CS230

neXt-Ray: Deep Learning on Bone X-Rays

Anfal Siddiqui
Department of Computer Science
Stanford University
anfal@stanford.edu
SUNet Id: anfal

Abstract

As the world population ages, the prevalence of Musculosketal conditions will continue to rise. To meet the need for an automated means of detecting such conditions from X-rays - particularly in developing countries with limited access to radiologists - we introduce neXt-Ray, a model that detects whether upper-extremity Bone X-Rays passed to it as inputs are abnormal. neXt-Ray differs from a prior DenseNet-169-based attempt by the Stanford ML Group by utilizing a single, more memory-efficient 152-Layer Residual Network as its architecture. Our model achieves an AUC of 0.89, as well as a sensitivity of 0.78, a specificity of 0.83, and overall accuracy of 0.81. Compared on the Cohen's kappa statistic, our model outperforms the Stanford baseline and radiologists on finger studies and is comparable to radiologist performance on humerus studies. However, it performs worse than the Stanford baseline and radiologists on wrist, shoulder, hand, forearm, and elbow studies, and overall.

1 Introduction

According to the World Health Organization, Musculoskeletal conditions are the world's leading cause of disability, affecting up to an estimated 1 in 3 people worldwide across all ages (2). These conditions are made up of over 150 different diagnoses, ranging from simple fractures that heal relatively quickly to longer-term conditions such as Osteoarthritis. As the global population continues to age, the prevalence of such Musculoskeletal conditions will only continue to increase, as will associated socioeconomic costs.

As such, the importance of easily diagnosing these conditions, which is often achieved by means of Bone X-Ray scans examined by trained radiologists, will only continue to rise over the next few decades. With only a limited number of trained radiologists in the world, with an especially small number found in developing countries, having an automated means of detecting bone abnormalities from X-Rays quickly and cheaply is critical. To this end, we introduce neXt-Ray, which takes as input a Bone X-Ray scan from the upper extremity, passes it through a Convolutional Neural Network (CNN), and outputs a prediction of whether the bone is normal or abnormal.

2 Related work

CNNs for Medical Tasks: CNNs have been used extensively for medical classification and detection tasks and have performed at levels matching or even exceeding specialists in areas including skin-cancer (6), pulmonary embolism (8), and heart arrhythmia (13). In particular, Residual Networks (ResNet) - the primary architectural focus of this project - have proven successful in binary classification tasks such as detection of Covid-19 in CT images, demonstrating their utility even on the most novel medical conditions (3).

CNNs for X-Rays: CNNs also have an established history of use with X-Ray images - the input to this project's model. CNNs provided promising results when used for disease classification from Chest X-Rays (16) and the CheXNet system exceeded the performance of the average radiologist on the F1 metric for detecting pneumonia in similar X-Ray images (15).

CNNs for Bone X-Rays: The source of inspiration for this project was the work done by the Stanford ML Group on abnormality detection in Bone X-Rays (14). Their 169-layer DenseNet ensemble model matched the best radiologists' performance on detecting abnormalities for both finger and wrist studies, but came up short on elbow, forearm, hand, humerus, and shoulder studies, as well as overall performance. Two other significant contribution made by the team were the production of the MURA Dataset and publishing data on the performance of radiologists' on abnormality detection. Due to the robustness of its reported metrics, the Stanford baseline serves as the standard of comparison for this report.

Other teams have attempted to improve upon the Stanford baseline, such as the 2019 follow-up by Dennis Banga and Peter Waiganjo. They used an ensemble neural network composed of a DenseNet-201, MobileNet, and NASNETMobile (4). Although their model outperformed the Stanford baseline on finger studies, it failed to match its performance in any other region or overall, demonstrating the on-going challenge of abnormality detection in Bone X-Rays. While some more recent attempts have eclipsed the performance of the Stanford baseline (9), virtually all have also used computationally expensive ensemble models, showcasing the need for projects like this one involving more easily deployable single model approaches.

3 Dataset and Features

The MURA (**mu**sculoskeletal **ra**diographs) Dataset was produced as part of the work done by the Stanford ML Group on abnormality detection in Bone X-Rays (1). It is made up of 40,561 multi-view Bone X-Rays from 14,863 studies conducted at the Stanford Hospital between 2001 and 2012 (14). Each was hand-labeled by a board-certified radiologist as normal or abnormal at the time of production. The scans are from one of seven regions: elbow, finger, forearm, hand, humerus, shoulder, and wrist. It is among one of the largest publicly available datasets of radiographs.

