Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)

doi:10.3233/SHTI241080

Using Deep Learning to Suggest Treatment for Proximal Humerus Fractures

Mohammadreza AZARPIRA[a,1], Ihssen BELHADJ[a] and Mohammed KHODJA[a]
[a]*Centre Hospitalier Intercommunal de Meulan-Les-Mureaux, Yvelines, France*
ORCiD ID: Mohammadreza Azarpira https://orcid.org/0000-0001-6990-2780

Abstract. Proximal humeral fractures are among the most common fractures seen in emergency departments. Accurately diagnosing and selecting the most appropriate treatment for these fractures can be challenging, and consultation with a senior orthopedic surgeon can be time-consuming for both the patient and the emergency unit. We developed a machine learning model for predicting the type of treatment based on injury radiographic images. The model distinguishes between nonoperative and operative treatment options, achieving an accuracy of 86% and an interobserver reliability (kappa) of 0.722 for test-dataset, which is more than the interobserver agreement between shoulder surgeons. This model has the potential to serve as a therapeutic decision support system for the practitioners in the emergency departments to expedite treatment decisions and to reduce patients' waiting time.

Keywords. Proximal humerus fracture, decision support system, deep learning, transfer learning, interobserver reliability

1. Introduction

Proximal humerus fractures account for approximately 5% of all fractures in adults, with a significant increase in incidence among the elderly population [1]. Several main classifications exist for this fracture, including Neer, AO, and Hertel [1]. To effectively manage a patient with a fracture, the first healthcare practitioner must accurately classify the type of fracture and be well-versed in treatments suited for different fracture patterns. Effective management of these fractures necessitates a thorough understanding of anatomy, familiarity with classification systems, and knowledge of available treatment options. Such expertise ensures that the patient receives optimal care tailored to their specific condition [1]. However, accurately classifying and selecting the most appropriate treatment can be challenging, especially for less experienced practitioners.

These fractures can be treated by two main modes: operative or non-operative [1]. The choice of treatment plan depends on age, medical history, and fracture classification. Orthopedic residents and emergency medicine specialists face a considerable learning curve when deciding the correct treatment plan based on radiologic images. However, this task is easier for experienced orthopedists. The ability to rapidly formulate an accurate treatment plan is timesaving for both patients and clinics. It can prevent over- or under-treatment, alleviate emergency department overcrowding, and enhance patient care [2].

[1] Corresponding Author: Mohammadreza AZARPIRA; E-mail: mohammadreza.azarpira@ght-yvelinesnord.fr

In our study, we aim to utilize deep learning through transfer learning for predicting the type of management (operative or non-operative) for proximal humeral fractures in anteroposterior radiographic images. We will also determine the interobserver reliability of the predictions of the model relative to specialist annotations.

2. Material and Methods

2.1. X-ray image dataset

X-ray images of proximal humerus fracture were selected from the MURA dataset [3]. MURA is a large public dataset containing 40,561 normal and abnormal x-ray images. We then reviewed each radiographic image to determine the type of treatment needed either operative or non-operative treatment. The images were saved in the appropriate directories: 'N' for non-operative treatment and 'O' for operative treatment. Several normal proximal humerus x-rays were added to 'N' group to have balanced groups and increase generalization capacity of the learnt model [4].

2.2. Deep learning libraries and transfer learning model

We used TensorFlow version 2.1 (https://www.tensorflow.org), Keras library for deep learning algorithm (https://keras.io/), Python programming language version 10 (https://python.org), Cohen-kappa, accuracy, precision, recall and f1 scores from Scikit-learn, Jupyter Notebook (https://jupyter.org). For Transfer Learning model, we used "MobileNetV2" pre-trained model. MobileNetV2 is a 53-layer deep lightweight CNN model which contains 3.5 million parameters. This model has been widely used for medical image analysis [5]. Apple M1 processor with enabled Metal device was used for the execution of the algorithms.

2.3. Pretreatment of the images

Each image was cropped by a rectangle containing the Coracoid process, Acromion, and proximal of the Humerus, showing the fracture fragments completely (Figure 1). This was to exclude extra variable parts and to include the same uniform parts as a pretreatment step of input data. We applied data augmentation techniques such as rotation, horizontal flipping, and brightness and contrast changes to enhance the dataset. The MobileNetV2 model expects the pixel values between [-1, 1]. A preprocessing method is included in the model to rescale the input images.

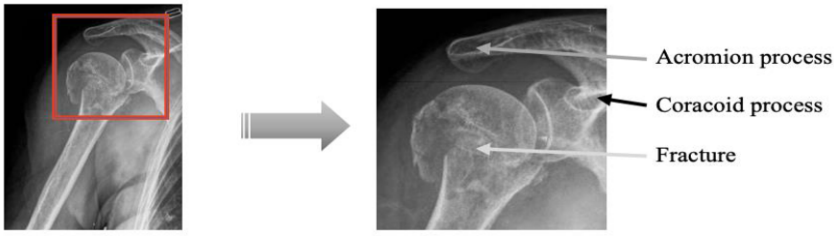

Figure 1. Left image shows a typical shoulder x-ray of proximal humerus fracture. The image includes a segment of humeral shaft, ribs, and soft tissue shadows which add significant noise to input data. Right image shows cropped image to a rectangle containing consistent anatomic landmarks (Acromion process, Coracoid process, and Fracture) to have more uniform and cleaner input data.

