RiskRay: detecting hips at risk of fracture using machine learning

Motivation

Millions of people suffer from hip fractures worldwide every year. They are a hallmark fracture of osteoporosis, with 70-90% of fracture cases stemming from this disease. According to a 1993 study, the first year of costs for a patient with a hip fracture in the US is $26,000 USD[1].

The current gold standard for detecting if a hip is at risk of fracture is using Dual Energy X-ray Absorptiometry, “DEXA”. The procedure involves using two different energies of x-rays, and without describing the entire process, produces a measurement of bone density. This measurement, however, is known to have significant variation (as much as 30%) across its many manufacturers and instantiations. Access to these expensive machines is limited, and may have patients waiting weeks or months for their appointment, which cost around $106 CAD per test, in 2006.

With all of this in consideration, and the fact that this technique was developed and deployed in the 1960’s, we believe there is an opportunity for this technology to be superseded. Our machine learning architecture presents an approach that is not only more accurate, but would be significantly faster and cheaper, with technicians/doctors executing the utility in tandem with a patient’s hip x-ray.

Architecture

Neural Network Model

Supervised machine learning (ML) is a style of ML where data is labeled by a human (or other means) from a limited set of labels. The ML model is then tuned or “trained” to identify the label on new data previously unseen. Ultimately, it is a form of pattern recognition, and part of our investigation was to answer the question: “Is there a visual pattern contained in x-ray images of at-risk hips vs. healthy hips that could be picked up by a ML model?” When optimized, ML image recognition can notice the subtlest of signatures and could be able to pick up on details that evade even a professional’s eye.

Getting a labeled dataset is often the road block in a supervised learning pipeline. We introduce a novel key insight for building up this dataset. If a hip fracture occurs on one side of the body, then the opposite hip can be marked as an at-risk hip (conditional on hip health being at least in some part responsible for the fracture). These at-risk hips that are the counterparts of fractured hips can form the at-risk portion of the dataset. We then complete the dataset with images of x-rays of healthy hips.

With this method, we compiled a dataset of 113 at-risk hip x-rays and 360 healthy, “control”, hip x-rays, for 473 images total which were divided as 378 Training images, and 95 Test images (30 at-risk, 65 healthy/control). Indeed, this approaches the lower bound of an acceptable dataset size in this context, but is enough to be worth the experiment.

We used a Mixed Data neural network (MDNN) architecture that had two main components, followed by a concatenating layer. The first main block was a regular image processing network of two Convolutional Neural Networks (CNNs) and a Multilayer Perceptron (MLP) layer. The second was a small MLP applied to image metadata including the patient age and a small number of x-ray fields: kVP, distance-to-detector, and exposure time. The output from these two main channels were concatenated in a layer that passed through another MLP before outputting the final logistic result in the range [0, 1] (at-risk or control). In future iterations of this architecture, adding more fields to the second block of metadata fields would be interesting. For example, the answer to “Is there a history of osteoporosis in your family?” Finally, a 3rd party open source software package Ray Tune was used to systematically optimize hyperparameters across various runs.

Final RiskRay Product

The final RiskRay product will consist of 3 main components that function to intake a hip x-ray/DICOM, and output a normalized category of risk, a “Risk Rating” or “Risk Score”, for the user. The components are:

1. An X-ray/DICOM cropping software utility that technicians can use to click & drag the hip region in a patient’s x-ray. This cropped portion is input into the MDNN for the next phase.
2. The MDNN intakes the hip image and outputs a its classification in the range [0, 1], where 0 is towards healthy hips and 1 towards at-risk hips.
3. The classification is further categorized using the mapping in Table 1 to output a normalized Risk Rating.

The network outputting in the range [0, 1] is quite natural, really, as even humans often only have a partial degree of certainty in their estimations. Note that the Risk Rating categorization matrix in Table 1 is a suggestion and is subject to change.