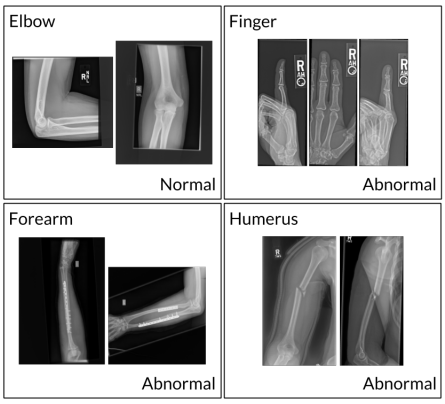


Figure 1: Samples from the MURA Dataset

The Stanford ML Group has already divided the dataset into a training and validation set of sizes 36,808 and 3,197, respectively. The remaining 556 images are reserved as a special test set used to evaluate models submitted to the on-going competition hosted by the group and have not been made public. For the purpose of this project, we randomly split the 3,197 images in the original validation set into our validation set of 2,557 images and a test set made of the remaining 640.

We applied some limited pre-processing to the raw MURA data. We resized all the variable-sized images in the dataset to a consistent 320x320 pixels, converted all of them to Greyscale, and normalized each to have the same mean and standard deviation as the images found in the ImageNet training set (5).

4 Methods

Our model takes as input an upper-extremity Bone X-Ray scan and runs it through a CNN, outputting the probability of abnormality. If the probability is greater than 0.5, we predict that the scan is abnormal.

Figure 2: Overview of neXt-Ray architecture

Our model makes use of a 152-Layer CNN in order to calculate the probability of abnormality. In particular, we use a ResNet architecture. ResNets are made up of "residual blocks", each 3 layers large in ResNet-152. Each of the blocks has a "skip connection," where the input to the first layer in the block is added onto the pre-activation output of its last layer, which makes training of deep CNNs viable (7).

We chose a ResNet architecture as they tend to be less memory-intensive than the DenseNets used in prior implementations while still offering strong performance (12). As one of the goals is to be able to use this model for X-Ray analysis in developing countries, having a less memory-hungry, single model architecture based on a ResNet is useful for deploying it on cheap, offline devices.

We replaced the final fully-connected layer of the ResNet-152 with a sequence of a dropout layer, a fully-connected layer with only one output, and a sigmoid nonlinearity. We made use of a standard binary cross-entropy loss function and utilized the Adam optimization algorithm with the default $betas$ of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We applied random horizonal flips to the images during training.

5 Experiments

DenseNet Baseline: To establish our own baseline, we first implemented a model using a DenseNet-169 architecture. Trained from scratch, it achieved 98% accuracy on the training set and 76% accuracy on the validation set - clearly exhibiting high variance. With this baseline established, we moved on to a ResNet implementation.

Initial Hyperparameter Tuning: A search for learning rate values between 0.001 and 0.00001 found that our best performance on the validation set came from $\alpha = 0.00001$. We initialized our ResNet model with pre-trained weights from a model trained on the ImageNet database and provided by PyTorch (10).

	Model	L2	Dropout	Frozen Layers	Aug.?	Batch Size	Train Acc.	Val Acc.	Variance	Val. Recall
1	ResNet-152	0.01	0	0	No	32	0.816	0.782	+0.034	0.725
2	ResNet-152	0.1	0.5	0	No	32	0.825	0.790	+0.035	0.650
3	ResNet-101	0.1	0.5	0	No	32	0.823	0.803	+0.02	0.681
4	ResNet-152	0	0	0	Yes	32	0.885	0.795	+0.096	0.675
5	ResNet-152	0.1	0.5	148	No	16	0.862	0.806	+0.056	0.768
6	ResNet-152	0.1	0.5	148	No	8	0.835	0.812	+0.023	0.721

Table 1: Results for different experiments. Hyperparameters that were tuned include number of layers in ResNet model, L2-Regularization factor, Dropout factor, number of layers from the start of the model that were frozen, whether data augmentation was used, and batch size. Reported metrics are accuracy on the training and validation sets, the difference between those accuracies, and recall on the validation set. All experiments were performed with a learning rate of 0.00001. The chosen configuration is in green.

Next, we worked to address the variance problem that carried over to the ResNet model, which proved to be quite challenging. We ran dozens of trials to alleviate overfitting without compromising overall

performance, experimenting with L2 Regularization, dropout, batch size, number of layers in our model, freezing pre-trained layers in our model, and data augmentation by adding 50% more data to our training set by randomly choosing images from our training set to rotate by up to 30 degrees. The results on the training and validation sets for a sample of our most successful experiments are shown in 1.

