ISBN: 978-981-18-7950-0
2023 the 13th International Workshop on Computer Science and Engineering (WCSE 2023)
doi: 10.18178/wcse.2023.06.015

A Comparative Study of Deep Convolutional Neural Networks for Musculoskeletal X-Ray Images

Nay Thazin Htun[1], Khin Mo Mo Tun[2+]

[1] Ph.D. Student, University of Information Technology, Yangon, Myanmar, naythazintun@uit.edu.mm

[2] Supervisor, University of Information Technology, Yangon, Myanmar, khinmo2htun@gmail.com, khinmomotun@uit.edu.mm

Abstract. Technological achievements in science and technology, especially deep learning models have brought about a promising overall performance in many medical image evaluation tasks. Evaluating abnormal conditions from Body part X-rays images is such a need for radiological examination. To fulfill the need of automated detecting of Musculoskeletal conditions, this paper experiments and explores the state-of-the-art deep neural networks which are most commonly applied for image processing. Some of the typical Convolutional neural networks are experimented and evaluated each of the model's classification performance using bone X-ray images. In this study, a typical CNN architecture, CNN architecture with Adabound optimizer, VGG16, ResNet50 and DenseNet architetures are experimented and evaluated. The effectiveness of these Deep CNN models are accessed and compared on the Cohen's kappa statistic and classification accuracy. DenseNet169 Model outperformed the other pre-trained models tested in this study, with the greatest Training Accuracy of 97.98%. However, the standard CNN model with Adam optimizer has a little higher kappa score than the DenseNet169 model. The DenseNet169 Model obtained roughly 20% in terms of Train and Test Loss, which is less than the training loss compared to the other pre-trained CNN models.

Keywords: component; convolutional neural network; MURA dataset; classification;

1. Introduction

In order to identify and cure abnormality of internal body parts, radiologists check X-rays to examine images of internal body sections. A radiologist examines the X-ray images to find an anomaly in the patient's body, but there aren't many radiologists with the necessary training in many remote areas. As a result, a system for automated X-ray image diagnosis is required to provide access to healthcare in regions where radiologists are in short supply [10]. The proper classification of the X-ray images is necessary for the construction of a successful X-ray image diagnosis system. Deep learning techniques are now being applied to create classification models that can categorize and automatically decipher X-ray pictures.

Convolutional Neural Networks (CNNs) [2] are one of the most popular and practical Deep Learning techniques currently being employed, mostly for image recognition. It was first introduced more than 20 years ago, when advancements in computer hardware and network architecture made it possible for Deep CNNs to train highly complexly and quickly. Large Datasets [3] have always been crucial to the development of Deep CNN (DCNN) [2], as have highly improvised Deep Learning techniques. The performance of these techniques is highly dependent on the data availability, data privacy concerns and tedious labelling.

The development of Deep learning techniques has advanced significantly as a result of the use of large, high-quality datasets. MURA, a sizable radiograph dataset encompassing 14,863 upper extremity musculoskeletal studies are tested in this experiment. Using a well-developed model to recognize and detect an image is still a challenge. Especially, Classifying and detecting the Radiograph images is still a challenge and deep neural networks, image recognition and computer vision techniques are widely applied in this area.

In this experiment, regions of interest from Bone X-ray images are tested to classify whether it is normal or not. X-ray images for elbow, finger, forearm, hand, humerus, shoulder, and wrist are used as the object to

+ Nay Thazin Htun. Tel.: +95 95125201.
E-mail address: naythazintun@uit.edu.mm

detect abnormality by utilizing convolutional neural network CNN) and some of the typical pre-trained CNN models. This experiment aims to select the reliable and optimal performance classifier for further processing. Hence, the performance of the employing classifiers is evaluated with different evaluation measures.

The paper is divided into the following sections: section 2 discusses the previous researches performed on deep learning classification, especially with X-ray images. Section 3 discusses the details of the MURA dataset, the typical CNN model and pre-trained CNN models; section 4 contains the model interpretation; section 5 discusses the overall results. Section 3 also compares the performance of the deep learning classifiers.

