Automatic Measurement of Patellar Tilt using Deep Learning Methods

Samuel Hunter
Stanford University
syhunter@stanford.edu

Tofunmi Omiye
Stanford University
tomiye@stanford.edu

Marissa Lee *
Stanford University
marissalee@stanford.edu

Abstract

Patellar tilt is a risk factor for patellar instability and is important in the clinical decision process for patellar surgery. Standards of care involves manually measuring patellar tilt on knee MRI and X-ray images. This is largely plagued by inter-observer variation and is relatively inaccurate. We propose the successful use of transfer learning on a convolutional neural network (CNN) model for the prediction of keypoints (segmentation) and tilt measurement (regression) for this task. We additionally used standard augmentation techniques including elastic deformations to expand our training dataset and further increase the accuracy of our model. Our model performed with a 55.5% reliability before data augmentation and 76.8% reliability after, as compared to radiologist measurements. This shows how data augmentation can overcome the obstacle of generalizing from a small labeled dataset.

1. Introduction

Patellar instability is a clinical syndrome where the patella bone disarticulates from the patellofemoral joint in the knee [19]. It is a significant risk factor for recurrent patellar dislocation and could necessitate surgical intervention. It is commoner in adolescents, especially within the ages of 14-18 years, with cases of about 148 per 100,000 [20]. Its peak prevalence in this age can be attributed to the rapid growth spurt experienced and increased activity. In addition, major risk factors for patellar instability includes connective tissue disorders like Ehlers-Danlos syndrome, ligament laxity, anatomical abnormalities, and acute trauma [26]. Also, a previously diagnosed patellar instability poses a high risk for persistent instability throughout ones lifetime [26].

Clinical presentation of this condition is usually knee pain, recurrent dislocation, signs of acute trauma, and knee rotation. Beyond clinical presentation, definitive diagnosis of patellar instability is via radiological imaging [5].

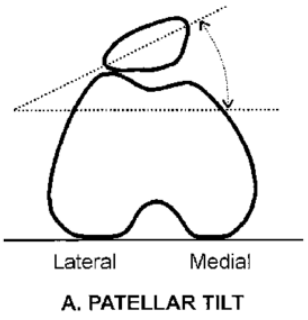


Figure 1. Measurement of patellar tilt from condyle and patella keypoints. Image credit [21].

This can be via X-ray and increasingly magnetic resonance imaging (MRI). MRI is useful for specific tissue delineation and shown to be best for assessing medial injuries. In evaluation of these imaging modalities, physicians heavily rely on a manual method for measuring the angle between the condyle (lateral to medial) and the patella (lateral to medial) line Figure 1. This is the patellar tilt. The patellar tilt measurement is made through different radiological parameters and has been shown to vary widely amongst observers with little to no consensus mechanism. Greater patellar tilt is more pathologic, indicative of imbalance and increased load on the patella compared to the medial quadriceps muscle, leading to a high risk of dislocation.

Beyond the diagnosis of patellar instability, patellar tilt is also useful for planning management interventions for patients with recurrent instability. Physicians rely on this tilt measurement to guide the decision to offer surgical treatment, amongst other factors [5]. The variability in patellar tilt measurement provides an opportunity to use deep neural networks to automate the process with an accurate and fast alternative. The standard values are typically: 10 ± 4.3 degrees, and surgically viable cases are within 16 ± 3.3 degrees range.

In this work, we utilized convolutional neural network (CNN) to automate this measurement process. Specifically,

*advisor, not enrolled in CS 231N

our input to our neural network were radiologist-labeled adolescent MRI images of the knee showing four keypoints: the lateral and medial aspect of the patella and the lateral and medial condyles. The raw measurements are stored in DICOM images that are volumes of 2D cross-sections (slices) along the knee. Our algorithm was based on a U-Net architecture to output predicted measurements of the keypoints coordinates and calculate the patellar tilt using the coordinates. Since we had a relatively small dataset of MRI images, we applied an elastic deformation method for data augmentation that increased our model accuracy to 76.8%, compared with manual measurement.

This project is particularly important as it shows the potential of CNN application in automating patellar tilt measurement and the utility of data augmentation for medical images. Also, we are unaware of studies applying deep neural networks to this problem.