2.4. Training the model and evaluation of its predictions

The MobileNetV2 with frozen convolutional base was used as the base-model. A classification head was added, and layers-dropout was set to 20%. We used a learning rate of 0.0005 and a batch size of 32. The 80%/20% split of the training images was chosen to ensure a sufficient validation set while maximizing the training data. The model was used to learn from the radiographic image data containing 'N' and 'O' classes. The resulting trained model distinguishes between x-ray images that need non-operative or operative treatment. After training the model, the accuracy of the model was determined on a test dataset including images never used during training or evaluation phases. We calculated the Cohen-kappa coefficient (k) to determine the interobserver reliability between predicted labels and true labels of the test dataset. Confusion matrix and other evaluation metrics of the model including precision, recall and f1 scores were also provided. The notebook and datasets are available at this repository: https://github.com/azarpira/Project_STC2024 .

3. Results

The train dataset included 155 images in N group and 159 images in O group and, the test dataset included 36 images. The model was used to train on 314 images during 100 epochs. At the end of training, the model achieved validation accuracy of 87% with training accuracy of 90% (Figure 2A). We applied the model to the test dataset. The results showed an accuracy of 86% and a kappa of 0.722. The confusion matrix (Figure 2B) illustrates the model's performance across different classes. Additionally, the model achieved a precision of 78%, a recall of 100%, and an f1-score of 88%.

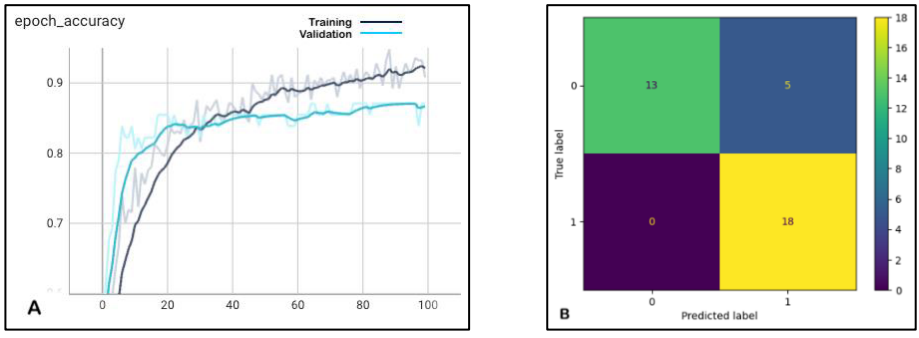

Figure 2. A: Training and Evaluation accuracy versus iterations of the learning process. B: Confusion matrix of classes, non-operative treatment (0) and operative treatment (1).

4. Discussion

Proximal humeral fractures are the third most common fractures in human beings [1]. Junior orthopedists and emergency medicine specialists face a learning curve to correctly manage these fractures without consulting a senior surgeon [6].

Supervised machine learning algorithms were used for the classification of x-ray images of proximal humeral fractures. In a study on 1891 proximal humerus x-ray images using neural networks and transfer learning the authors showed 96% accuracy in detecting proximal humerus fractures from non-fractured shoulders. Their model achieved 65-86% accuracy in distinguishing the type of proximal humerus fracture. To produce clean input data (data pretreatment), the authors manually cropped the proximal humerus x-ray images in a rectangle containing the head and neck. They used a data augmentation method to increase the input data from 1891 images to more than 40,000 images [7].

In another study, using the MURA database, the authors trained a 169-layer DenseNet baseline model to detect and localize abnormality in x-rays in a supervised methodology [4]. The x-rays in the train dataset were manually labeled by radiologists as either normal or abnormal. For the test dataset, 207 x-rays, annotated by board-certified radiologists, were used as gold-standard. They calculated Cohen-kappa score to find the agreement of the model output and the radiologist annotations with the gold-standard. Their model performance was comparable to best radiologist performance for detection abnormality in finger and wrist x-rays, however, it was worse than the best radiologist performance in detecting abnormality in hand, forearm, elbow, humerus, and shoulder x-rays [3].

In another study, convolutional neural networks (CNN) were used to classify fractures of the proximal humerus, as well as the humeral shaft, clavicle, and scapula, from a dataset containing 6,172 radiographic examinations [8]. Their input data included different radiographic projections without any selection of specific areas of the radiographs. The authors reported an overall good to excellent area under the curve (AUC) score for different fracture classifications [8].

Shoulder x-rays almost always include variable parts of nearby structures including lung, ribs, cervical spine, arm, monitoring electrodes, etc. These extra parts add considerable noise and may damper the efficiency of a training model. We cropped the x-ray images to reduce the noise and achieve cleaner, more uniform input data [9].

