Bone X-Ray Classification For Upper Extremity Radiographs

Riya Ghalyan[1*], Anjali Singh[2], Karuna Kadian[3], Vijay Kumar[4], Disha Yadav[5]

[1*,2,3,4,5]CSE Department, Indira Gandhi Delhi Technical University for Women, Delhi, India

Email: [1*]riya150btcse@igdtuw.ac.in, [2]anjali133btcse@igdtuw.ac.in, [3]karunakadian@igdtuw.ac.in, [4]vijaykumar@igdtuw.ac.in, [5] disha143btcse@igdtuw.ac.in

***Corresponding Author:** Riya Ghalyan
*CSE Department, Indira Gandhi Delhi Technical University for Women, Delhi, India, riya150btcse@igdtuw.ac.in

Abstract
Accurate identification of abnormalities in the upper limb is crucial for effective diagnosis and treatment in musculoskeletal conditions. This research paper presents a deep learningbased approach aimed at surpassing human performance in detecting abnormalities in radiographs of the upper limb. The dataset used for training comprises 40,561 multi-view radiographic images from approximately 14,000 musculoskeletal investigations, with expert radiologists manually labeling each study as normal or abnormal. Through extensive testing and meticulous model tuning, remarkable results are achieved, with an accuracy of 78.57% and an F1 score of 78.43%, rivaling the kappa score reported by the MURA dataset producer. This study contributes to advancing the field by demonstrating the potential of deep learning techniques to outperform humans in identifying abnormalities in upper limb radiographs. The practical applicability of deep learning models is highlighted, showcasing their exceptional accuracy in supporting medical professionals in diagnosing and managing musculoskeletal disorders. This research represents a testament to the ongoing advancements in artificial intelligence, benefiting patient care and outcomes in the medical domain.

INTRODUCTION

Musculoskeletal conditions affect a significant portion of the global population, causing severe and long-term pain, disability, and reduced quality of life. In 2012, it was estimated that 1.7 billion people worldwide were affected by these conditions (Stuart I. Weinstein, 2014). However, accurately identifying abnormalities in the upper limb, a crucial step in effective diagnosis and treatment, can be challenging using traditional radiographic interpretation methods, which rely heavily on human expertise. Moreover, the demanding workload experienced by radiologists can lead to decreased precision and potential risks to patient well-being (Krupinski and Kim, 2010).

To address these challenges, recent advancements in Deep Learning techniques have shown promise in various medical applications, including the detection of Diabetic Retinopathy (Gul- shan, 2016). Leveraging these developments, this research paper presents a deep learning-based approach to surpass human performance in detecting abnormalities in radiographs of the upper limb. The goal is to develop a model that can accurately identify abnormalities, potentially exceeding human accuracy and saving crucial human hours and lives. The methodology employed in this study involves a comprehensive analysis of the model on a region-wise basis, testing the model with different optimizers, learning rates, and layers to identify the optimal configuration. Furthermore, the dataset used for training consists of a large collection of 40,561 multi-view radiographic images from approximately 14,000 musculoskeletal investigations. Expert radiologists meticulously labeled each study as normal or abnormal, providing a reliable ground truth for training and evaluation.

Extensive testing and model tuning were conducted to optimize the performance of the deep learning model.(Ronneberger et al., 2015) The results achieved demonstrate a remarkable accuracy of 78.57% and an F1 score of 78.43%, which rivals the kappa score of 0.778 reported by the MURA dataset producer. These findings underscore the significant usability of contemporary deep learning methods in the medical field, particularly in improving diagnostic accuracy.

In addition to model analysis and testing, this research paper addresses various key aspects of the implementation. This includes the process of loading the dataset and converting it into PyTorch dataloader in batches, which enables efficient and effective data processing. Moreover, a pre-trained DenseNet model was utilized, and the study explores the fine-tuning of some or all of its layers to adapt it specifically for the task of identifying abnormalities in upper limb radiographs..(Huang et al.)

To evaluate the model's performance and understand its behavior, accuracy metrics were created, fit curves were plotted, and data visualizations were generated.(Litjens et al.) These analyses provide insights into the model's strengths, weaknesses, and generalizability. Furthermore, the model was tested with various optimizers, learning rates, and layers to assess its robustness and explore different configurations that optimize its performance.

The research conducted in this paper significantly contributes to the field by demonstrating the potential of deep learning techniques to outperform human capabilities in identifying abnormalities in upper limb radiographs.(Esteva et al., 2017.) The exceptional accuracy achieved, coupled with the practical applicability of deep learning models, showcases their ability to support medical professionals in diagnosing and managing musculoskeletal disorders.

