Exercise 3 Report:
Representation Learning for Bone Fractures

Linoy Elimelech (ID. 319122610), Ron Azuelos (ID. 207059114)

AI for Healthcare course with Dr. Ayelet Akselrod-Ballin
Reichman University, 2023

1 Introduction

The goal of our project is to provide an overview of the last part of the course material, in which we discussed the state-of-the-art (SOTA) of artificial intelligence (AI) for the healthcare domain.

In this report, we implement and compare two approaches of bone fracture X-Ray classifiers utilizing a using semi-supervised learning (SSL) approach. You may view our code in the following Google Colab notebook: https://colab.research.google.com/drive/1jR-wp1r-yAo2saAEC3CRhj_HQTNWa2Jl

1.1 The Problem

Bone X-Ray, also known as radiography, is a medical imaging technique that uses X-Ray radiation to capture images of bones, aiding in the diagnosis and evaluation of fractures, infections, tumors, and other bone-related conditions. The task of bone fracture X-Ray classification is important to aid in the medical diagnosis and treatment planning for patients.

A successful model that can distinguish normal and abnormal bones, can help radiologists by providing a preliminary assessment or a second opinion assessment. This potentially will reduce human error, and save time and resources for medical professionals, which can lead to a more effective healthcare treatment.

In our assignment, we suggest a potential model as described above, that can quickly analyze a large number of images. As requested, we first implemented a naïve baseline based on the research of Rajpurkar et al. [3] We then compared the result to a representation learning approach based on SimCLR, suggested by Chen et al. [1]

1.2 Tools

The code was written in Python using the Google Colab Pro platform.

2 The Data

We were instructed to base our implementation on the MURA dataset, presented in our baseline's paper [3], and use it to train and evaluate our model. MURA is a large dataset of musculoskeletal radio-graphs containing 40,561 images from 14,863 studies from 12,173 patients, where each study is manually labeled by board-certified radiologists from Stanford Hospital as either normal or abnormal. Each study belongs to one of seven standard upper extremity radio-graphic study types: elbow, finger, forearm, hand, humerus, shoulder, and wrist. In our project, we focused on three specific types of bones — elbow, hand and shoulder – as was done in the baseline paper we based on. MURA dataset contains only training and validation data. To overcome the lack of a test set, we split the training set into two sets, where 80% of the data were dedicated to the actual training process, and the remaining 20% were dedicated to the testing process of our model. Table 1 shows the final amount of images per bone type, after splitting the original train data into test and train sets.

	Train data	Validation data	Test data
Hand	4436	460	1107
Elbow	3935	465	996
Shoulder	6711	563	1668
Total	**15082**	**1488**	**3771**

Table 1: Amount of images in each dataset

2.1 The Images

As mentioned earlier, we chose to train and evaluate the model using only three bone types: elbow, hand, and shoulder. Figures 1 and 2 present examples of both normal and abnormal X-Ray scans of the above-mentioned types of bones, used to evaluate our models.

Since the provided images do not have the same resolution (or pixel size), we pre-processed the images by resizing them to a unified size of 224x224 pixels, converting them to grayscale (from RGB color), normalizing their pixel values, and transforming them to Tensors. Lastly, we provided the model with the preprocessed images and their corresponding labels.

Figure 1: Examples of normal hand, elbow, and shoulder X-Rays

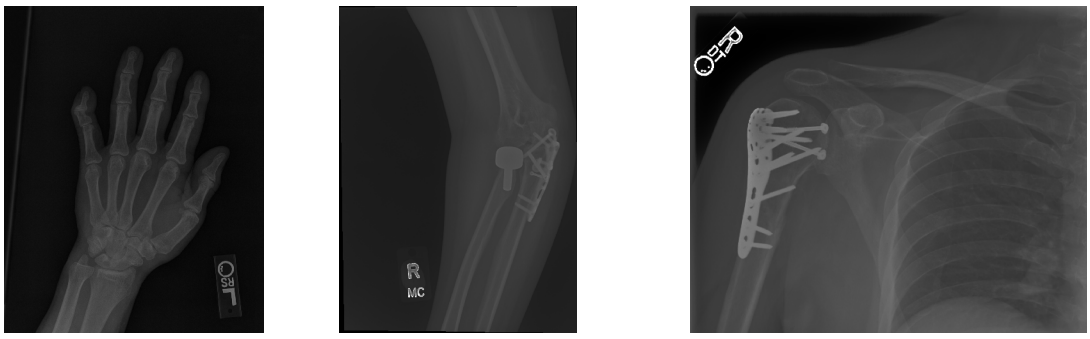

Figure 2: Examples of abnormal hand, elbow, and shoulder X-Rays

2.2 Label Formatting

The original labels are provided in a comma-separated values (*.csv*) file containing a record for each study, which describes the path to the folder where the X-Ray images are located, and the label for the entire folder. This means that all images in a study have the same label. The label $y \in 0, 1$ represents the bone status, when *0* is normal and *1* is abnormal. Since a study may contain multiple X-Ray images, we created a new file containing a record for all images, where the label derived from the original study record's label. We performed the same process for all datasets and eventually created separate train, validation, and test sets for each type of bone.

