Using Deep Learning Approaches to Detect Osteoarthritis in Radiographs

Defining the Problem

- Given radiographs of a patellar joint, can I use Deep Learning techniques to detect and diagnose Osteoarthritis (OA)?

The Data

- Kaggle: Knee Osteoarthritis Dataset with Severity Grading
- URL: https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity/data

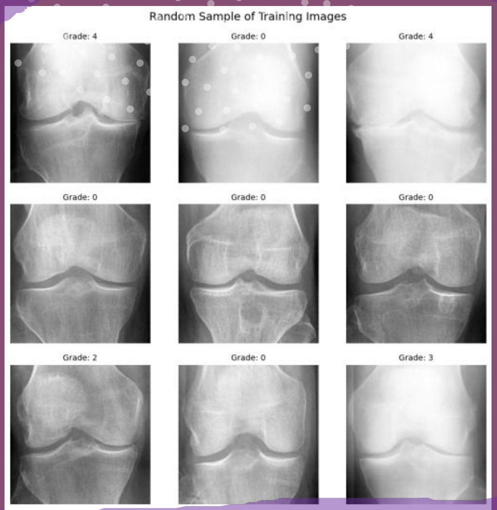
- Future aspirations:
  - I would love to expand this project into the veterinary space and hope to gather a comparable veterinary OA radiograph dataset in the future to apply this work to.

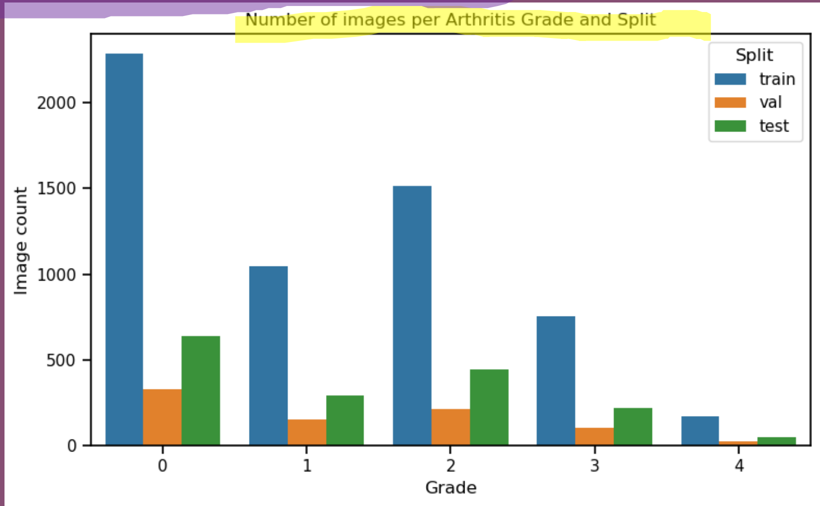
The Process

- EDA (exploratory data analysis)
- CNN (convolutional neural network)
- RNN (recurrent neural network)
- CNN version 2
- CNN version 2 – tuned
- Comparison of Approaches Used vs Human Accuracy Data
- Recommendations for Future Application & Research
- Code & slides have been made freely available on Github at https://github.com/Kate-Zilla/deep-learning-arthritis-detection

EDA (exploratory data analysis)

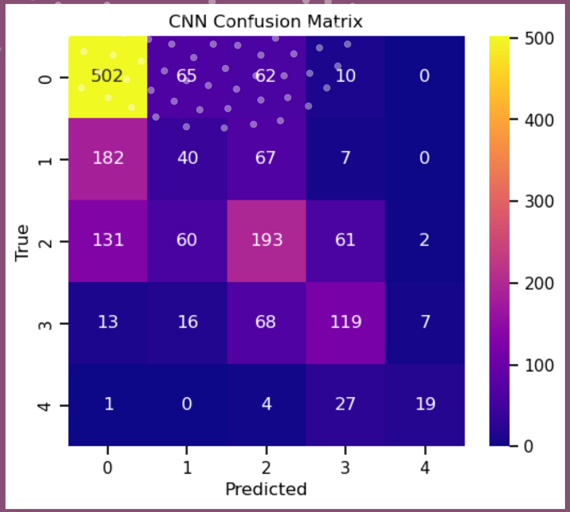
- Determined basic counts per class per split
- Sampled a 3 x 3 grid of images of random OA grades
- Confirmed all images are 224 x 224





Initial CNN model (Convolutional Neural Network)

- Utilized 3 convolutional layers with maxpooling, output achieved using softmax
- Trained on 30 epochs
- Not the best performance, especially when classifying into the 5 grades of human OA
- Took a little less than 40 minutes to train.