2. Related Work

A crucial radiological task is determining whether a radiographic examination is normal or abnormal; a scan that is viewed as normal excludes disease and can spare patients from having to undergo additional diagnostic tests or therapies. Given that musculoskeletal problems impact more than 1.7 billion individuals worldwide, the challenge of musculoskeletal anomaly diagnosis is very important. This section provides an overview of Deep learning and several methods for bone abnormality detection related investigations conducted by various researchers.

The MURA data is recognized and classified based on DenseNet169. The MURA dataset [5] is made up of X-ray images. The images' quality vary depending on the resolution, exposure, and object orientation. This sets the bar for all techniques used to categorize and/or find region proposals. Additionally, "identification marks" (flags) that are utilized on X-ray pictures as Left-Right position identifiers or that display patient identification letters also demonstrate to interfere with the classification and recognition of objects.

In their study, Rajpurkar P. et al. [6] employed a 169-layer CNN to identify anomalies in the upper extremities in musculoskeletal imaging, although the model's accuracy in the case of a finger radiograph was only 38.9%. In order to achieve a better outcome than before, I. D. Apostolopoulos [7] proposed a CNN model employing deep transfer learning in his article. Using a DensevCNN model, Verma M. et al. [8] investigated the automatic detection of abnormalities in radiographs of the lower extremities. In [2], the author developed a deep CNN model with VGG-19 and ResNet architecture that had an accuracy of 82.13%. In order to automate the detection process, the author of [5] constructed a deep learning-based model using ensembles of Efficient-Net topologies.

Some studies concentrated on evaluating and fine-tuning current deep learning models using X-ray images. A review of various fine-tuning deep learning models was conducted in [2]. ResNet-50 , DenseNet-201, ResNet-18, ResNet-101 , and VGG19 are some of the pre-trained deep learning models. Mistreatment resulting from misclassification is crucial for that patient in any circumstance. Due to this these circumstances, a precise machine learning model that can perfectly identify the state of the bone in a quicker, more in-depth, and more understandable manner is unquestionably required.

This experiment aims to experiment a Deep Convolutional Neural Network (DCNN) Architecture and training it on the MURA dataset using a typical CNN model with Adam Optimizer and three pre-trained CNN models- DenseNet169, ResNet50 and VGG166 models. Then, accuracy, loss, AUC, and kappa values are used to compare and analyse each model's classification performance.

3. Materials and Methods

The methodology and dataset utilized in this study are covered in this section.

3.1. Musculoskeletal radiographs dataset

The musculoskeletal radiographs datasets (MURA) used in this experiment [5] was introduced by A Stamford Machine learning (ML) group.

This dataset supports abnormality detection of bone X-Rays images. Seven areas of upper body parts namely shoulder, forearm, hand, humerus, elbow and wrist and finger are involved in this X-Ray dataset [9].

This MURA dataset includes 14,863 musculoskeletal studies of the upper extremity. The Expert of the Radiologists have classified these images into two classes. namely normal and abnormal.

The MURA dataset is one of the publicly available datasets of X-Rays images and composed of 40,561 Bone X-Rays images from 14,863 studies. The dataset is made up of 7 distinct categories, namely XR_ELBOW, XR_HAND, XR_HUMERUS, XR_FOREARM, XR_FINGER, XR_SHOULDER, and XR_WRIST, which correspond to the 7 different human body parts.

Table 1 and Figure 2 show the number of body parts for each category. Every category is further divided into subcategories by patient IDs, and each of these contains either Study_positive (which contains normal images) or Study_negative (which contains abnormal images), or both. The distribution of the normal and abnormal image is displayed in Figure 1. Hence, this dataset can be experimented with the task of binary classifications.