We used Transfer Learning. With this method, one can retain the architecture of a high-performing and large model and retrain it for our specific classification problem (for example: our specific images of proximal humerus fractures) to obtain an optimal model with a limited amount of data [9,10]. We used the MobileNetV2 pre-trained model from Tensor Hub [5]. This model, produced by Google, is trained on 1.8 million images across 1000 classes of web images.

There is moderate interobserver reliability between shoulder surgeons for the choice of treatment of these fractures [11]. In a study using 40 x-rays of proximal humerus fracture, the interobserver reliability (kappa) for choosing between “operative or nonoperative” options, was between 0.528 and 0.578. More experienced surgeons had more interobserver agreement (kappa) score [11]. Our classification model has a kappa score of 0.722 which is superior to interobserver agreement between shoulder surgeons.

Previous studies have demonstrated the potential of machine learning models in predicting hospital admissions and aiding decision-making in emergency departments. For instance, Feretzakis et al. (2022) developed a machine learning model to predict hospitalization outcomes for emergency department patients, highlighting the value of such models in improving clinical workflows and patient outcomes [12]. From the practical standpoint, our model can be helpful in emergency units by decreasing the time of therapeutic decision process [2].

This study has limitations, one limitation is exclusion of patient-specific factors such as age and medical history, which are crucial in clinical decision-making. Future work should incorporate these variables to enhance model accuracy and applicability. Another limitation is the limited number of x-ray images which were used for this study. Training a model on a larger number of images would lead to a more accurate and reliable model with better generalization capacity.

5. Conclusions

We developed a deep learning model that predicts proximal humerus fracture treatment plans from injury x-ray images. Our model achieved an accuracy of 86% and a kappa score of 0.722, demonstrating its potential to assist in clinical decision-making. Future research should focus on integrating this model into hospital information systems and validating its performance on larger, more diverse datasets.

References

[1] Baker HP, Gutbrod J, Strelzow JA, Maassen NH, Shi L. Management of Proximal Humerus Fractures in Adults-A Scoping Review. J Clin Med. 18 Oct 2022;11(20):6140. doi: 10.3390/jcm11206140.
[2] Sartini M, Carbone A, Demartini A, Giribone L, Oliva M, Spagnolo AM, et al. Overcrowding in Emergency Department: Causes, Consequences, and Solutions—A Narrative Review. Healthcare (Basel). 25 Aug 2022;10(9):1625. doi: 10.3390/healthcare10091625.
[3] Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, et al. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs [Internet]. arXiv; 2018 [cited 20 May 2024]. Available from: http://arxiv.org/abs/1712.06957
[4] Google for Developers [Internet]. [cited 28 Sept 2024]. Datasets: Imbalanced datasets | Machine Learning. Available from: https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets
[5] Shamrat FJM, Azam S, Karim A, Ahmed K, Bui FM, De Boer F. High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images. Comput Biol Med. 2023 Mar;155:106646. doi: 10.1016/j.compbiomed.2023.106646.
[6] Hao KA, Patch DA, Reed LA, Spitler CA, Horneff JG, Ahn J, et al. Factors influencing surgical management of proximal humerus fractures: do shoulder and trauma surgeons differ? J Shoulder Elbow Surg. June 2022;31(6):e259-69. doi: 10.1016/j.jse.2021.11.016
[7] Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. Aug 2018;89(4):468-73. doi: 10.1080/17453674.2018.1453714.
[8] Magnéli M, Ling P, Gislén J, Fagrell J, Demir Y, Arverud ED, et al. Deep learning classification of shoulder fractures on plain radiographs of the humerus, scapula and clavicle. PLOS ONE. 30 Aug 2023;18(8): e0289808. doi: 10.1371/journal.pone.0289808
[9] Shah RM, Wong C, Arpey NC, Patel AA, Divi SN. A Surgeon's Guide to Understanding Artificial Intelligence and Machine Learning Studies in Orthopaedic Surgery. Curr Rev Musculoskelet Med. 1 Apr 2022;15(2):121-32. doi: 10.1007/s12178-022-09738-7
[10] Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. BMC Med Imaging. 13 Apr 2022;22(1):69. doi: 10.1186/s12880-022-00793-7
[11] Gracitelli MEC, Dotta TAG, Assunção JH, Malavolta EA, Andrade-Silva FB, Kojima KE, et al. Intraobserver and interobserver agreement in the classification and treatment of proximal humeral fractures. Journal of Shoulder and Elbow Surgery. 1 June 2017;26(6):1097-102. doi: 10.1016/j.jse.2016.11.047
[12] Feretzakis G, Sakagianni A, Loupelis E, Kalles D, et al. Prediction of Hospitalization Using Machine Learning for Emergency Department Patients. Stud Health Technol Inform. 2022 May 25; 294:145-146. doi: 10.3233/SHTI220422.