This study serves as a testament to the ongoing advancements in artificial intelligence and its integration into the medical domain, ultimately benefiting patient care and outcomes. By leveraging deep learning methods and advancements in computer vision, healthcare professionals can improve diagnostic accuracy, leading to more effective treatment strategies and improved patient outcomes.

In conclusion, this research paper presents a deep learning-based approach to accurately identify abnormalities in radiographs of the upper limb. The methodology encompasses region-wise analysis, model testing with various configurations, and visualization of results. The achieved results demonstrate the potential of deep learning models to surpass human accuracy, saving valuable human hours and enhancing patient care. The integration of artificial intelligence in the medical domain holds significant promise for advancing diagnostic precision and improving the management of musculoskeletal conditions, ultimately benefiting individuals affected by these conditions.(Shen et al.)

PROBLEM STATEMENT

Accurately classifying musculoskeletal radiograph studies of the upper extremity into normal or abnormal categories is a challenging task that is crucial for effective diagnosis and treatment. Existing approaches often fall short of achieving the level of accuracy demonstrated by human radiologists. The aim of this research is to develop a classification system that can accurately identify abnormalities in musculoskeletal radiographs, matching or surpassing the accuracy level of a radiologist. The dataset consists of 40,561 labeled musculoskeletal radiograph images obtained from over 14,000 studies, providing a comprehensive representation of different body parts. Each study contains multiple images captured from different angles, allowing for a more comprehensive assessment of the patient's condition. By accurately classifying these radiographs, the system will assist medical professionals in their diagnostic process, leading to improved patient care and outcomes in musculoskeletal disorders.

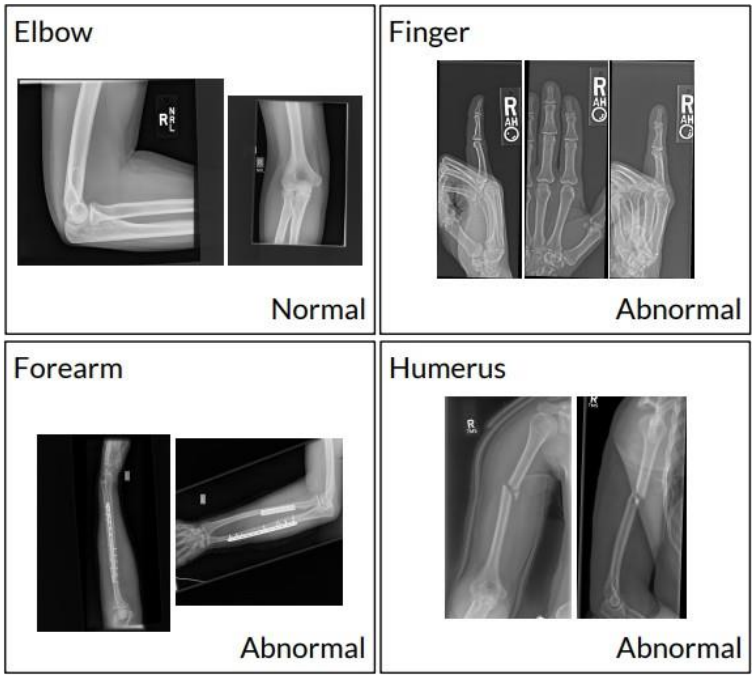

Figure 1: Different studies and their corresponding labels provided by radiologists

DATASET

The Mura dataset consists of 40,561 images from 14,863 studies. The dataset is made public by stanford university with their baseline model. The model contains images from 14k studies of approx 13k patients and it is labeled by the radiologists of those patients as normal and abnormal. We were able to find the data from this site : (stan- ford, 2018).

RELATED WORK

Mura

Considering the importance of deep learning in mind, this paper contains the study for classification of musculoskeletal radiographs into normal or abnormal using deep learning techniques. This paper introduces a dataset of 40k images over 14863 studies done by Stanford University. They took the help of several radiologists to prepare 207 more studies for the test set and proposed a 169-layer densenet baseline model to train the data.

Study	Train		Validation		Total
	Normal	Abnormal	Normal	Abnormal	
Elbow	1094	660	92	66	1912
Finger	1280	655	92	83	2110
Hand	1497	521	101	66	2185
Humerus	321	271	68	67	727
Forearm	590	287	69	64	1010
Shoulder	1364	1457	99	95	3015
Wrist	2134	1326	140	97	3697
Total No. of Studies	8280	5177	661	538	14656

Figure 2: All regions and their corresponding number of studies

The authors were able to achieve an AUROC of 0.929. The model performs really well for the finger and wrist region. However, their performance for the elbow, forearm, hand, humerus, and shoulder was not at par with that of the finger and wrist region. (Rajpurkar, 2018).