2.3 Data Distribution

Figures 4, 3 and 5 present the data distribution in each dataset, describing the amount of normal and abnormal bones per type of bone. As we can notice, across all parts of the data, the shoulder data is almost equally distributed between normal and abnormal samples. Opposing to the shoulder data, the hand data have a significant imbalance, where the abnormal data is only 25-40% of the data in each set. The elbow data have only a moderate imbalance between the two classifications. There different types of distributions between the classes may have different impacts on the performance of the trained models.

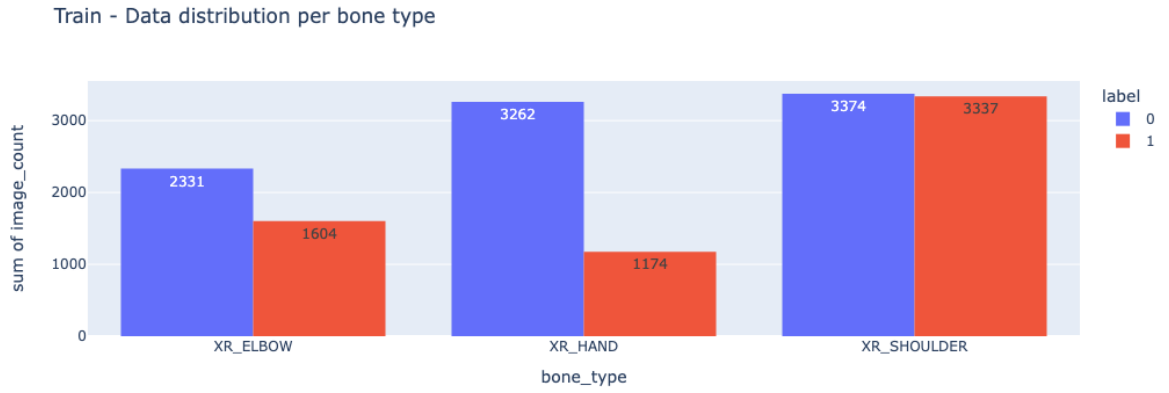

Figure 3: Train - Data distribution per bone type

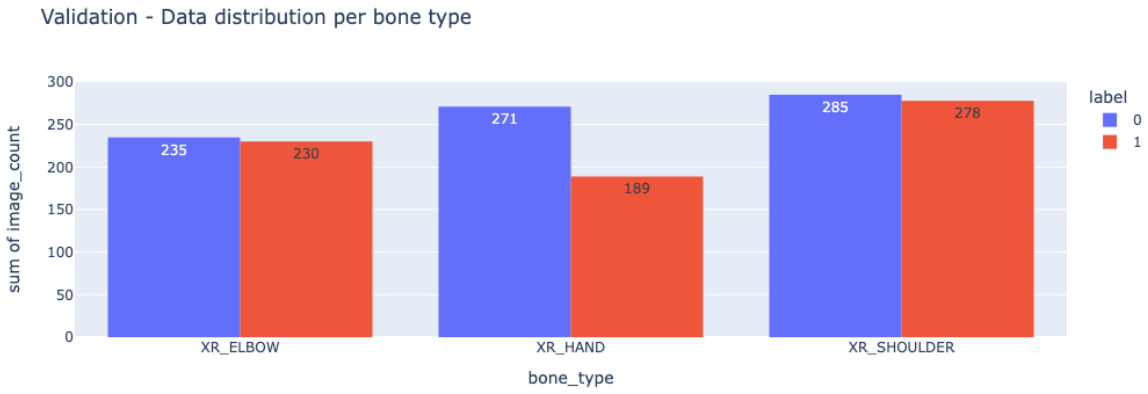

Figure 4: Validation - Data distribution per bone type

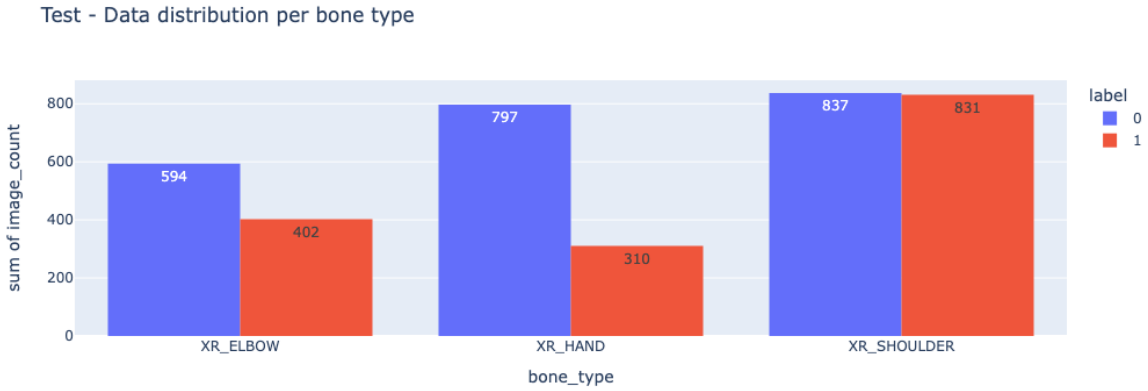

Figure 5: Test - Data distribution per bone type

3 The baseline

In the first part of the assignment, we were requested to implement a baseline. A baseline model provides a benchmark against which the performance of the implemented solution can be compared. It helps to determine if the proposed approach provides an improvement over existing methods, similar to other methods, or even worse than them.

Based on the baseline's paper, we initially tried to train our model using the entire dataset, containing the three types of bones combined. This approach produced unsatisfactory results. We assume that the poor results were caused by the labeling method, which provided the same label value for different types of bones. This means that the model cannot distinguish correctly between abnormal X-Rays of two different types of bones (e.g. an abnormal hand X-Ray and an abnormal shoulder X-Ray both have the same label value of *1*).

Therefore, we decided to train a separate model for each bone type to increase the performance and get better results. To implement the baseline we use DenseNet-169, as was used in the baseline's paper.