As shown in 1, tuning batch size and the number of layers to freeze had the most impact on both the variance problem and accuracy/recall. A smaller batch size likely best addressed overfitting due to the regularization effect the frequent, noisy updates had. Freezing many layers likely preserved the general learnings the pre-trained weights held from the ImageNet database, simultaneously allowing for higher performance on the validation set and preventing overfitting to the training set.

From 1, the 5th and 6th models had the best performance, overall, on the validation set. We ultimately chose the 5th model, which marginally sacrificed on the overall accuracy and variance in favor of a nearly 5% increase in recall. As having false negatives can be life threatening to the patient in the context of abnormality detection, maximizing this metric was the most important factor in our decision making.

6 Results and Discussion

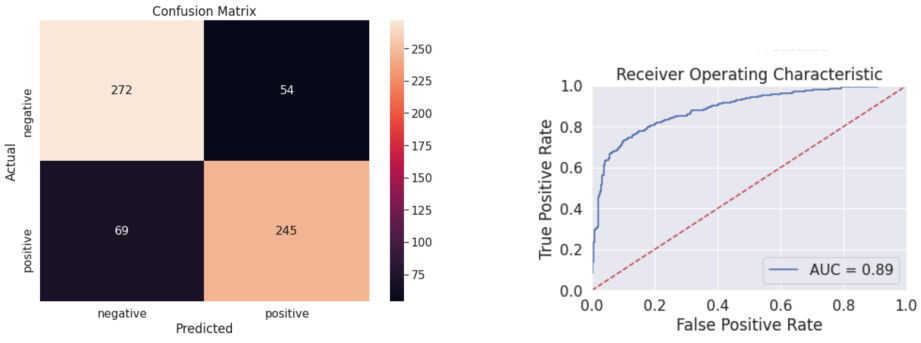
6.1 Model Metrics

We assessed our model's performance on the test set using the standard metrics of accuracy, sensitivity/recall, specificity, precision, and F1. Additionally, we also used the Cohen's kappa statistic, which expresses the agreement between our model and the test set, to better compare with the results from the original MURA paper (14). We calculated these metrics both overall and for each of the different bodily regions represented in the MURA dataset. The results are summarized in 2.

	Counts	Specificity	Sensitivity/Recall	Precision	F1	Accuracy	Cohen's kappa
Wrist	134	0.91	0.81	0.87	0.84	0.87	0.72
Shoulder	113	0.70	0.78	0.75	0.76	0.74	0.48
Humerus	63	0.94	0.82	0.92	0.87	0.89	0.77
Hand	87	0.77	0.78	0.74	0.76	0.77	0.53
Forearm	59	0.84	0.63	0.77	0.69	0.75	0.48
Finger	94	0.86	0.77	0.90	0.83	0.81	0.61
Elbow	90	0.82	0.82	0.82	0.82	0.82	0.64
Overall	640	0.83	0.78	0.82	0.80	0.81	0.62

Table 2: Performance on test set, by upper-extremity region and overall. Best performance was found on humerus radiographs, while worst performance was found on forearm radiographs.

Our model achieved an overall accuracy of 0.81, a sensitivity of 0.78, and Cohen's kappa Score of 0.62. It's best performance, as measured by Cohen's kappa score and F1, was on humerus X-Rays, while it's worst performance was on forearms scans. Additionally, our model achieved an AUC score of 0.89. The Confusion Matrix and ROC Curve can be found in 3.



(a) **(b)**

Figure 3: Confusion Matrix and ROC Curve

6.2 Model vs Stanford Baseline vs Radiologists

	Worst Radiologist Score	Best Radiologist Score	Stanford Baseline	neXt-Ray
Wrist	0.79	0.93	0.93	0.72
Shoulder	0.79	0.86	0.73	0.48
Humerus	0.73	0.93	0.60	0.77
Hand	0.66	0.93	0.85	0.53
Forearm	0.80	0.80	0.74	0.48
Finger	0.30	0.41	0.39	0.61
Elbow	0.71	0.85	0.71	0.64
Overall	0.73	0.78	0.71	0.62

Table 3: Comparisons of the Cohen's kappa score between our model (nExt-Ray), the Stanford baseline, and the best/worst radiologist scores overall and on each bodily region. Best performance for each row is in green and worst is in red. Radiologist and baseline data from (14).