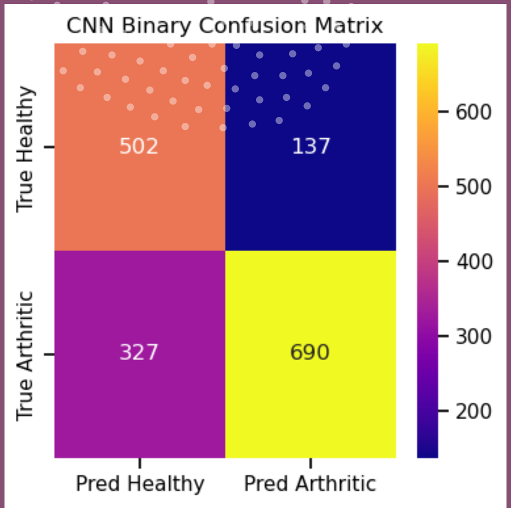


CNN – Test accuracy: 0.527

	precision	recall	f1-score	support
0	0.61	0.79	0.68	639
1	0.22	0.14	0.17	296
2	0.49	0.43	0.46	447
3	0.53	0.53	0.53	223
4	0.68	0.37	0.48	51
accuracy			0.53	1656
macro avg	0.51	0.45	0.46	1656
weighted avg	0.50	0.53	0.50	1656

Initial CNN model- collapsed to binary diagnosis

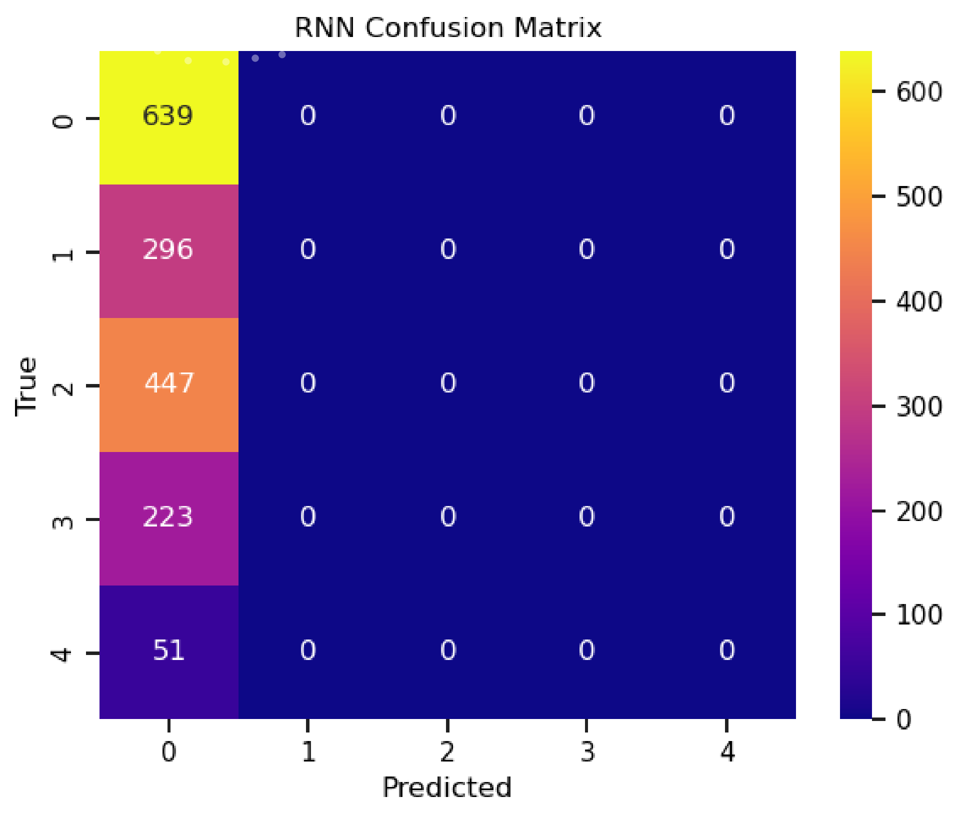
- I did collapse the model to the binary of “healthy vs arthritic” to compare with human performance.
- Those diagnostic metrics are shown here:




```
=== CNN Diagnostic Metrics (Binary: Healthy vs Arthritic) ===
Sensitivity: 0.678
Specificity: 0.786
PPV:         0.834
NPV:         0.606
Accuracy:    0.720
```