3.1 The DenseNet-169 Model

DenseNet-169 is a convolution neural network (CNN) architecture that was proposed in the DenseNet paper by Huang et al. [2] It is specifically designed for image classification tasks. DenseNet-169 consists of multiple dense blocks. Each dense block contains a series of dense layers, which are composed of multiple convolution layers followed by a batch normalization layer and a non-linear activation function (ReLU).

In our baseline, we used a DenseNet variation with 169 layers to predict the probability of abnormality for each image in a study. The network uses above mentioned Dense CNN architecture, which connects each layer to every other layer in a feed-forward fashion to make the optimization of deep networks tractable. We replaced the final fully-connected layer with one that has a single output, after which we applied a sigmoid non-linearity. For each image of each study in the training set, we optimized the weights using binary cross-entropy loss with logits.

3.2 Experimental Results

To evaluate the performance of our suggested method, we created a baseline for each type of bone and trained them with the relevant images. Each model was trained for 10 epochs, with *Adam* optimizer and a learning rate of *0.00001*. Figures 6, 7 and 8 present the accuracy and loss results of each epoch during the learning process of the baseline models. The shoulder model presents the most significant learning, while the loss improvement is similar between the shoulder and elbow models.

Figure 9 presents the final results of the evaluation process for each model, evaluated using the relevant test sets. Overall, all of the models produced fairly good results. Surprisingly, although supplied the worst results during the training process over the validation set, the hand model achieved the highest accuracy on the test set.