2. Related work

Increasingly, deep learning and particularly CNN have been applied to solving medical problems with medial imaging taking significant attention compared to other medical fields [23]. Neural networks have been ubiquitously applied to chest X-ray imaging for relatively simple tasks like identifying the orientation of an X-ray to more complex ones like pneumonia diagnosis or heart failure prediction [15]. Other imaging tasks like the diagnosis of skin cancer [4] and detection of diabetic retinopathy from retinal scans have been successful [6]. For MRI, neural networks have also been utilized for various tasks like contrast enhancement classification and even reconstruction [13].

Virtually, all the imaging tasks described above have utilized CNN for working with medial images with varying degrees of complexity. The potential for pre-trained neural networks for reconstruction of MRI images was explored by Knoll et al. [11]. U-Net is a popular neural network architecture developed to address the nuances of image segmentation in medical imaging [16]. U-Net is particularly suitable for this case as it can create highly detailed segmentation maps and since its release, adaptations of the architecture have been applied to various tasks in medical imaging, including X-rays and MRI. This made it our choice of architecture to adapt to this task.

A problem with deep learning architectures is that it needs massive amounts of data for training, this is especially worrisome in medical imaging as while there are many images available, there's not a corresponding amount of labeled data. This is even more important in rare diseases where data is hard to come by. Various augmentation techniques have been described in medical image analysis for deep learning. Basic transformations like flipping, cropping, and other geometric transformations have been used with varied success. When applying augmentation, it is important to consider the specific use case, especially for medical images where particular transformations can have the opposite desired effect on the model. We used augmentation techniques that have been described in the literature and applied them to our specific task. Elastic deformations are generated from random displacement fields that are then smoothed with a Gaussian. It is superior in image classification tasks. For example, Platt et al. found that while affine transformations reduced error on an image classification task for both fully-connected and CNNs to under 1%, elastic deformations further reduced error to 0.4% [18]. Similarly, Brox et al. applied elastic deformations to the segmentation of biomedical images, which improved invariance to deformed tissue [16].

U-Net is also particularly famous for introducing how elastic deformation could be very powerful in medical imaging. UOLO [2] was a framework based on U-Net [16] for simultaneous detection and segmentation of the fovea and optic disc. It took advantage of U-Net's autoencoder structure with skip connections that allow for training on fewer training samples, unlike Mask R-CNN [7]. The soft intersection over union (IoU) was used as their loss function. By using multi-scale information, UOLO became robust to low contrast input images and achieved 0.82 IoU for segmentation in the Messidor dataset.

While there have been several deep learning applications for MRI, none of these has applied a deep learning approach for the task of patellar tilt measurement. We implement variations of the methods described above, and details are revealed in the following sections.

3. Dataset

We obtained data from the JUPITER (JUstifying Patellar Instability Treatment by Early Results) group [3], which is a subset of the data from Cincinnati Children's Hospital Medical Center. JUPITER is a multi-institutional, multi-armed, prospective cohort study specially designed to obtain sufficient number of subjects to better describe the clinical characteristics and predictors of clinical outcome in patients with patellar instability.

The patients in this dataset have age range from 10 to 25 years with sustained patellar dislocation/subluxation. Our dataset is comprised of 216 lateral pre-treatment MRIs with keypoints and patellar tilt computed. The slices of each MRI move from the distal end of the femur through the proximal end of the tibia. Depending on the setting of the MRI, some images may be brighter around areas of concentrated fat, and there might be some artifacts in some images. Pre-processing should account for these variations.

The sizes of the MRI in our dataset ranged from 256×256 to a maximum of 1024×1024. This was also resized during our preprocessing. 80% of the data (173 MRIs) was used for training, and the remaining 20% (43 MRIs) for test-

ing. There were too few MRIs for validation, and since there were few hyperparameters and the model was quick to train. We also applied elastic deformation as a data augmentation technique based on the reasons described earlier in this paper.

For each image, a radiologist had labeled the coordinates of each of the four keypoints on both the condyles and patellas, as well as the relevant slices indices that the measurements were taken.