Table 1: MURA dataset body-part data

Body_part	Abnormal	Normal	Total	%
XR_ELBOW	2006	2925	4931	13.4
XR_FINGER	1968	3138	5106	13.87
XR_FOREARM	661	1164	1825	4.96
XR_HAND	1484	4059	5543	15.06
XR_HUMERUS	599	673	1272	3.46
XR_SHOULDER	4168	4211	8379	22.76
R_WRIST	3987	5769	9756	26.5

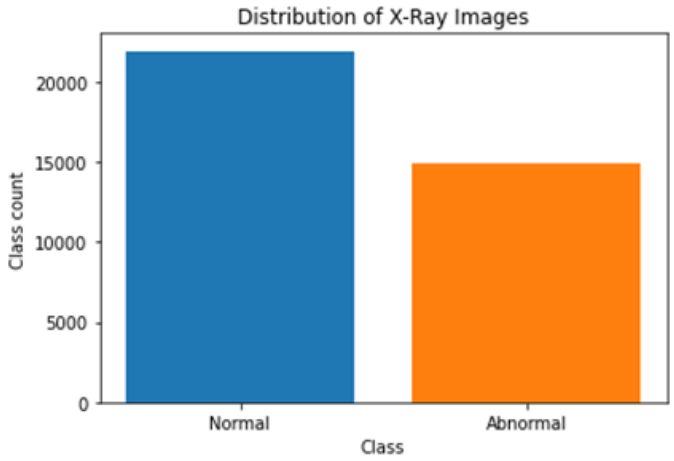

Fig. 1: Class Distribution from the MURA Dataset

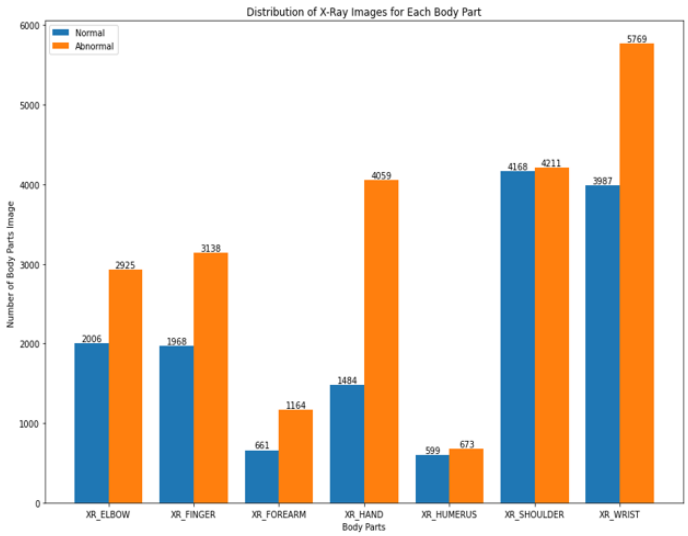

Fig. 2: Body-Part X-Ray Image Distribution from the MURA Dataset

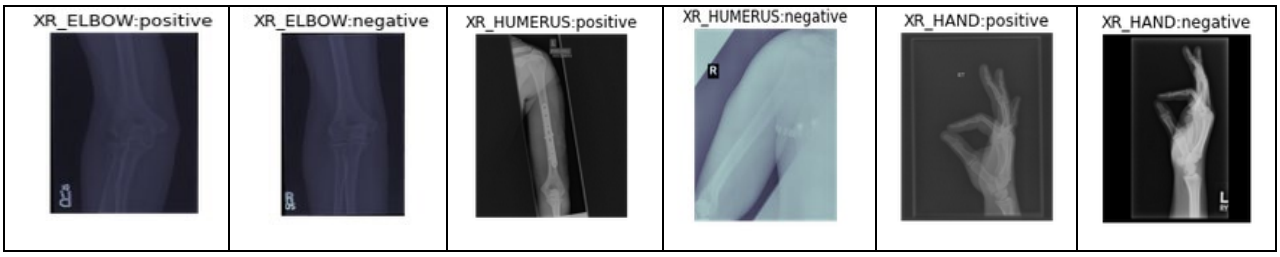

Fig. 3. Sample X-Ray Images from the MURA Dataset

3.2. Convolutional neural networks (cnn) model

Computer vision applications typically use the class of artificial neural networks known as convolutional neural networks (CNN). Each neuron in a neural network gets an input, conducts some computation, and then transmits the output to neurons in the consecutive layer. This is similar to how neuronal networks function in the actual brain.