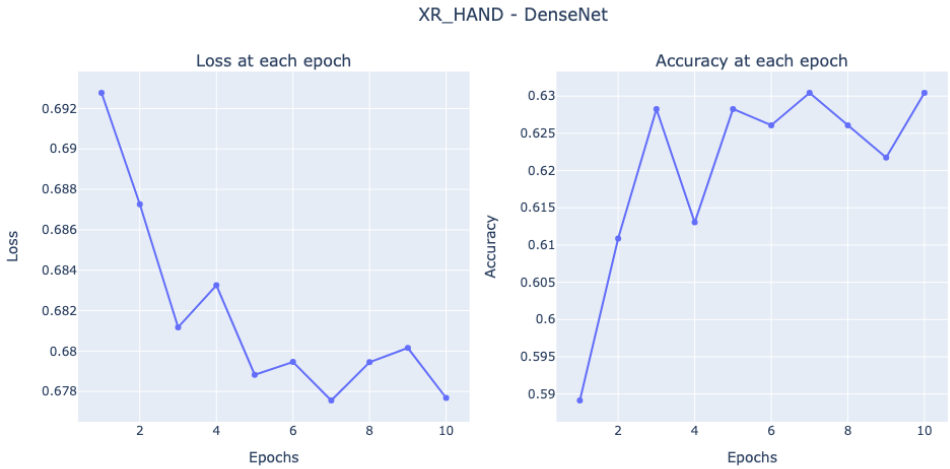

Figure 6: Baseline: Loss & Accuracy for hand model

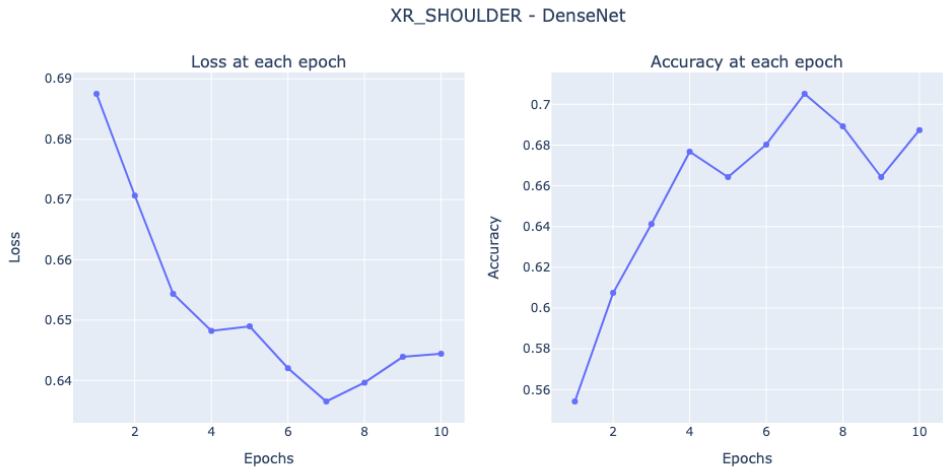

Figure 7: Baseline: Loss & Accuracy for shoulder model

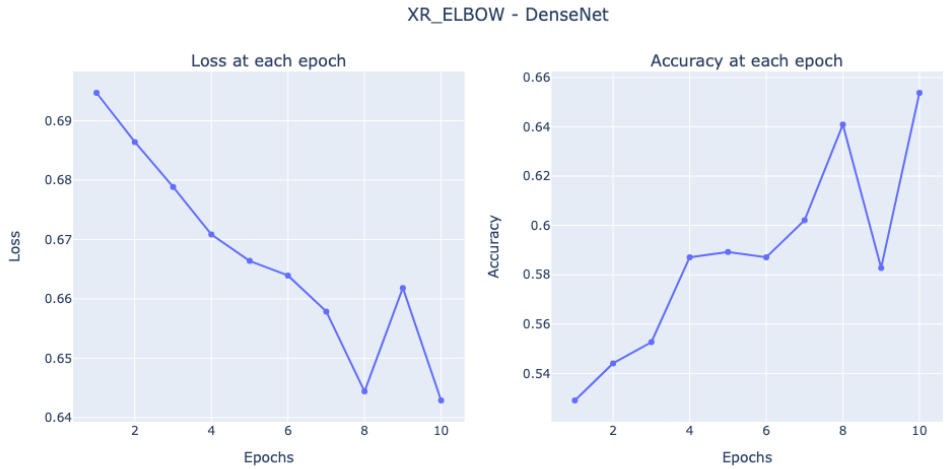

Figure 8: Baseline: Loss & Accuracy for elbow model

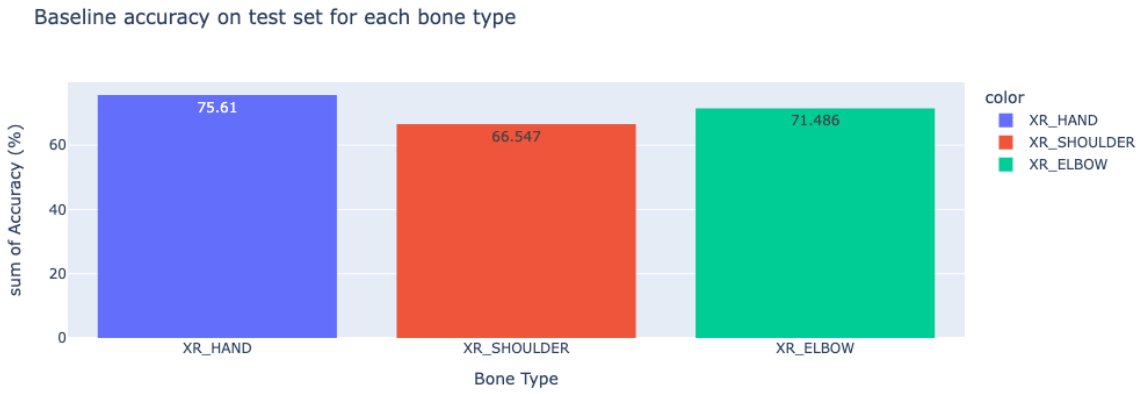

Figure 9: Baseline: Final accuracy

4 Representation Learning

4.1 The SimCLR Model

SimCLR is a deep learning model designed for self-supervised representation learning and is used for downstream tasks like image classification, object detection, and image retrieval. The main principle of SimCLR is contrastive learning, a technique that aims to learn representations by contrasting similar and dissimilar samples. SimCLR relies on data augmentation to create diverse views of the same image. Each input image is transformed using various data augmentation techniques. The core idea is to maximize agreement between positive pairs (augmented views of the same image) and minimize agreement between negative pairs (views from different images). This is done using a contrastive loss function. SimCLR incorporates a projection head on top of the CNN backbone. The projection head takes the high-dimensional output features from the backbone and maps them to a lower-dimensional representation space, known as the "projection space." The projection head enhances the discriminative power of the learned representations. During training, SimCLR randomly samples pairs of augmented views from the data set and feeds them through the CNN backbone and projection head. The contrastive loss is computed for each pair, and the model parameters are updated using back-propagation and gradient descent optimization.

Due to high GPU memory consumption, our implementation is based on ResNet-18, which is a smaller variant of ResNet-50 used in the referenced paper. The model was fed with a batch size of 32 images. We also used a binary cross entropy with a logit loss function, the same one we used in the baseline model.

4.2 Experimental Results