4. Methods

4.1. Preprocessing

The MRIs were collected from many patients and MRI machines, so the measurements varied greatly in range, variance, and resolution. To standardize the input, we rescaled the MRI images to 128×128 pixels, subtracted the minimum value and divided by the maximum value. Finally, all images of right legs were horizontally flipped to match the left legs, since the patella and condyle are nearly symmetrical. This avoids the CNN needing to predict the same keypoint on both the left and right halves of the images.

The same transformations were applied to the keypoint mask and keypoint coordinates. The OpenCV python library was used.

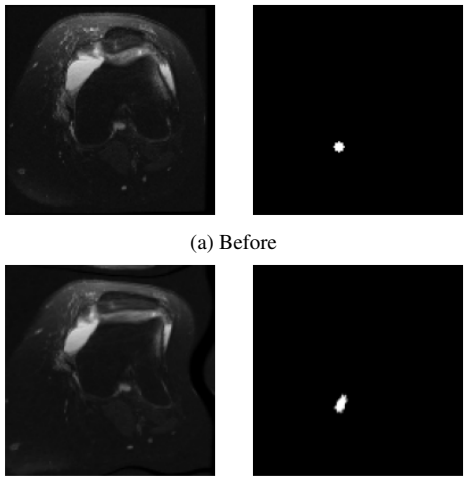
One simplification we made was only using the relevant labeled MRI slices for the condyle and patella. This reduced the amount of extraneous data the model would need to be trained on.

4.2. Data Augmentation

With so few annotated training images, we needed to augment our dataset for the model to generalize better. As recommended by U-Net [16], we employed elastic deformations to generate synthetic training samples. First, we generate a grid of random displacement vectors from a Gaussian distribution. Next, we blur the displacement vector grid for a smooth result. The original MRI and keypoint coordinates are warped according to this displacement. During every training batch, a randomized elastic deformation was applied to each image. See Fig. 2.

4.3. ResNet

ResNet is a popular deep neural network for image classification [8]. Its architecture includes residual connections, which are shortcuts between convolutional layers than prevent vanishing gradients during backpropogation. The advantage of transfer learning from a well-established model is good feature extraction and significantly reduced training time. However, since the architecture is designed for classification, we retrained a pretrained ResNet with 2 output units. The input was a 3 channel 512×512 pixel MRI slice and the output was 2D coordinates of the keypoint. The optimizer was stochastic gradient descent with a learning rate of 2×10^{-3} and momentum of 0.9. For the loss function, we simply took the L2 distance between the predicted and actual keypoint. This model produced predictions with an average distance of 40 pixels to the actual keypoint. This accuracy was not sufficient.



(a) Before

(b) After

Figure 2. Elastic deformation. This is an extreme deformation for illustrative purposes; the actual dataset was augmented with smaller parameters.

4.4. U-Net

U-Net is a fully convolutional architecture with symmetrical contractive and expansive paths that downsample then upsample the input [16], Fig. 3. The final layer is a 1x1 convolution that maps the features into the segmentation classes. It was designed for biomedical images. Also, since there are no fully connected layers, the input can be of any size.

Instead of directly predicting the coordinates as with ResNet, we generated a target keypoint mask which was a binary image of the same dimension as the MRI. It was black everywhere except a small circle around the keypoint, where it was white. The output of U-Net is the same height and width as the input, but with k output channels corresponding to the unnormalized probability of the corresponding pixel in the input belonging to the kth class. To match this, we encoded the single channel binary keypoint mask as a two channel image of one-hot vectors, where the 0th channel corresponds to the background, and the 1st channel corresponds to the keypoint.