DenseNet

In recent works, we have seen that convolutions networks are becoming deeper and deeper to get more accurate. However as the input or the gradients pass through many layers, they can vanish or "wash out" by the time they reach the end (or beginning) of the network. This is called the vanishing gradient problem. Many solutions have been proposed in recent times such as ResNets (Kaim- ing He and Sun., 2015.) and Highway Networks (Rupesh K Srivastava and Schmidhuber., 2015.), all of which try to make shorter connections between layers close to the input and output. DenseNet embraces this observation by connecting each layer to every other layer in a feed-forward fashion. The traditional convolutional networks with L layers have L connections one between each layer, in DenseNet each layer is connected to all layers after it, having $^{L(L+1)}$ direct connections. Since there is no need to re-learn redundant feature maps it requires fewer parameters too. DenseNet obtains significant improvements over most state-of-the-art models. (Gao Huang and Weinberger., 2018).

ImageNet

ImageNet is a collection of the majority of the 80,000 sysnets of WordNet structure with an aver-age of 500-1000 clean and full resolution images of those words. This provides a way to index, retrieve, organize, and interact with images and multimedia data. The current 12 subtrees consist of a total of 3.2 million clearly annotated images spread over 5247 categories. On average, over 600 images are collected for each sysnet. The MURA model initialized the weights for its model with the weights of a pre-trained model on ImageNet. (Jia Deng and Li., 2009.).

ChestX-ray8

This paper presents a new chest X-Ray database and tries to classify the X-Ray into 8 disease classes. The paper contains X-ray radiographs of people who were suffering from one of the following 8 diseases : Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia and Pneumathorax and their labels based on the radiologist's feedback. Then they try to train a model which can classify the images into the above 8 categories. Their multi-label CNN architecture is implemented using Caffe framework. They are using imageNet pre-trained models i.e. AlexNet, GoogLeNet etc. Their model was able to recognize Cardiomegaly and Pneumothorax labels with an avg AUC of 0.8 while other labels had a low AUC of approx 0.6. This dataset and model are good benchmarks for further improving the fullyautomated medical diagnosing systems. (Wang and Summers, 2017.)

Learning Deep Features for Discriminative Localization

The following paper presents a technique to localize the discriminative regions of images. A vital step in the classification problem is to identify the discriminative regions of the image. Once the discriminative regions are identified, the model can then use this information to classify the whole image into one of the given labels. Besides its utility in learning, this localization also leads to better model interpretation by users. In the paper, Global Average Pooling (GAP) (with adequate tweaking), clubbed with a Class Activation Map (CAM) has been used to identify the prominent regions in a single forward pass. The authors of the (Rajpurkar, 2018), also use the following technique to highlight the salient regions of radiographs which contribute the most to their abnormality detection prediction. (Zhou and Torralba, 2016)

Current Applications and Future Impact of Machine Learning in Radiology

The following paper walks through the possible (and currently used) application of ML/DL in radiology. With the likes of Detection and Interpretation of Radiology findings, the authors enlighten the readers about the promises of ML/DL techniques and the future they could hold. Additionally, reviews of a number of currently used ML/DL techniques in radiology is also presented. The current challenges of such techniques and the expected future (along with possible extension) for them are also well discussed. (Garry Choy, 2018)

EXPERIMENT

Experimental Methodology

We investigated carefully in this study to find the best model architecture and hyperparameter values for our task. By experimenting with several models and hyperparameter combinations, we aimed to create a model that performs well. The DenseNet architecture was used as the basis model for the tests, with additional layers being placed on top. The image was prepossessed by resizing them to 224 * 224, followed by normalization between -1 to 1. For training batches of 64 were made and sent to the model. The model consisted of DenseNet followed by a number of fully connected layers. Different weight initialization techniques like, He and Xavier were also employed. Different combinations of Batch Normalisation, Dropout (varied layers, different probabilities) Optimiser and learning rate were also enforced.

Model Architectures

We used a variety of changes to evaluate the influence of model complexity. After the DenseNet basis, some models had just one completely connected layer, while others had up to three of these levels. We were able to investigate the trade-off between model capacity and computing efficiency.

Freezing Layers

We experimented with a number of combinations to examine the impact of freezing layers. This required freezing all layers, just certain subsets of layers, and just the extra layers that were put on top of the DenseNet basis. In situations where transfer learning is advantageous or when computational resources are constrained, layer freezing can be helpful.

