow images, as the elbow was one of the classes with the worst performance in their model, and one of the best in our work. Our models performed in a similar fashion with regards to the wrist, forearm, and humerus images.

Conclusions

Future Work and Model Improvements

Due to the similarity in poor performance in forearm and humerus images in both projects, we think it would be extremely beneficial to expand the size of the dataset. Images classified as forearm or humerus had the lowest number of sample images, at roughly 1,500 each, while other classes, such as elbow, consisted of more than double that, with an approximate 3,500 images. While increasing the overall sample size may also improve model performance, increasing image classes with a significantly lower number of images may make more of an impact.

Although we do believe we are on the right track with the two-stage classification model for this research, there do exist improvements to be made in our models as well. For future work, we believe further modifying our models will increase performance - specifically, further customization within the normal/abnormal classification models. It may prove advantageous to create various models with various specifications, custom to each body part class. This may result in seven distinct models under the umbrella of the “Stage 2” normal/abnormal classifier.

Lessons Learned

Through this project, we were able to see how advanced technology, in this case densely connected convolutional networks, has the potential to make a profound impact in the real-world. If we, as mere students, are able to create a fairly successful model to help x-ray technicians, hospitals, and patients during a period of radiologist shortage, it is exciting and inspiring to know the large potential for these algorithms moving forwards.

Code and Data

If interested in learning more or pursuing improvements to this project, feel free to visit our GitHub repository, where the entirety of our code is located. The data for this project is publicly available on the Stanford Machine Learning Group website.
Github Repository: https://github.com/alseekford/xray_abnormalities
Data: https://paperswithcode.com/dataset/mura

References

Alkoby. (2023, January). Alkoby/bone-fracture-detection: Bone Fracture Detection Using Deep Learning (RESNET50) - final project in the fourth year of the degree. GitHub. Retrieved April 17, 2023, from https://github.com/Alkoby/Bone-Fracture-Detection

Alzubaidi, M. S., Shah, U., Zubaydi, H. D., Dolaat, K., Abd-Alrazaq, A. A., Ahmed, A., & Househ, M. (2021). The Role of Neural Network for the Detection of Parkinson's Disease: A Scoping Review. *Healthcare*, *9*(6). https://doi.org/10.3390/healthcare9060740

American Academy of Orthopedic Surgeons. (2017, June). *X-rays, CT scans, and MRI scans.* OrthoInfo. Retrieved March 15, 2023, from https://orthoinfo.aaos.org/en/treatment/x-rays-ct-scans-and-mris/#:~:text=X%2Drays%20(radiographs)%20are,get%20an%20X%2Dray%20first

MetaAI. (n.d.). *Papers with Code - MURA dataset*. Papers With Code. Retrieved March 15, 2023, from https://paperswithcode.com/dataset/mura

Mina S. Makary, M. D., &; Noah Takacs, B. S. (2022, January 20). *Are We Prepared for a Looming Radiologist Shortage?* Diagnostic Imaging. Retrieved March 15, 2023, from https://www.diagnosticimaging.com/view/are-we-prepared-for-a-looming-radiologist-shortage-

Rajpurkar, P., Irvin, J., et al., (n.d.). *MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs*. Stanford ML Group. Retrieved March 15, 2023, from https://stanfordmlgroup.github.io/competitions/mura/

Tsang, S.-H. (2018, November 25). *Review: DenseNet - Dense Convolutional Network (Image Classification)*. Medium. Retrieved March 15, 2023, from https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., &; Summers, R. M. (2017, December 14). *Chestx-Ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*. arXiv.org. Retrieved March 15, 2023, from https://arxiv.org/abs/1705.02315

World Health Organization. (2022, July 14). *Musculoskeletal health*. World Health Organization. Retrieved March 15, 2023, from https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions#:~:text=Musculoskeletal%20conditions%20are%20typically%20characterized,form%20of%20non%2Dcancer%20pain