As with the original paper, we trained the U-Net models with cross entropy loss. The input is a 3 channel MRI slice. The target is a 2 channel 128×128 pixel mask. We found that a higher resolution input of 512×512 pixels

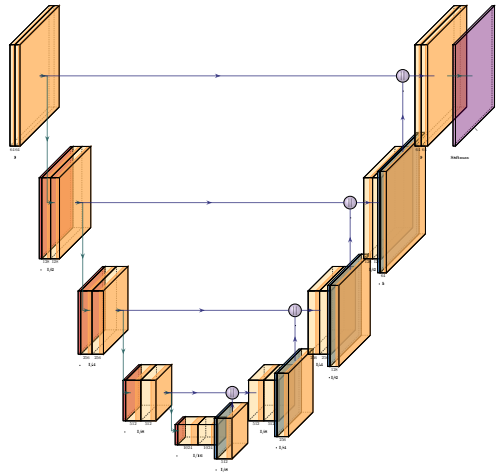


Figure 3. U-Net architecture. Image credit [9].

did not increase training accuracy and only slowed down training time. In our first attempt, we used Mean Square Error (MSE) loss between the predicted and target keypoint mask. While it reduced initial error more rapidly than cross entropy loss, it failed to produce more accurate results with additional epochs. The cross entropy loss between the predicted segmentation map and the target mask is computed as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{1}$$

where y is the true label and $\hat{y}$ is the predicted label.

To predict the coordinates of the keypoint given its predicted mask, we first took the softmax over the two channels, which normalizes the class probabilities. We then only needed to consider the 1st channel since it corresponds to the probability of a pixel in the MRI slice belonging to the keypoint class. Next, we used OpenCV to blur and locate the position of the brightest pixels. These coordinates were the predicted keypoint location. Blurring made the max location robust to noise in the predicted mask. This approach for object detection was inspired by [2]. See Fig. 4.

4.5. Combined Keypoint Architecture

To predict patellar tilt, we combined four separate U-Net models to produce a segmented outputs for each of the four keypoints from the MRI slices. The keypoint coordinates were extracted from the segmented outputs as described above, then two lines were drawn. The lateral and medial condyle form the condyle line, and the lateral and medial patella form the patella line. The angle between these lines is the patellar tilt. A drawback to this approach is that with four separate models, if any one of the keypoint predictions is off, the angle error will be very high.

5. Experiments and Results

We trained each keypoint predictor model with weights initialized from a pretrained U-Net [1] for 100 epochs in pytorch [17]. We used the Adam [10] optimizer with a learning rate of 5×10^{-3} and a batch size of 32. Initially, we tried a batch size of 1 with fewer (10-30) epochs, but the training loss plateaued. Training took about 35 minutes on a GTX 1660 Super.

Early in training, we observed that occasionally the U-Net mask predictions would have several or no bright regions corresponding to a keypoint. This produced an outlier keypoint prediction that would drastically skew the computed patellar tilt outside of a reasonable range. To account for this, any keypoint greater than a maximum distance (50 pixels) from the median training keypoint coordinates would be replaced by the median. This replacement technique removed angle errors of over 40 degrees, bringing the standard deviation in angle error down to under 5 degrees. See Figs. 6 and 7. However, as the number of epochs increased, the number of outliers that needed replacement by the median dropped.

Our metric of performance of the patellar tilt predictions was intraclass correlation coefficient (ICC) [22]. ICC measures the reliability of measurements or ratings from several raters (e.g., human vs human or model vs human). ICC ranges from 0, indicating no reliability among raters, to 1, indicating perfect reliability among raters. We used ICC2, which measures absolute agreement in the ratings for a single rater, between the model and the radiologist labels.

$$\text{ICC}(2,1) = \frac{\text{BMS} - \text{EMS}}{\text{BMS} + (k-1)\text{EMS} + k(\text{JMS} - \text{EMS})/n} \tag{2}$$

where BMS is between-targets mean square, EMS the residual mean square (error), JMS the between-rater mean square, k the number of raters an n the number of subjects.

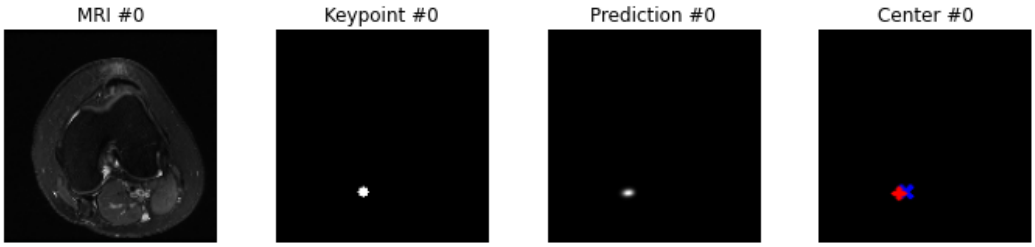
ICC values between 0.75 and 0.9 indicate good reliability [12]. For reference, the ICC2 for patellar tilt between two human raters (Marissa Lee and a radiologist) was 0.816. See Tab. 2.