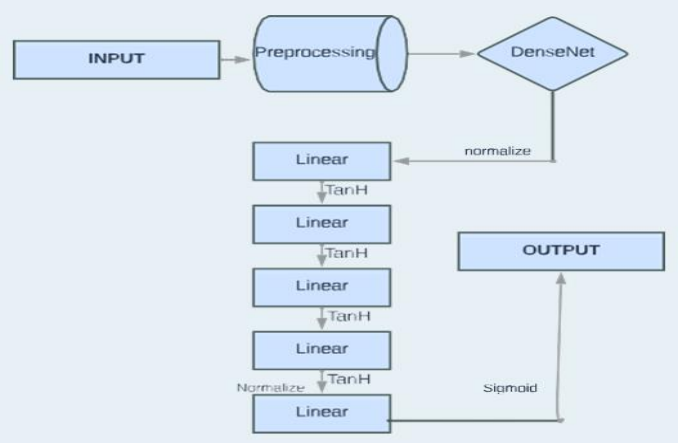

Figure 3: Model Architecture

Weight Initialization Techniques

We also looked into the function of weight initialization strategies like He, Xavier, and Random. To comprehend their effect on the model's convergence and performance, we combined these strategies with fine-tuning other hyperparameters like the learning rate. In order to mitigate the problem of vanishing or expanding gradients during training, weight initialization is essential.

Dropout and Batch Normalization

We tested several Dropout and Batch Normalisation combos to enhance model generalization and avoid overfitting. While training, the regularisation method Dropout randomly deactivates neurons, whereas Batch Normalisation normalizes each layer's activations. We sought to identify the ideal balance between model complexity and regularisation by altering the use of both strategies.

Hyperparameter Search

We used an iterative method of hyperparameter adjustment throughout the experiments. In order to evaluate multiple hyperparameter combinations, we mixed learning rate, regularisation strength, and batch size. As a result, we were able to assess the models performance throughout a range of training times and pinpoint the most promising configurations.

Table 1: Results(Accuracy and F1 Score) of All the Experiments.

Model Configuration	Layers Freezed	Accuracy	F1 Score
Dropout(0.2)+Batch-Norm	All	67.0	67.0
Dropout(0.2)+Batch-Norm + He	None	77.6	77.6
Dropout(0.2)+Batch-Norm+ Xavier	None	58.0	43.0
Dropout(0.2)+ Xavier + lr=10^{-4}	Last 2 Dense block Unfreeze	75.0	75.0
Dropout(0.2)+ Batch-Norm + He	Last 1 Dense block Unfreeze	74.0	74.0
Dropout(0.2)+ Random weight+ lr=10^{-4}	Last 3 Dense block Unfreeze	77.1	77.0
More linear layers on top of DenseNet	None	78.5	78.5

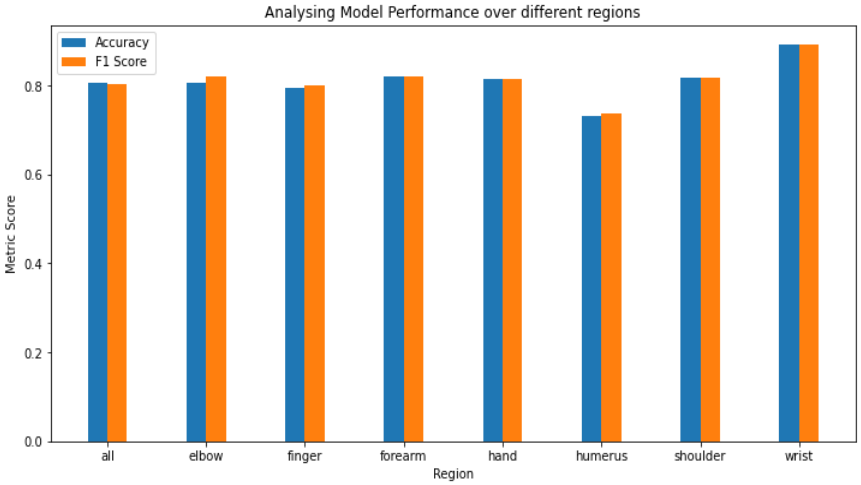

Figure 4: Best model Accuracy and F1 Score on different regions

RESULTS AND ANALYSIS

Our best model was able to achieve best performance of 78.57 % accuracy and 78.43 % F1 Score, which is on par with the MURA dataset producer's kappa statistic of 0.778. Furthermore, even on least performing regions the model accuracy was more than 70%. Comparing the best model with other models, we can see how adding more linear layers, i.e. making the model deep, was crucial for gaining performance. One reason we think might be the cause that freezing certain layers of DenseNet didn't increase performance, could be that the images we had were the x-ray images, but DenseNet was trained over ImageNet which contained general pictures and not the x-ray film like pictures.