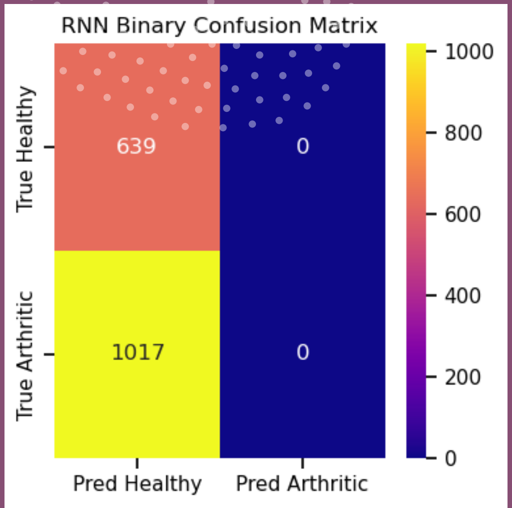
RNN (Recurrent Neural Network)

- Trained on 30 epochs, just like the initial CNN model (for comparability purposes)
- Fully trained in about 2 minutes (compared to CNN's ~40)
    - Red flag?
- RNN test accuracy was 0.386
    - Not great
- Poor performance shown on confusion matrix, where it oversimplifies and collapses



RNN diagnostic metrics (collapsed to binary)

- Again, I collapsed the RNN metrics to binary to compare with human diagnostic performance
- Results were atrocious; this model cannot be recommended for use in this purpose



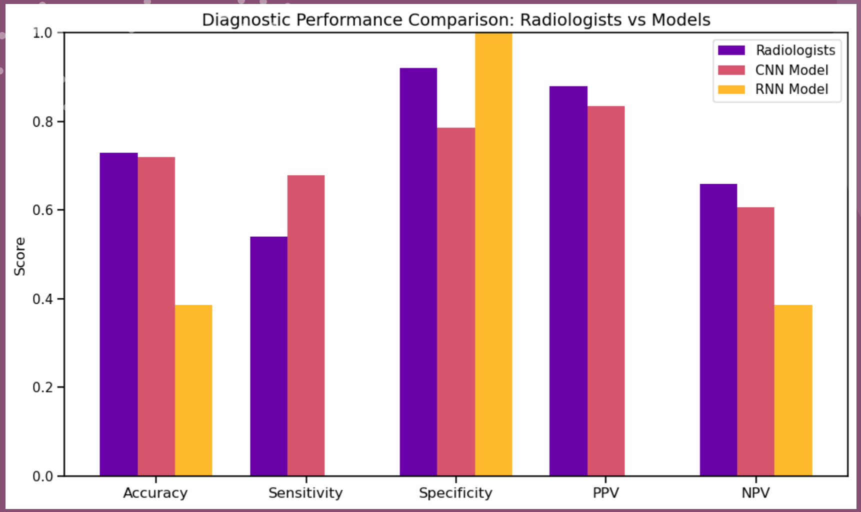

```
=== RNN Diagnostic Metrics (Binary: Healthy vs Arthritic) ===
Sensitivity: 0.000
Specificity: 1.000
PPV:         nan
NPV:         0.386
Accuracy:    0.386
```

Human Metrics

- According to the study at the URL below, radiographic diagnosis of knee arthritis in humans has been shown to have
  - Overall accuracy 73%
  - PPV 88%
  - NPV 66%
  - Sensitivity 54%
  - Specificity 92%

- https://acrjournals.onlinelibrary.wiley.com/doi/full/10.1002/art.42368

Initial Comparisons

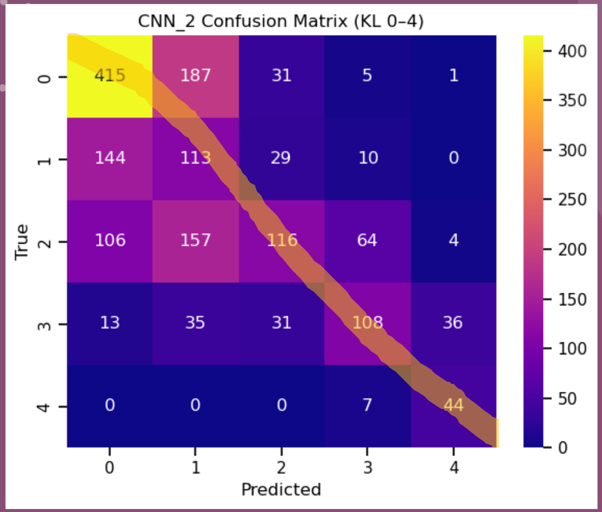


	Metric	Doctor	CNN Model	RNN Model
0	Accuracy	0.73	0.720	0.386
1	Sensitivity	0.54	0.678	0.000
2	Specificity	0.92	0.786	1.000
3	PPV	0.88	0.834	NaN
4	NPV	0.66	0.606	0.38[illegible]