Also, with such a small set of training images (170), the elastic deformations enhanced the keypoint prediction accuracy by providing randomized inputs during each training epoch. Data augmentation increased the ICC2 by 0.213.

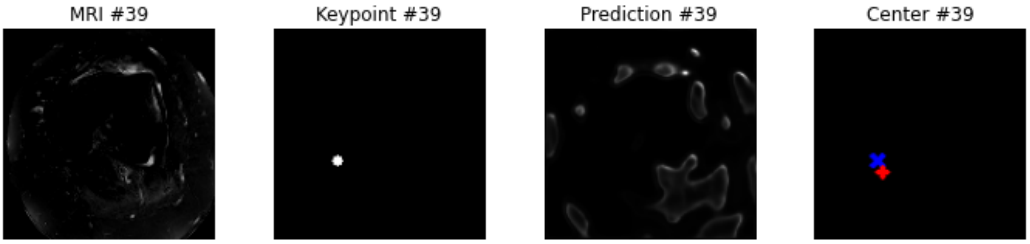
A keypoint prediction from the combined keypoint architecture with data augmentation can bee seen in Fig. 5.

6. Conclusion and Future work

For predicting keypoints in medical images, the segmentation approach proved more flexible and generalizable to augmented data than direct coordinate prediction. Data augmentation exposed the model to far more training samples



(a) Successful predicted mask with one bright spot.



(b) Unsuccessful predicted mask with multiple bright spots. Note the predicted coordinates have been replaced by the median.

Figure 4. From left to right: Input, Target output, Predicted output, Actual (blue) vs. Predicted (red) lateral condyle keypoint coordinates.

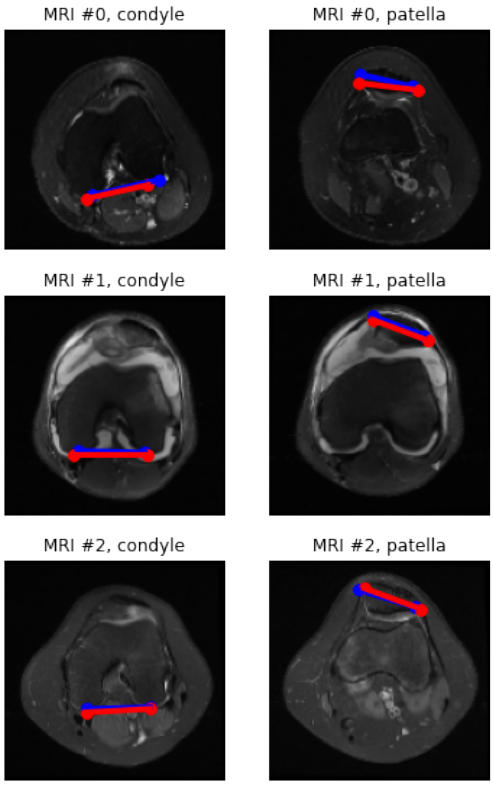


Figure 5. Actual (blue) vs. Predicted (red) keypoints for several condyle (left) and patella (right) test MRI slices. Notice the variety of contrast levels, bone shapes, and fat tissue (white) in the original MRIs.

Keypoint model	Number of Outliers	Error (pixels) Avg.	Max
Lateral condyle	1	3.395	26.653
Medial condyle	3	2.885	27.625
Lateral patella	2	2.445	8.285
Medial patella	1	2.568	15.987

Table 1. Keypoint prediction error statistics.