The MURA paper calculated the Cohen's kappa Scores for its model on each body region, as well as reporting it for 3 different radiologists. A summary comparing our model with these results is found in 3. Our model did better than the Stanford baseline and the best radiologist on finger studies (0.61 vs 0.39, 0.41) and did better than the Stanford baseline and the worst radiologist on humerus studies (0.77 vs 0.60, 0.73). On all other body regions, our model did worse than the Stanford baseline and the worst radiologist performance. Overall, our model did worse than the Stanford baseline, which itself did worse than the worst radiologist performance.

Our model's overall sensitivity of 0.78 was close to the Stanford baseline's 0.815, but its specificity of 0.83 was less than the Stanford baseline's 0.887. Our AUC Score of 0.89 was a bit lower than the baseline's of 0.929. Overall, just as with the Stanford baseline, our model does not detect abnormalities in Bone X-Rays as well as radiologists, showing this continues to be a challenging problem and that more complex models may be needed.

7 Conclusion/Future Work

We introduce neXt-Ray, which takes as input Bone X-Rays from 7 upper-extremity regions of the human body (wrist, shoulder, humerus, hand, forearm, finger, elbow) and detects whether or not they are abnormal. Our model, which uses a single ResNet-152 architecture, outperforms the DenseNet-169 baseline produced by the Stanford ML Group and the best radiologists on abnormality detection in finger radiographs and outperforms the baseline and the worst radiologist performance on humerus studies. On all other regions and overall, it performs worse than the Stanford baseline and the worst radiologist performance. Moreover, with a recall value of 0.78, neXt-Ray is in need of further improvements before it is reliable enough to be used in any clinical settings.

The Stanford baseline and follow-ups that have beaten both its performance and the performance of the best radiologists have made use of complex, resource-intensive ensemble models. As such, it may be that the task of abnormality detection from Bone X-Rays requires network architectures that are more complicated than a single ResNet in order to reach clinically-acceptable performance. However, increasingly large or demanding architectures may preclude the possibility of deploying these models on cheap, offline devices in developing countries where they are most needed.

As such, working to further iterate upon a single model implementation is a natural step for future work. One such step may entail using a more robust Loss function than simple binary cross-entropy that weighs positive and negative examples or those from the different body regions differently, which could allow more emphasis to be put on positive examples or those from the regions our model struggled with - like the forearm. Given more team members and resources, we could also try exploring a different single model architecture, like a DenseNet-169 or a DenseNet-201, or even work to gather more X-Ray data to better combat the slight overfitting that was still evident in our final model. Finally, Class Activation Maps could be used to localize the abnormalities in the scans, further boosting the utility of neXt-Ray.

8 Contributions

This was a one-person project, so all aspects of the project were handled by this report's author.

References

[1] Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiograph. https://stanfordmlgroup.github.io/competitions/mura. Accessed: 2020-11-01.

[2] Musculoskeletal conditions. URL: http://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions.

[3] Ali Abbasian Ardakani, Alireza Rajabzadeh Kanafi, U Rajendra Acharya, Nazanin Khadem, and Afshin Mohammadi. Application of deep learning technique to manage covid-19 in routine clinical practice using ct images: Results of 10 convolutional neural networks. *Computers in Biology and Medicine*, page 103795, 2020.

[4] Dennis Banga and Peter Waiganjo. Abnormality detection in musculoskeletal radiographs with convolutional neural networks (ensembles) and performance optimization. *arXiv preprint arXiv:1908.02170*, 2019.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] 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(7639):115–118, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Shih-Cheng Huang, Tanay Kothari, Imon Banerjee, Chris Chute, Robyn L Ball, Norah Borus, Andrew Huang, Bhavik N Patel, Pranav Rajpurkar, Jeremy Irvin, et al. Penet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric ct imaging. *npj Digital Medicine*, 3(1):1–9, 2020.

[9] Kwun Ho Ngan, Artur d'Avila Garcez, Karen M Knapp, Andy Appelboam, and Constantino Carlos Reyes-Aldasoro. Making densenet interpretable a case study in clinical radiology. *medRxiv*, page 19013730, 2019.

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Geoff Pleiss, Danlu Chen, Gao Huang, Tongcheng Li, Laurens van der Maaten, and Kilian Q. Weinberger. Memory-efficient implementation of densenets, 2017. arXiv:1707.06990.

[13] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

[14] 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, 2018. arXiv:1712.06957.

[15] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[16] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[17] Michael Waskom and the seaborn development team. mwaskom/seaborn, September 2020. doi:10.5281/zenodo.592845.