Initial Results & plans for CNN version 2

- Initial CNN model is very close to human accuracy
  - I want to try to beat it
- CNN_2 will have:
  - Improved data augmentation
  - Class weights
  - Deeper filter size (128)
  - 10 additional epochs

CNN_2 Performance Analysis

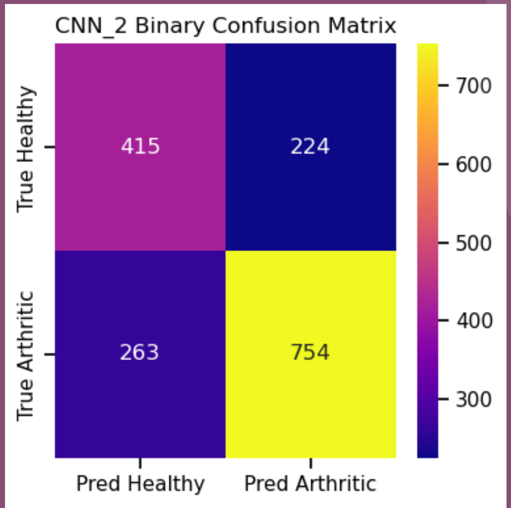



```
CNN_2 — Test accuracy: 0.481
Classification report (CNN_2):
              precision    recall  f1-score   support

     Grade 0       0.61      0.65      0.63       639
     Grade 1       0.23      0.38      0.29       296
     Grade 2       0.56      0.26      0.35       447
     Grade 3       0.56      0.48      0.52       223
     Grade 4       0.52      0.86      0.65        51

    accuracy                           0.48      1656
   macro avg       0.50      0.53      0.49      1656
weighted avg       0.52      0.48      0.48      1656
```

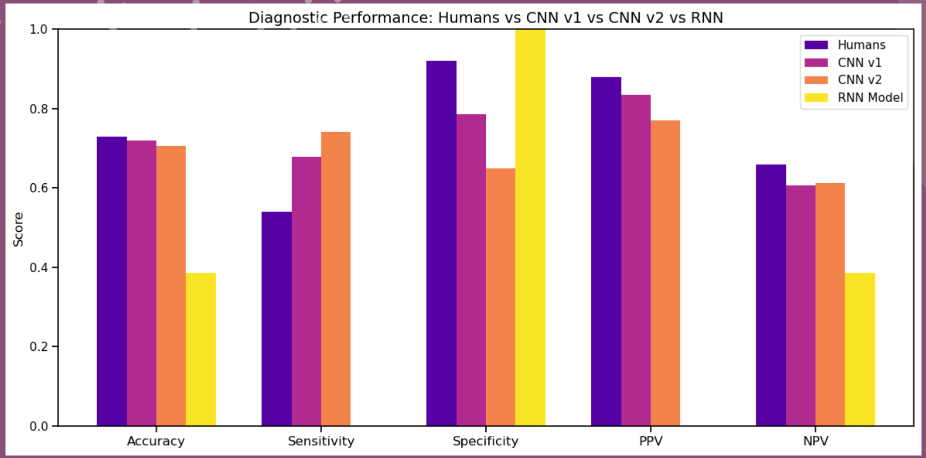
CNN_2, as binary




```
=== CNN_2 Diagnostic Metrics (Binary: Healthy vs Arthritic) ===
Sensitivity: 0.741
Specificity: 0.649
PPV:         0.771
NPV:         0.612
Accuracy:    0.706
```

More comparisons

Not what I expected to see, so I will investigate further.