Output	Risk Rating
[0, 0.4]	Healthy
[0.4, 0.6]	Moderate
[0.6, 1.0]	At risk

Table 1. MDNN Output to Risk Rating

Results & Discussion

Performance

The Ray Tune library was used to run many different sets of hyperparameters to determine the model achieving the highest accuracy on the test dataset (Fig. 1). We then evaluated the selected model against the following standard ML evaluation metrics: mean accuracy, precision, recall, and f1-score. These simple metrics can be summarized quickly. Mean accuracy is the percentage of labels the model correctly guessed on images unseen during training. As is standard practice, when the MDNN output is below 0.5, we classify its estimation as control, if the output is 0.5 or greater,

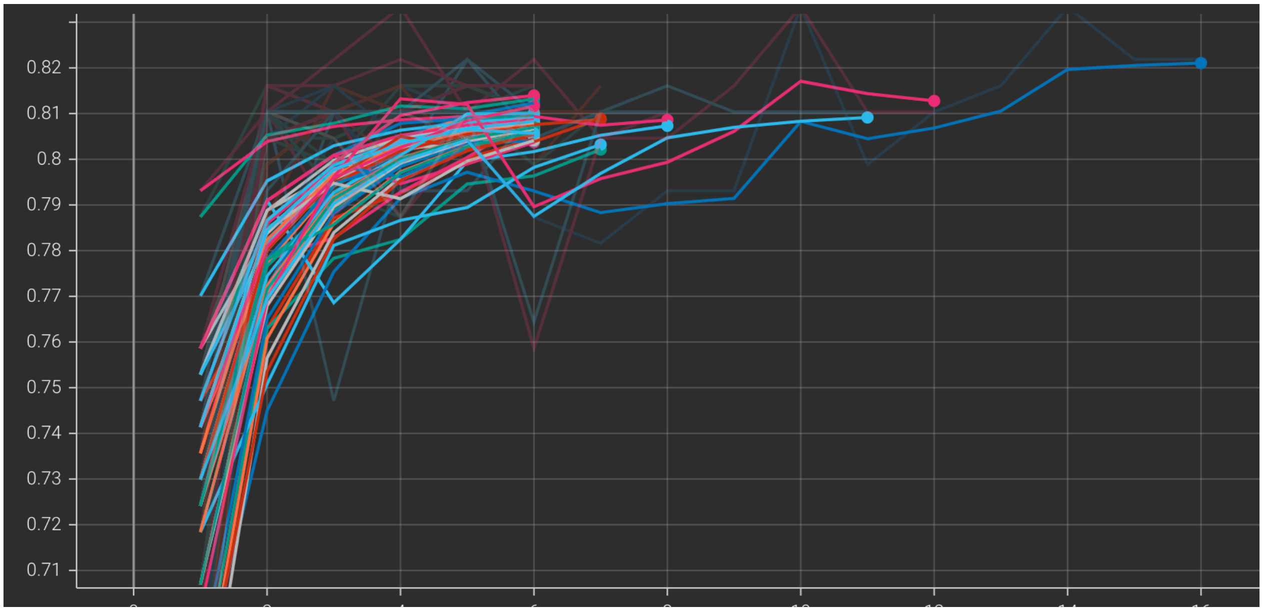

Fig 1. Mean Test set Accuracy across training time. Each line represents a different Ray Tune session.

we classify the estimation as at-risk. Precision and recall look at this labelling in a more useful way. Where we let a “group” be all the at-risk images or all the control images, and **TP, FP, TN,** and **FN** be the number of true-positives, false-positives, true-negatives, and false-negatives. Then, precision is the accuracy in a given group: **TP / [TP + FP].** Recall is how many of a given group were correctly labelled: **TP / [TP + FN].** Finally, f1-score is the “harmonic mean” and accumulates both precision and recall: **f1 = 2 x [P x R] / [P + R]**.

As indicated in Figure 1, our chosen model achieved a mean accuracy of 82% on the test dataset. This means, on the 95 test images (those it had not seen during training), it labeled “at-risk” or “control” correctly on 82% of them. Evaluation of this selected model across the remaining performance metrics is shown in Table 1. “Support” is the number of images of that type in the Test dataset. “Macro Average” is the flat average of the metric values, and “Weighted Average” weights the metrics by their Support count.

From Table 2, we can notice a slight bias in the model towards evaluating hips as control, most noticeably in Recall, and significantly less so in Precision. These numbers, along with the 82% mean accuracy strongly suggest the neural network is indeed picking up on some features in these images and metadata.

Table 2. Performance metrics on selected model.