To enable a valid comparison, we chose the same approach of building a separate model for each bone type in this section. In addition, we performed three different experiments per bone type as requested, and compared the results when using 1%, 10%, and 100% of the training set data while training. Eventually, we tested nine different models, where each one was trained for 10 epochs with a learning rate of *0.00001*.

4.2.1 Hand

Figures 10, 11 and 12 present the accuracy and loss results of each epoch during the learning process of SimCLR models for the hand bone images. As expected, the model that used 1% of the data or training reached the lowest final accuracy. It is noticeable that the model that used 10% of the data didn't improve its accuracy at all, but it did reach the most significant improvement regarding the loss. To evaluate these models we used the test set and the results are in Section 4.2.4.

4.2.2 Shoulder

Figures 13, 14 and 15 present the accuracy and loss results of each epoch during the learning process of SimCLR models for the shoulder bone images. In that case, the 1% model presents poor results with no improvement neither for the accuracy nor for the loss. The 10% model, on the other hand, presents a descent improvement in both

accuracy and loss. The 100% model presents a learning pattern that is very similar to the 10% model, achieving much better results with an accuracy of almost 70% at the final epochs, and converging faster than the previous model. To results of the evaluation process for these models are also in Section 4.2.4.

4.2.3 Elbow

Figures 16, 17 and 18 present the accuracy and loss results of each epoch during the learning process of SimCLR models for the elbow bone images. The 1% model and the 10% do not present a stable learning process, with sharp declines in the accuracy graphs. Surprisingly, the 1% shows an improvement regarding the loss, unlike the 10% model. Regarding the 100% model, the learning process is similar to the 100% model of the shoulder, with an improvement in both measures. The test results for these models are presented in Section 4.2.4 as well.