In contrast to normal neural networks, CNN accepts a multidimensional tensor as input. The standard neural network would need a big number of parameters and be prone to over-tuning, but the CNN design can be significantly better tailored to handle images.

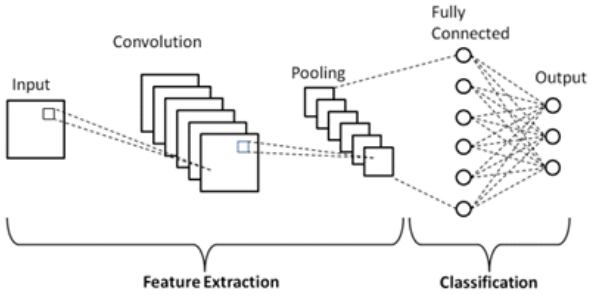

Fig. 4: A Typical CNN Model

3.3. Convolutional neural network architecture of the experiment

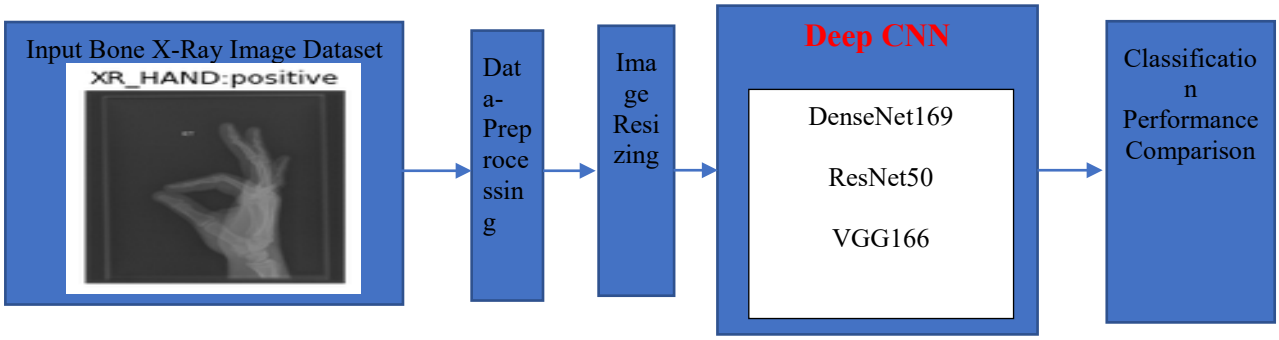

Fig. 5: Network Architecture for the Comparative Study

This experiment aims to experiment a Deep Convolutional Neural Network (DCNN) Architecture and training it on the MURA dataset using a typical CNN model with Adam Optimizer and three pre-trained CNN models- DenseNet169, ResNet50 and VGG166 models. Then, accuracy, loss, AUC, and kappa values are used to compare and analyse each model's classification performance.

A rescaling layer is used to scale the image data values from 0-255 to a float value between 0-1, after the image data has been input. To avoid previously reported internal covariate shifts and vanishing/exploding gradients, batch normalization layers are used throughout the Model to normalize the data.

To extract image features, Convolution Layers and Pooling Layers are used. General kernel, size of 3x3 and rectified linear activation function were employed. Following flattening layer, a string of layers that are all connected together is utilized to extrapolate to a final categorization (fractured or not).

Each of the Convolutional Layers in our DCNN architecture is followed by a Max Pooling Layer. Because RELU performs better than RMS/Tanh, ReLU is used as an activation function in each of the Convolutional Layers. There are respectively 32, 64, 128 and 256 filters employed in those respective levels, each having a 3*3 filter size. For the Max Pooling Layers, a 2*2 filter with a 1 stride is used as the filter size. The result is a Feature Map with Input Features, which is supplied as an input to the Flatten Layer. The Flatten Layer then converts the Feature Map into Nodes, and Nodes of the Flatten Layer are passed to the next step. The completely fully connected layers included dropout to add variety and promote greater use of all the nodes.