CONCLUSION

Our research has provided us with some important findings based on our experiments and the outcomes we received. First of all, with an accuracy of 78.57% and an F1 score of 78.43%, our finest model performed admirably. These outcomes confirm the efficiency of our strategy and are equivalent to the kappa statistic of 0.778 of the creators of the MURA dataset. The accuracy remained continuously above 70%, even in the areas where the model performed the least well.

We found that improving the depth of the model by adding more linear layers was essential in getting higher performance when comparing the performance of various models. This shows that the additional layers increased complexity helped the model better capture and depict the underlying patterns in the data.

Surprisingly, performance did not increase when some Dense Net base layers were frozen. The difference between the characteristics of our x-ray images and the generic images, such as those in the ImageNet dataset, on which the DenseNet was initially trained, could be a feasible explanation for this. The advantages of freezing specific layers may be limited because the pre-trained weights were not perhaps optimised specifically for x-ray pictures.

It is important to note that the precise dataset and task that were employed in our research are what led to these conclusions. It would be beneficial to conduct additional testing and validation on larger and more varied datasets in order to confirm the generalizability of our findings and investigate additional opportunities for raising performance in this area.

There is always some scope for improvement in the models for such predictions, this can be achieved by trying different network architectures, varying the depth, tuning the hyperparameters etc. We can also try to categorize the abnormalities further into their types such as fractures, degenerative joint diseases, hardware, and other miscellaneous abnormalities (like lesions and subluxations).

We can try and train such deep learning models on other medical classifications and try to automate the medical diagnosing system, which can be very helpful to analyse a patient without any human interventions.

REFERENCES

1. Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* pages 115–118.
2. Laurens van der Maaten Gao Huang, Zhuang Liu and Kilian Q. Weinberger. 2018. Densely connected con- volutional networks. arXiv:1608.06993.
3. Mark Michalski Synho Do Anthony E. Samir Oleg S. Pianykh J. Raymond Geis Pari V. Pandharipande James A. Brink Keith J. Dreyer. Garry Choy, Omid Khalilzadeh1. 2018. Current applications and future impact of machine learning in radiology. *Radiology 2018, 288,* page 318–328.
4. Peng Lily Coram Marc Stumpe Martin C Wu Derek Narayanaswamy Arunachalam Venugopalan Subhashini Widner

Kasumi Madams Tom Cuadros Jorge et al. Gulshan, Varun. 2016. Development and vali- dation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama, 316(22)*, page 2402–2410.

5. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolu- tional networks. pages 4700–4708.
6. Richard Socher Li-Jia Li Kai Li Jia Deng, Wei Dong and Fei-Fei Li. 2009. Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 10.1109/CVPR.2009.5206848:248–255.
7. Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun. 2015. Deep residual learning for image recogni- tion. arXiv: 1512.03385.
8. Berbaum Kevin S Caldwell-Robert T Schartz Kevin M Krupinski, Elizabeth A and John. Kim. 2010. Long radiology workdays reduce detection and accommo- dation accuracy. *Journal of the American College of Radiology, 7(9)*, page 698–704.
9. Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud A. A. Setio, Francesco Ciompi, Mohsen
10. Ghafoorian, ..., and Clara I. Sanchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.
11. Irvin J. Bagul A. Ding-D. Duan T. Mehta H. Yang B. Zhu K. Laird D. Ball R. L. Langlotz C. Shpanskaya K. Lungren M. P. Ng A. Y. Rajpurkar, P. 2018. Mura: Large dataset for abnormality detection in muscu- loskeletal radiographs.
12. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. pages 234–241.
13. Klaus Greff Rupesh K Srivastava and Jürgen Schmidhuber. 2015. Training very deep networks. *Advances in Neural Information Processing Systems*, page volume 28.
14. Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248.
15. stanford. 2018. Mura dataset. https://docs.activeloop.ai/datasets/mura-dataset.
16. Sylvia I. Watkins-Castillo. Stuart I. Weinstein, Edward H. Yelin. 2014. The big picture. *BMUS: The Burden of Musculoskeletal Diseases in the United States*.
17. Peng Yifan Lu Le Lu Zhiyong Bagheri-Mohammadhadi Wang, Xiaosong and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv preprint*, arXiv:1705.02315.:248–255.
18. Khosla Aditya Lapedriza Agata Oliva Aude Zhou, Bolei and Antonio. Torralba. 2016. Learning deep features for discriminative localization. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 2921–2929.