Label	Precision	Recall	f1-score	Support
Control	0.84	0.91	0.87	65
At-risk	0.76	0.63	0.69	30
Macro Avg	0.80	0.77	0.78	95
Weighted Avg	0.82	0.82	0.82	95

Sample Inspection

In the images below, Figures 2 & 3, two batches of images are taken from the Test set for a visual inspection on network output. For the given input images, the model output percentages are overlain in red text along with the category of image: “Ctrl” or “Risk”. Most of the output samples are ideal, approaching the “Healthy” Risk Score for the control batch, and the “At Risk” Risk Score for the at-risk batch. However, exceptions can be found in either batch. Of course, this is quite likely in part because of the limited number of images to train our model. However, it is important to remember the inherent, and limiting, chaos of the data—at least in this pilot context. From the data gathered for this study, there is contamination across both control and at-risk groups because, for example, a hip might be *in reality* at risk, but still be in the control set because the patient hasn’t fractured their hip yet. Conversely, a control hip could be in the at-risk group if the patient’s fracture was due to external circumstances and not hip-health per se.

References

[1] Sernbo and Johnell., 1993. Osteoporosis International 3, 148-53

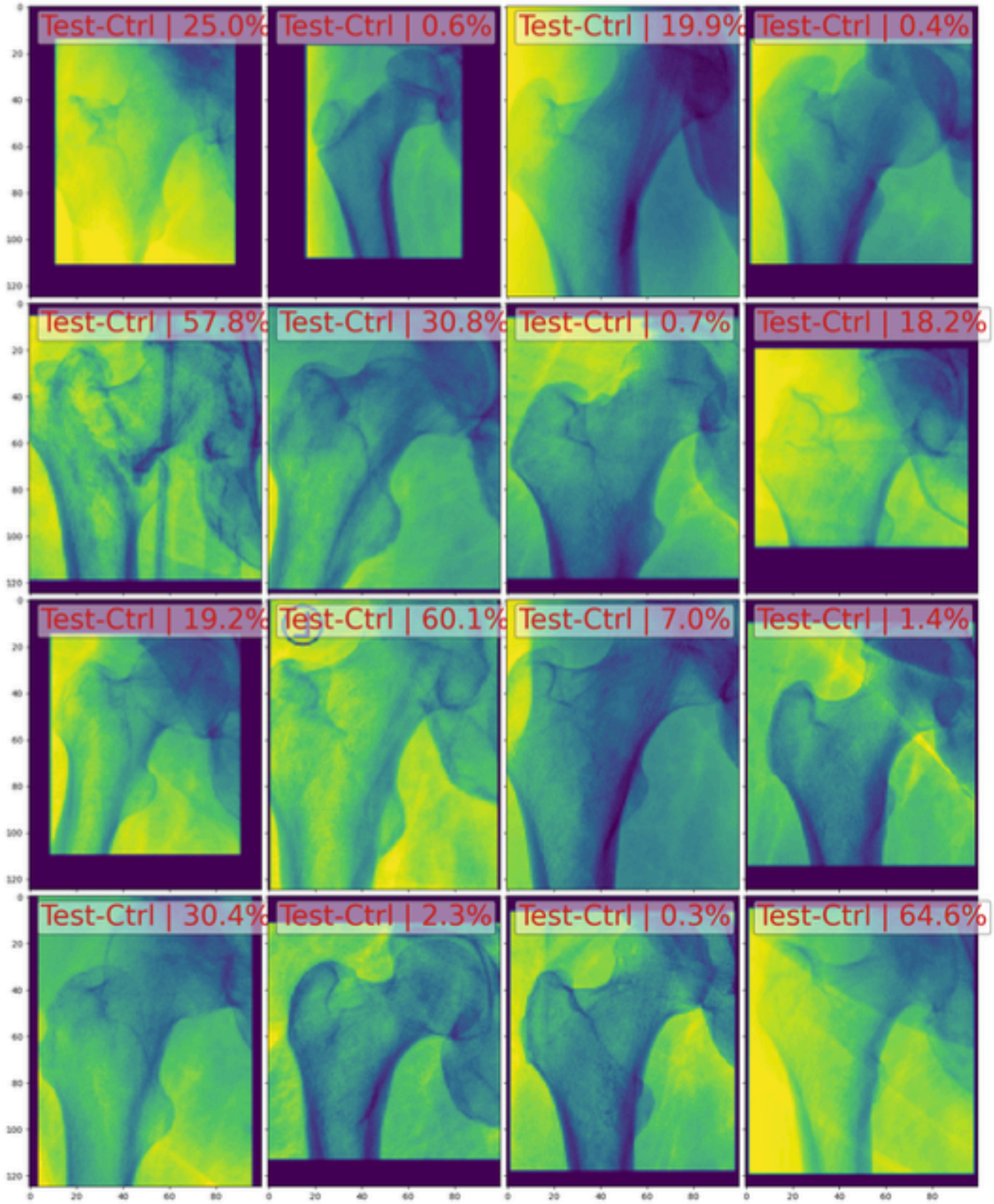

Fig. 2. Test-set control images and the selected model estimations.