In the first convolutional layer, a kernel was added to the input picture matrix during the convolution process. ReLU activation function was then employed. The pooling layer was used after the activation layer to lower the network's processing load and picture size on the feature map. The max-pooling approach was employed. After passing through convolution layers, flattening was used, and the following stage of the process resulted in the formation of a fully linked layer. Dropout was carried out to enhance model performance and avoid network overfitting.

Comparing the performance and capacity to highlight particular better performer model, pre-trained models were also examined. In this investigation, the three various pre-trained models were used. These models include ResNet50, VGG16, and DenseNet169. The Oxford University Visual Graphics Group (VGG) created the VGG model in 2014. The main difference between the numerous VGG versions that were put into use, including VGG16 and VGG19, was the amount of convolution layers. The residual model used by ResNet50, which was created at Microsoft and described in 2015, employs shortcut connections.

Three datasets, training, validation, and test are split from the bone X-ray images. The model was first developed using the training set, and it was then examined using the images from the validation set. Testing data is used to determine how effectively the model can predict the abnormalities.

3.4. Evaluation metrics

The performance of the classifier is determined using a variety of evaluation metrics. The study's evaluation measures are described in more detail below [10]. Accuracy, Loss, Area under the ROC Curve (AUC) and kappa evaluation measures are used to determine the classifier's performance.

To calculate the Cohen's kappa coefficient, accuracy, and loss rate for both training and validation data in order to evaluate the performance of the experimental deep CNN models. The analysis results include model performance for every evaluation measure. The experimented results demonstrate that which model performs rather well when categorizing the training data. Higher k-value shows the reliability of the model's performance. The analysis of training and testing loss demonstrates that as training epochs are increased, training and testing errors are decreased.

4. Results Comparison and Discussion

CNN is a popular tool for classifying images. In this study, the classification accuracy of the typical CNN model with Adam optimizer and pre-trained models (VGG16, ResNet50, and DenseNet169 are compared. The performance of the seal CNN architectures for the classification of bone X-ray images are examined. Each model underwent the same optimization process, and the results of these models' performances were evaluated.

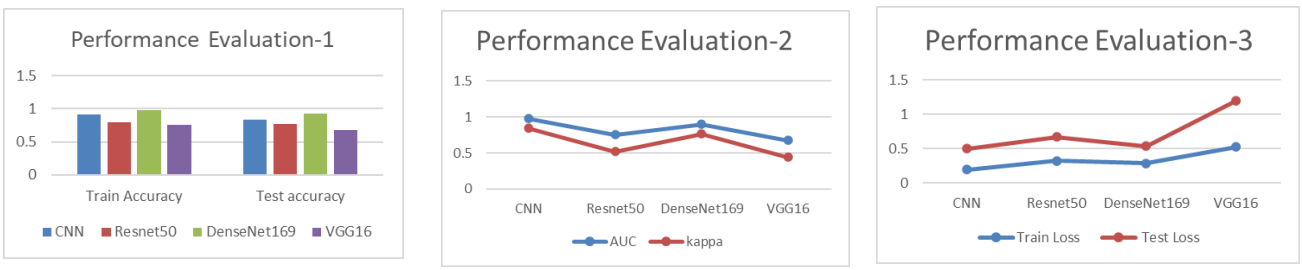

Fig. 6: (a) Comparison of Classification Performance in Accuracy (b) Comparison of Classification Performance in AUC and kappa Scores (c) Comparison of Classification Performance in Loss Rate

In this study, Deep Convolutional Neural Network (CNN) models are applied over a large size dataset known as MURA (Musculoskeletal Radiographs Abnormality) and attempted to enhance the model performance (in terms of maximization of accuracy and minimization of loss) and the use of Dropout Regularizers (to prevent overfitting of data).

Examining the classification quality of the models and selecting a model with the optimal performance is the main objective of this experiment. The model with high accuracy can detect and classify the Bone X-Ray images whether it is normal or abnormal more precisely. The standard evaluation metrics of accuracy, loss rate, Area under the ROC Curve (AUC) scores are used to test the quality of the CNN classifiers. In addition, the Cohen's kappa score is also evaluated. The evaluation scores for each measure tested in this classification is varied depending on the batch sizes, number of iterations/epoch and amount of testing images.