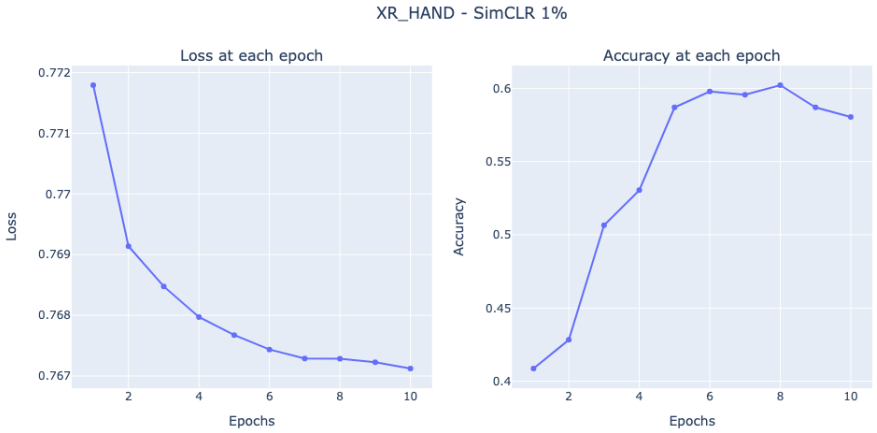

Figure 10: SimCLR: Loss & Accuracy for hand model 1%

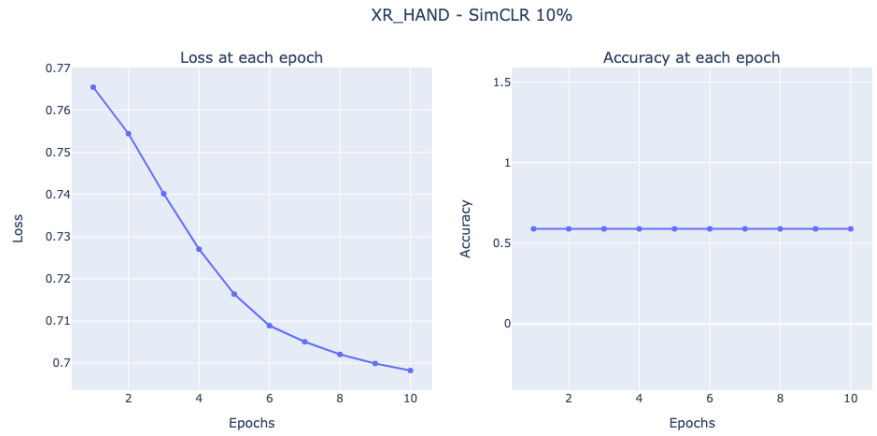

Figure 11: SimCLR: Loss & Accuracy for hand model 10%

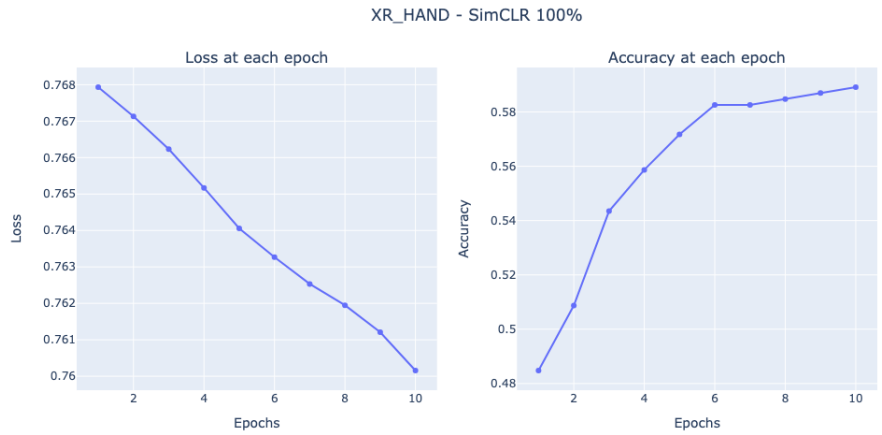

Figure 12: SimCLR: Loss & Accuracy for hand model 100%

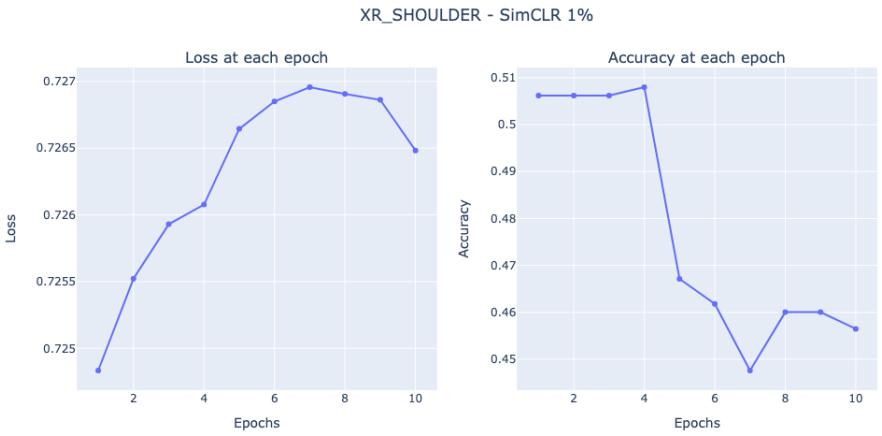

Figure 13: SimCLR: Loss & Accuracy for shoulder model 1%

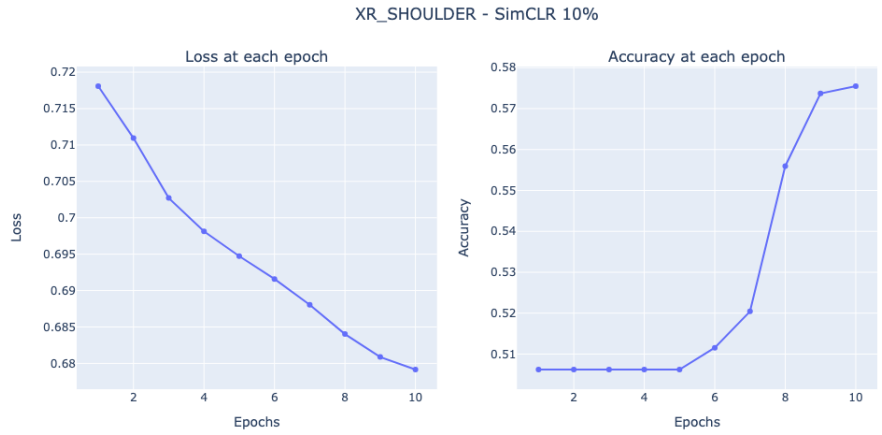

Figure 14: SimCLR: Loss & Accuracy for shoulder model 10%

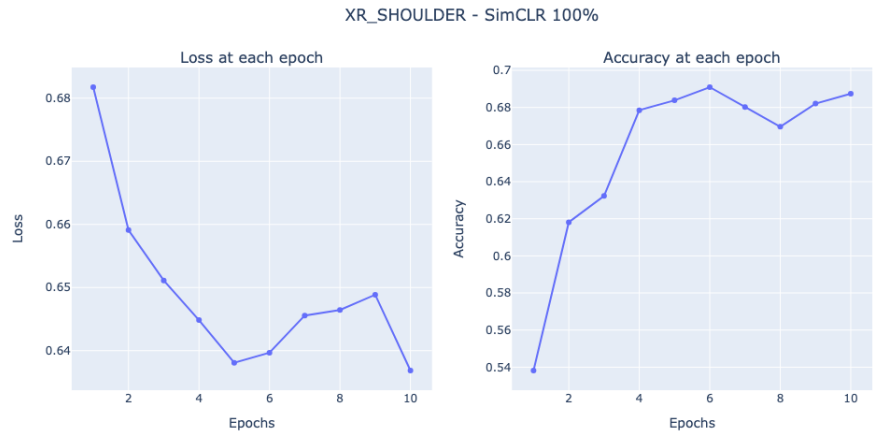

Figure 15: SimCLR: Loss & Accuracy for shoulder model 100%

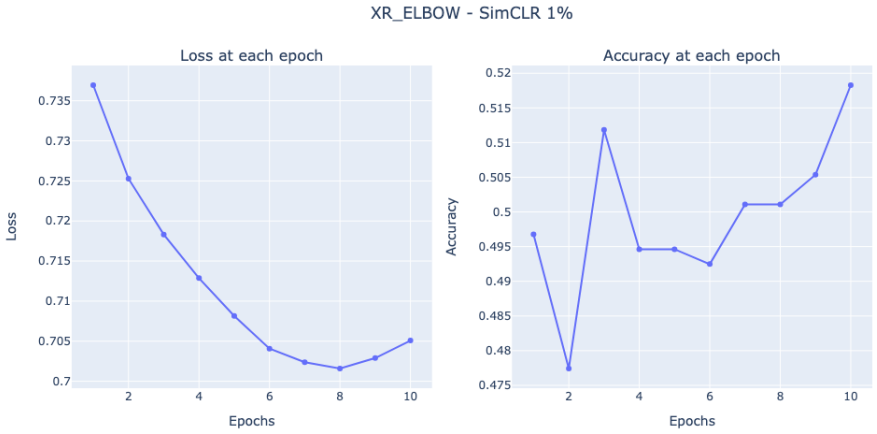

Figure 16: SimCLR: Loss & Accuracy for elbow model 1%

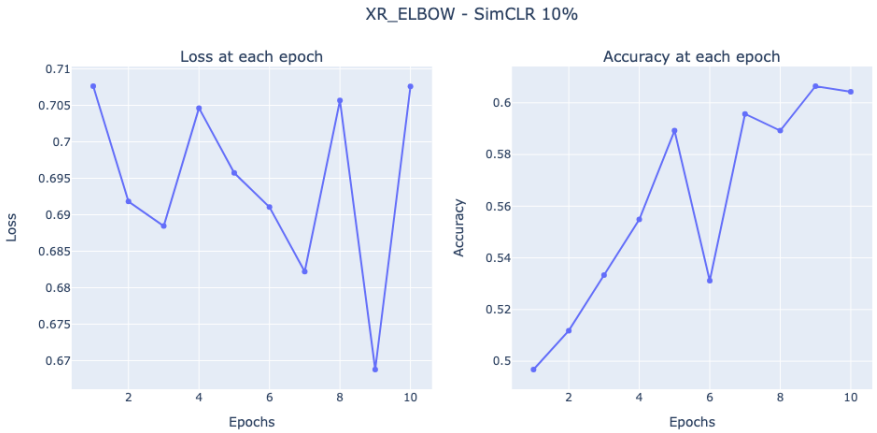

Figure 17: SimCLR: Loss & Accuracy for elbow model 10%

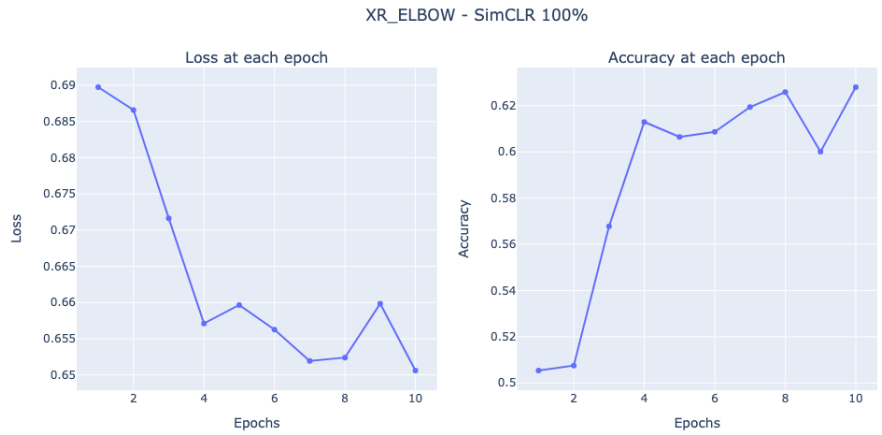

Figure 18: SimCLR: Loss & Accuracy for elbow model 100%

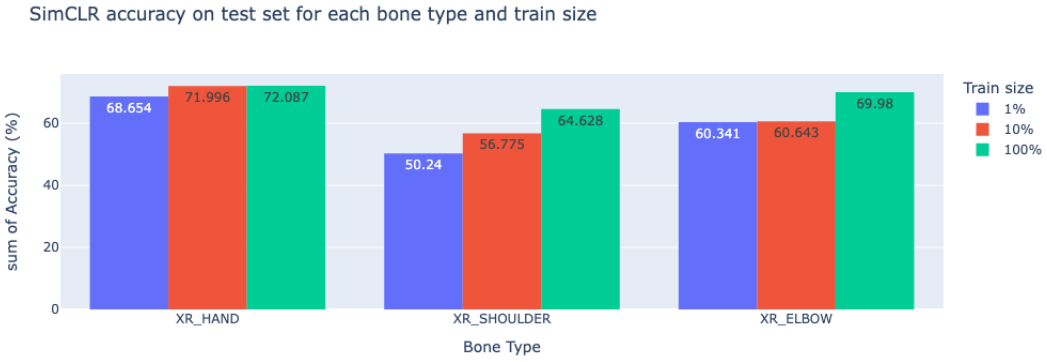

Figure 19: SimCLR - Accuracy comparison

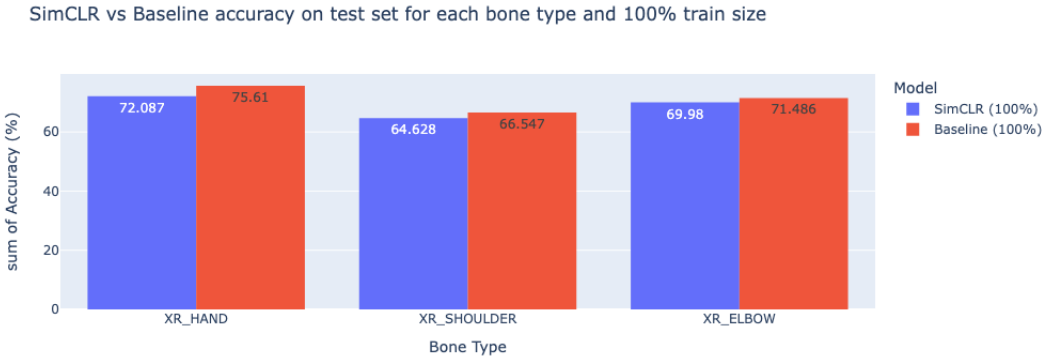

Figure 20: Baseline & SimCLR comparison