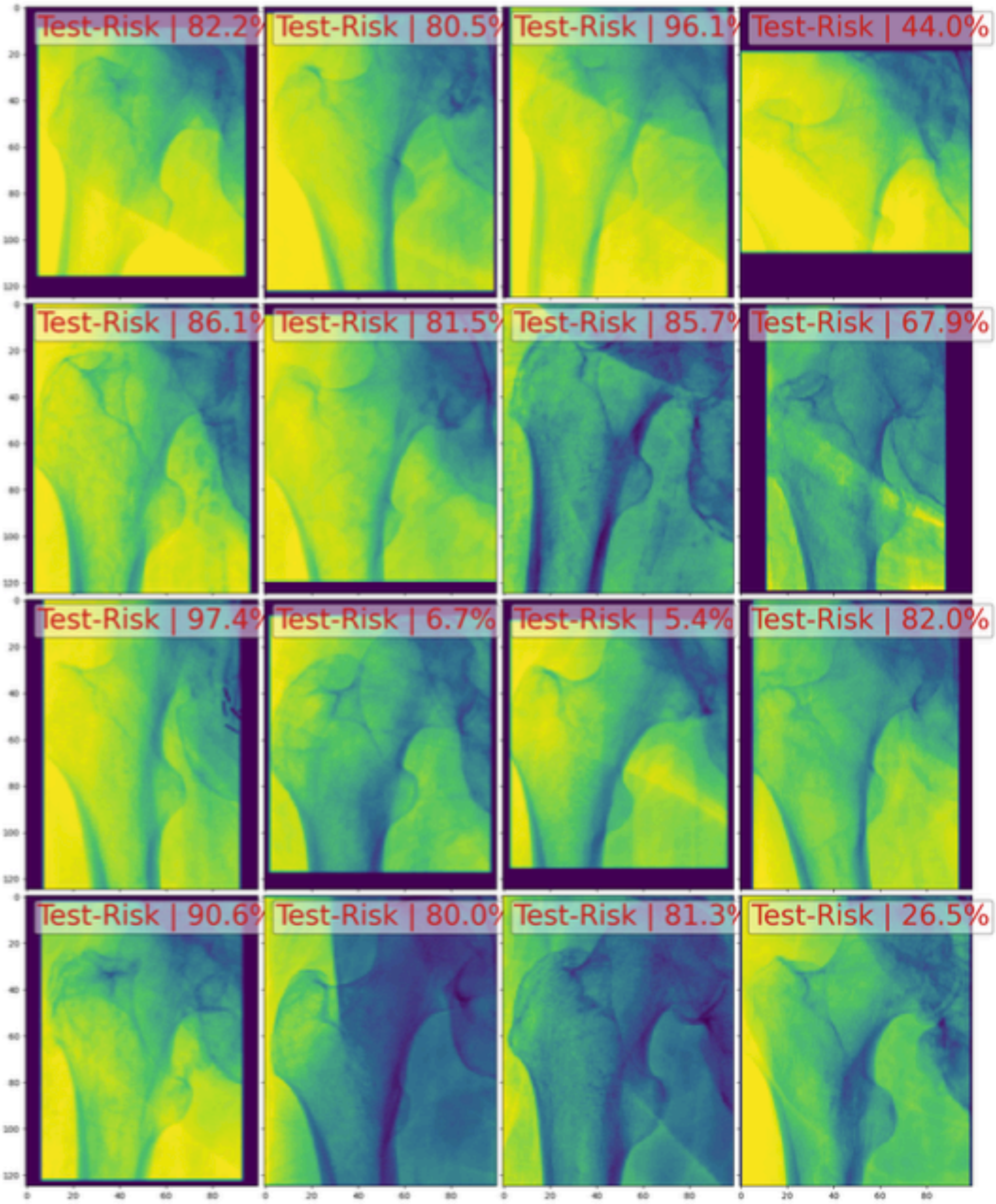

Fig 3. Test-set at-risk images and the selected model estimations.