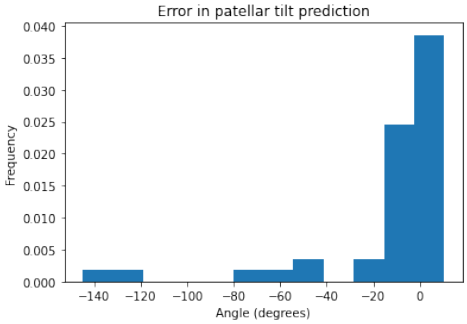
Model	ICC2
U-Net vs. Human	0.555
U-Net (augmented data) vs. Human	**0.768**
Human vs. Human	0.816

Table 2. Patellar tilt reliability results.

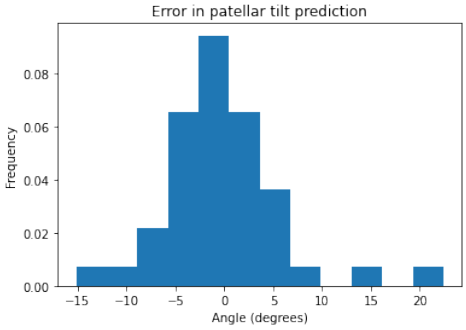
and allowed the training loss to continue to fall after far more epochs.

In the future, a model that simultaneously predicts all of the keypoints would be faster to train, but less straightforward to tune. In our experience, predicting multiple keypoints, then localizing the brightest regions in the predicted mask sometimes produced fewer keypoints than necessary to make the angle prediction, which meant unpredictably low or high error. An alternative model would be to train a single keypoint predictor for a large number of epochs, then training the remaining predictors from its learned weights for less time.

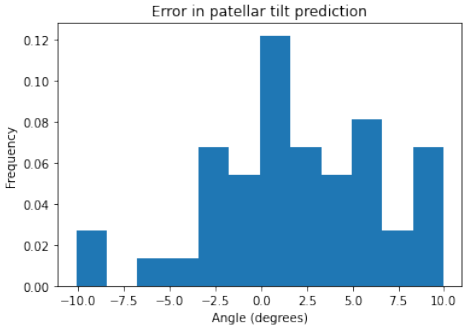
A simplification that was made for our model was pre-



(a) Without outlier replacement.
Standard deviation: 31.725 degrees



(b) With outlier replacement.
Standard deviation: 6.013 degrees



(c) With outlier replacement and data augmentation.
Standard deviation: 4.658 degrees

Figure 6. Tilt error with and without outlier keypoint replacement.

filtering the MRI slices that corresponded to the patella and condyle lines. These slices of the MRI were labeled and extracted before feeding them to the model, which reduces the automation of the process. A future model could either take all of the slices as input or have a separate predictor for selecting the relevant slices for measurement.

7. Contributions and Acknowledgements

Tofunmi Omiye augmented the dataset by applying elastic transformations to the MRI images. This preprocesing step was crucial in allowing the CNN to generalize on such a small pool of images. Samuel Hunter adapted the U-Net

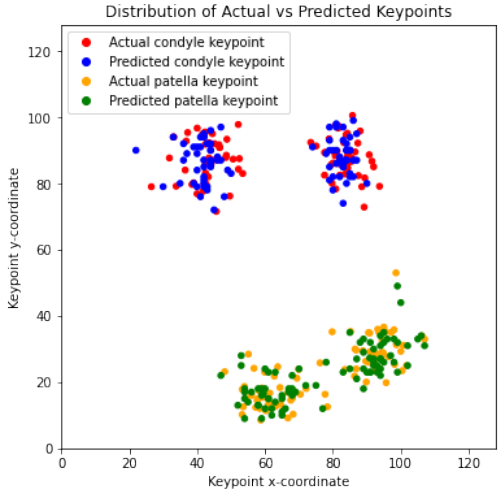


Figure 7. Distribution of keypoint predictions after outlier replacement.

architecture to predict the segmentation maps for the keypoints and aggregated the predictions for visualization and angle prediction. He also tuned the training parameters. Tofunmi Omiye and Samuel Hunter wrote the paper. Marissa Lee, PhD candidate in Mechanical Engineering, proposed and advised on the project and dataset access.

We used several open source python libraries: numpy and pandas for general computation, matplotlib and PlotNeuralNet [9] for visualization, pydicom [14], cv2, and elasticdeform [25] for MRI manipulation, pingouin [24] for statistics, pytorch [17] for creating and training the model, and Pytorch-UNet [1] for pre-trained U-Net weights.