4.2.4 Comparison

In Figure 19 we present a comparison between the evaluation process results for all nine models. It is noticeable that the hand models achieved the best results, even for the 1% models.

Based on the assumption that the more data used in training, the better results achieved, the elbow model presents the most reasonable results, with a small difference between the 1% and the 10% models, and a significant improvement in the 100% model.

The shoulder model also presents a similar pattern with a higher difference between the 1% and the 10% models. It also has the most significant difference between the models, achieving a very low accuracy for the 1% model, as expected.

5 Conclusions and discussion

In the previous Section, we saw that the hand models reached the best results compared to the other bone types. Looking at the data distribution of the training set and the test set (described in Section 2.3), we see that the train data contains almost three-times more images of normal bones than abnormal ones, with a similar case for the test set. Although that ratio may reflect a real-world situation (we tend to believe that there are more normal X-Ray images than abnormal ones), we can assume that the models were slightly overfitted or biased, which can explain the results differences between the bones.

Figure 20 presents a comparison between the final accuracy of the baseline models and the 100% SimCLR models. We can observe that the SimCLR almost achieved the same results as the baseline, with a slightly better accuracy for the baseline. Considering the fact that the baseline is based on a supervised learning approach, and the SimCLR is based on unsupervised learning, we can be satisfied by the SimCLR results that present almost the same performance without using the labels.

References

[1] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).

[2] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4700–4708.

[3] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R. L., et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957* (2017).