As we can see from Figure 6, figure 7 and figure 8, DenseNet169 has outperformed the other three optimizers in terms of accuracy and kappa scores. As with the Train & Test Losses, CNN model with Adam optimizer has the lowest percentages in comparison to the other three. In terms of Accuracy and Losses, VGG16 and ResNet50 have done poorly. With the highest training accuracy of 97.98%, test accuracy of 92.34%, train loss of 28%, and test loss of 25%, DenseNet169 has been shown to be the best classifier among the three CNN models. VGG16 with the lowest Train and Test Accuracy is the worst performer that has been demonstrated for this dataset. Hence, for this experiment, DenseNet169 model is the optimal classifier with the highest accuracy score of 97% . Its Cohen's kappa measure is 0.76 and this score is slightly lower than the kappa score of CNN model. In terms of kappa score, CNN model with Adam optimizer is the best performer.

5. Conclusion

In this study, Deep learning technique for diagnosing and classifying upper body X-ray images, which can aid medical professionals in assessing patient studies. The experiment aims to get the best model which are easily superior to the level of performance required for the classification tasks of abnormality detection in the human musculoskeletal system based on radiographs. DenseNet169 Model outperformed the experiment by achieving the greatest Training Accuracy of 97.98% with 0.76 kappa scores. Additionally, it is observed that an improving trend in model reliability as the number of training epochs increased. Based on these findings, it is found that, with more computational power, these pre-trained models can rank among the most trustworthy classifiers in the categorization of the MURA dataset. In the future, it is intended to apply an image enhancement technique for more precise image detection of these bone X-ray images.

6. Acknowledgment

My family plays the most important part in the avocation of this work. I am also thankful to the supervisor for her supportive motivation, unrestricted support and valuable and timely advice and suggestions for the completion of this work.

7. References

[1] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. 2020. Automated detection of COVID-19cases using deep neural networks with X-ray images. Computers in biology and medicine 121 (2020), 103792. 7

[2] Mehr, Goodarz. Automating Abnormality Detection in Musculoskeletal Radiographs through Deep Learning. Medicine arXiv preprint arXiv:2010.12030 (2020).

[3] Ahmad J, Saudagar AK, Malik KM, Ahmad W, Khan MB, Hasanat MH, AlTameem A, AlKhathami M, Sajjad M Disease Progression Detection via Deep Sequence Learning of Successive Radiographic Scans. International journal of environmental research and public health. (2022) 19(1):480. 2

[4] Chada, G. Machine learning models for abnormality detection in musculoskeletal radiographs. MDPI Reports

2019, 2, 26. [CrossRef] 6

[5] Mura Dataset. Available online: https://stanfordmlgroup.github.io/competitions/mura/ (accessed on 24 September 2021). 3

[6] Rajpurkar at al., MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. Medical Physics http://arxiv.org/abs/1712.06957

[7] I. D. Apostolopoulos and T. Bessiana, Covid-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural network. Physical and Engineering Sciences in Medicine, arXiv preprint arXiv: 2003.11617, 2020.

[8] Maya Varma, Mandy Lu1, Rachel Gardner, Jared Dunnmon, Nishith Khandwala, Pranav Rajpurkar, Jin Long, Christopher Beaulieu, Katie Shpanskaya, Li Fei-Fei1, Matthew P. Lungren and Bhavik N. Patel. Automated abnormality detection in lower extremity radiographs using deep learning. *Nature Machine Intelligence volume* 1, pages578–583 (2019)

[9] Rajpurkar, Pranav, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *Medical Physics, arXiv preprint* arXiv:1712.06957 (2017).

[10] Ghosh, M., Hasan, S., and Debnath, P., Ensemble Based Neural Network for the Classification of MURA Dataset, *Journal of Nature, Science & Technology*, 3(2021), 25-61.