	Metric	Humans	CNN v1	CNN v2	RNN Model
0	Accuracy	0.73	0.720	0.706	0.386
1	Sensitivity	0.54	0.678	0.741	0.000
2	Specificity	0.92	0.786	0.649	1.000
3	PPV	0.88	0.834	0.771	NaN
4	NPV	0.66	0.606	0.612	0.386

Further Investigation: Confusion Breakdowns

- My changes created a trade-off situation:
- The second version missed fewer cases of OA, while the first had fewer false OA diagnoses

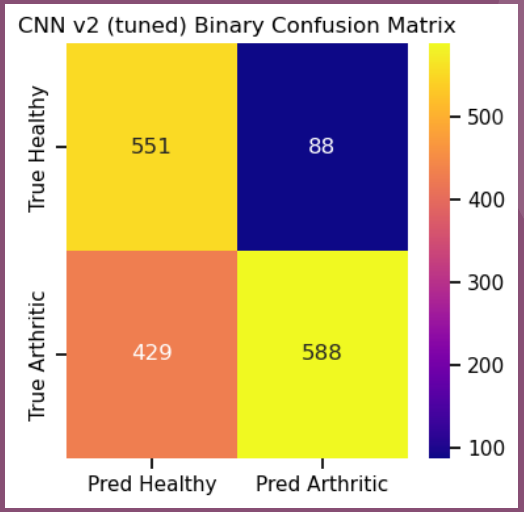
```
CNN v1 (binary):
  TN (true healthy)   : 502
  FP (false arthritic): 137
  FN (missed OA)      : 327
  TP (correct OA)     : 690

CNN v2 (binary):
  TN (true healthy)   : 415
  FP (false arthritic): 224
  FN (missed OA)      : 263
  TP (correct OA)     : 754
```

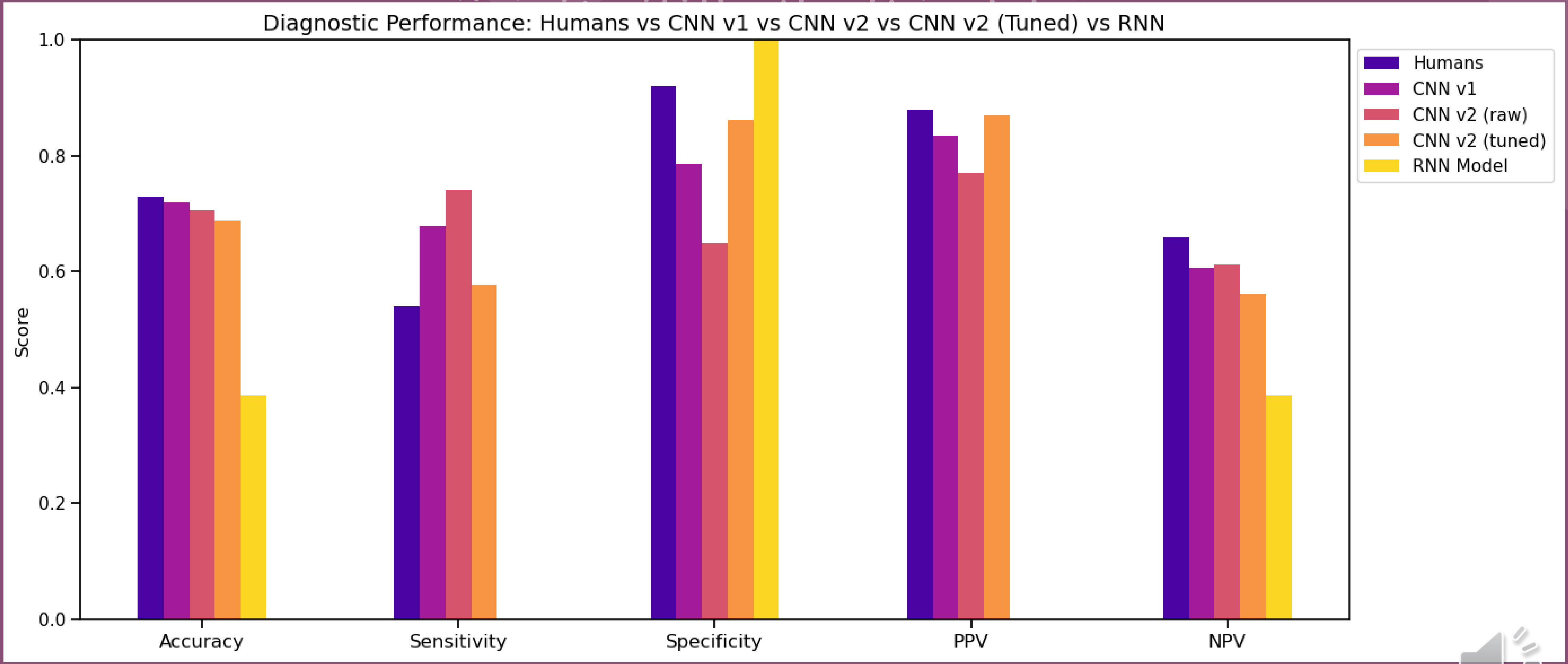
Tune the decision threshold

- It may be beneficial to tune the decision threshold based on the validation data
  - This is standard practice with radiology AI, pathology AI, and other clinical applications
  - It also compensates for model biases
  - Allows CNN to transform from classification experiment to diagnostic tool