References

[1] Milesi Alexandre. Pytorch-unet. https://github.com/milesial/Pytorch-UNet, 2022. 4, 6

[2] Teresa Araújo, Guilherme Aresta, Adrian Galdran, Pedro Costa, Ana Maria Mendonça, and Aurélio Campilho. UOLO - automatic object detection and segmentation in biomedical images. *CoRR*, abs/1810.05729, 2018. 2, 4

[3] Meghan E Bishop, Jacqueline M Brady, Simone Gruber, Matthew Veerkamp, Joseph T Nguyen, Daniel W Green, Eric J Wall, Shital Parikh, Beth E Shubin Stein, and JUPITER Group. Descriptive epidemiology study of the justifying patellar instability treatment by early results (jupiter) cohort. *Orthopaedic Journal of Sports Medicine*, 9(7_suppl3):2325967121S00144, 2021. 2

[4] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017. 2

[5] Peter D Fabricant, Madison R Heath, Douglas N Mintz, Kathleen Emery, Matthew Veerkamp, Simone Gruber, Daniel W Green, Sabrina M Strickland, Eric J Wall, Beth

E Shubin Stein, et al. Many radiographic and magnetic resonance imaging assessments for surgical decision making in pediatric patellofemoral instability patients demonstrate poor interrater reliability. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 2022. 1

[6] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016. 2

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3

[9] Haris Iqbal. Plotneuralnet. https://github.com/HarisIqbal88/PlotNeuralNet, 2020. 4, 6

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4

[11] Florian Knoll, Kerstin Hammernik, Erich Kobler, Thomas Pock, Michael P Recht, and Daniel K Sodickson. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magnetic resonance in medicine*, 81(1):116–128, 2019. 2

[12] Terry Koo and Mae Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 03 2016. 4

[13] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019. 2

[14] et al Mason, D. L. pydicom: An open source dicom library. https://github.com/pydicom/pydicom, 2022. 6

[15] Takuya Matsumoto, Satoshi Kodera, Hiroki Shinohara, Hirotaka Ieki, Toshihiro Yamaguchi, Yasutomi Higashikuni, Arihiro Kiyosue, Kaoru Ito, Jiro Ando, Eiki Takimoto, et al. Diagnosing heart failure from chest x-ray images using deep learning. *International Heart Journal*, 61(4):781–786, 2020. 2

[16] Thomas Brox Olaf Ronneberger, Philipp Fischer. U-net: Convolutional networks for biomedical image segmentation. *LNIP*, 9351, 2015. https://arxiv.org/abs/1505.04597. 2, 3

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. 4, 6

[18] John C. Platt Patrice Y. Simard, Dave Steinkraus. Best practices for convolutional neural networks applied to visual document analysis. *ICDAR*, 2003. https://www.researchgate.net/publication/220860992_Best_Practices_for_Convolutional_Neural_Networks_Applied_to_Visual_Document_Analysis. 2

[19] Daniel E Redziniak, David R Diduch, William M Mihalko, John P Fulkerson, Wendy M Novicoff, Shahin Sheibani-Rad, and Khaled J Saleh. Patellar instability. *JBJS*, 91(9):2264–2275, 2009. 1

[20] Thomas Sanders. Incidence of first-time lateral patellar dislocation: A 21-year population-based study. *Sports Health*, 10(2):146–151, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5857724/. 1

[21] Vinayak Sathe, Mary Ireland, Bryon Ballantyne, Nancy Quick, and Irene Davis. Acute effects of the protonics system on patellofemoral alignment: An mri study. *Knee surgery, sports traumatology, arthroscopy : official journal of the ESSKA*, 10:44–8, 02 2002. 1

[22] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86 2:420–8, 1979. 4

[23] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019. 2

[24] Raphael Vallat. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026, 2018. 6

[25] Gijs van Tulder. elasticdeform. https://github.com/gvtulder/elasticdeform, 2021. 6

[26] Steve Wolfe, Matthew Varacallo, Joshua D Thomas, Jeffrey J Carroll, and Chadi I Kahwaji. Patellar instability. *StatPearls [Internet]*, 2018. 1