Results of tuning CNN_2




```
=== CNN v2 (tuned threshold) Diagnostic Metrics ===
Sensitivity: 0.578
Specificity: 0.862
PPV:         0.870
NPV:         0.562
Accuracy:    0.688
```


Diagnostic Performance: Humans vs CNN v1 vs CNN v2 vs CNN v2 (Tuned) vs RNN
Score
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Sensitivity
Specificity
PPV
NPV
Humans
CNN v1
CNN v2 (raw)
CNN v2 (tuned)
RNN Model

Results

- The first CNN model, as well as the tuned version of the second CNN model are comparable with human doctor/baseline results. The raw CNN version 2 lost some accuracy because it favored sensitivity over specificity.
- Sensitivity: This is the ability to actually detect the arthritis. CNN version 2 (raw) was the best overall at detecting the disease, closely followed by CNN version 1. CNN version 2 still performed higher than the human level.
- Specificity: This is the ability to detect healthy joints. Humans scored the highest here, at 92%. CNN version 2 (tuned) was the best ML performer here, scoring a respectable 86%. Without tuning, the second version scores the lowest in this category, insinuating that it is likely overpredicting instances of OA.
- PPV (Positive Predictive Value): This is the chance that the knee is actually arthritic if the model predicts it to be. Humans score the highest at 88%, while the second version of the CNN model score extremely close behind, at ~87%.
- NPV (Negative Predictive Value): This is measuring how reliable a 'healthy' prediction is. Humans, at 66%, are not very reliable in this instance; however none of the models score any better. All CNN models, however, all score above 55%, giving reasonably close to human performance.
- NOTE: RNN does not appear in the categories of Sensitivity and PPV, above, as it collapsed to a trivial classifier, giving perfect specificity but terrible sensitivity, and no true positives, so the PPV is undefined. It is therefore considered to be an inappropriate method of radiographical OA classification.

Future Applications, Implications, and Takeaways

- Tuned second CNN iteration shows the most promise
  - Achieved nearly human-level confidence with positive diagnoses (PPV)
  - While it did favor sensitivity over specificity, I believe it to be the best baseline for future iterations of applying deep learning to radiographic OA diagnosis
- My hope is to adapt this model for future use on radiographs of canine and equine patellar joints, once an appropriate dataset can be sourced

Potential further refinements

- Given greater amounts of resources (including computing power, time, etc), I would consider looking into:
  - different methods for cropping the images, possibly employing a simple bounding-box or even a separate segmentation model
  - higher resolution images to preserve views of smaller osteophytes and more subtlety in general
  - effects of ensembling, different pooling methods, residual/skip connections
  - a larger and more diverse dataset (including the multi-species goal)
  - greater computational power, use a stronger GPU to experiment with higher resolutions, larger batch sizes, deeper models, etc
  - acquire more data on human accuracy instead of just relying on the one study I was able to find. I would love to instigate my own study into the matter to personally have more confidence in my own baseline data.

References

- Chen, Pingjun. “Knee Osteoarthritis Severity Grading Dataset.” *Data.mendeley.com*, vol. 1, 4 Sept. 2018, data.mendeley.com/datasets/56rmx5bjcr/1, https://doi.org/10.17632/56rmx5bjcr.1.
- “Knee Osteoarthritis Dataset with Severity Grading.” *Www.kaggle.com*, www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity/data.
- Sirotti, Silvia, et al. “Reliability and Diagnostic Accuracy of Radiography for the Diagnosis of Calcium Pyrophosphate Deposition: Performance of the Novel Definitions Developed by an International Multidisciplinary Working Group.” *Arthritis & Rheumatology*, vol. 75, no. 4, 19 Jan. 2023, pp. 630–638, https://doi.org/10.1002/art.42368. Accessed 4 Mar. 2024.