Journal of Digital Imaging
https://doi.org/10.1007/s10278-022-00671-2

# Artificial Intelligence System for Automatic Quantitative Analysis and Radiology Reporting of Leg Length Radiographs

**Nathan Larson[1] · Chantal Nguyen[2] · Bao Do[3] · Aryan Kaul[4] · Anna Larson[1] · Shannon Wang[4] · Erin Wang[5] · Eric Bultman[3] · Kate Stevens[7] · Jason Pai[7] · Audrey Ha[8] · Robert Boutin[7] · Michael Fredericson[2] · Long Do[6] · Charles Fang[3]**

Received: 15 November 2021 / Revised: 19 May 2022 / Accepted: 7 June 2022


**Abstract**
Leg length discrepancies are common orthopedic problems with the potential for poor functional outcomes. These are frequently assessed using bilateral leg length radiographs. The objective was to determine whether an artificial intelligence (AI)-based image analysis system can accurately interpret long leg length radiographic images. We built an end-to-end system to analyze leg length radiographs and generate reports like radiologists, which involves measurement of lengths (femur, tibia, entire leg) and angles (mechanical axis and pelvic tilt), describes presence and location of orthopedic hardware, and reports laterality discrepancies. After IRB approval, a dataset of 1,726 extremities (863 images) from consecutive examinations at a tertiary referral center was retrospectively acquired and partitioned into train/validation and test sets. The training set was annotated and used to train a fasterRCNN-ResNet101 object detection convolutional neural network. A second-stage classifier using a EfficientNet-D0 model was trained to recognize the presence or absence of hardware within extracted joint image patches. The system was deployed in a custom web application that generated a preliminary radiology report. Performance of the system was evaluated using a holdout 220 image test set, annotated by 3 musculoskeletal fellowship trained radiologists. At the object detection level, the system demonstrated a recall of 0.98 and precision of 0.96 in detecting anatomic landmarks. Correlation coefficients between radiologist and AI-generated measurements for femur, tibia, and whole-leg lengths were $>0.99$, with mean error of $<1\%$. Correlation coefficients for mechanical axis angle and pelvic tilt were 0.98 and 0.86, respectively, with mean absolute error of $<1°$. AI hardware detection demonstrated an accuracy of 99.8%. Automatic quantitative and qualitative analysis of leg length radiographs using deep learning is feasible and holds potential in improving radiologist workflow.

**Keywords** Artificial Intelligence · Deep Learning · Leg Length Discrepancy · Radiography

## Introduction

Leg length discrepancies (LLD) are common orthopedic problems associated with altered weight-bearing and biomechanical changes. In the pediatric population, LLD has been associated with musculoskeletal disorders such as gait disturbances and scoliosis [1, 2]. In the adult population, unmanaged LLD can be associated with long-standing disability and poor functional outcomes after hip and knee joint arthroplasties [3, 4]. In preoperative planning for total knee

✉ Charles Fang
Charles.Fang@va.gov

Nathan Larson
larsnathanson@gmail.com

[1] Computer Science Department, Brigham Young University, Campus Dr, Provo, UT 3361 TMCB84604, USA

[2] Department of Orthopedic Surgery, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA

[3] Department of Radiology, Palo Alto VA Medical Center, 3801 Miranda Ave, Palo Alto, CA 94304, USA

[4] University of California, Los Angeles, Los Angeles, CA 90095, USA

[5] Harvey Mudd College, Claremont, CA 91711, USA

[6] San Jose, CA, USA

[7] Department of Radiology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA

[8] Menlo-Atherton High School, Atherton, CA 94025, USA

Published online: 06 July 2022

arthroplasty, however, leg length measurements alone have been found to be insufficient to guide surgical planning—the literature suggests that excessive angulation can lead to early wear, for instance [5]. Given the potential impact on clinical outcome, accurate leg length measurements with minimal inter- and intra-reader variability can be critical to treatment planning, which can be sensitive to small differences under 5 mm [6].

Bilateral long-leg radiographs, or leg length views, have been essential in preoperative assessment [7]. Simple qualitative, leg length discrepancy is insufficient—comprehensive assessment requires evaluating multiple measures to exclude other considerations such as surgical change, pelvic tilt, and angulation that may result in false leg length discrepancy. These examinations, however, are laborious for radiologist interpretation given the number of potential parameters to assess.

Patient care can thus be improved by using a system that automatically quantifies these metrics as a preliminary step for the clinical radiologist, potentially improving efficiency and reducing intra- and inter-observer variability. Toward this end, there have been recent achievements in AI to detect leg length discrepancy and leg angles [8] as well as other more qualitative features such as knee arthritis from leg length examinations [9]. These advances demonstrate promise in automating this process of normal leg length examinations. However, real-world leg length examinations are also complex, involving hardware, severe deformity, and joint fusion; hence, designing a robust system to handle these more difficult cases is crucial for adoption into the routine workflow.

The goal is to build and validate an AI for interpreting and pre-dictating these studies into the clinical workflow. Previous work can be improved by designing a system that is capable of handling hardware and deformity. Utilizing anatomic landmarks as targets or keypoints for object detection allows for a more intuitive annotation approach that is distinctly less labor intensive compared to the arduous traditional segmentation-based approach. Lastly, an object classifier neural network is added for surgical hardware detection to enable the system to report post-surgical changes and convey appropriate limitations.

## Materials and Methods

### Data Collection

The images for the training and testing datasets were retrospectively acquired from a tertiary referral center Picture Archiving and Communications System (PACS) using a search criteria of age > 16 and a 1-year acquisition window. Images were exported in full-resolution Joint Photographic Experts Group (JPEG) format. To ensure complete deidentification of metadata, JPEG image format was chosen. Images were manually reviewed to ensure no burned-in private health information (PHI) was present. Only the anterior–posterior (AP) full-length lower extremity view was included in the dataset—lateral views and piecemeal images used to assemble the AP view or lateral views were excluded. A total of 863 consecutive studies were exported. This was randomly partitioned into a training and validation set of 643 images and a leave-out test set of 220 images. Nine patients had 2 images, which were randomized together in either the training/validation or the test set. The training and validation set was used to train and validate each of

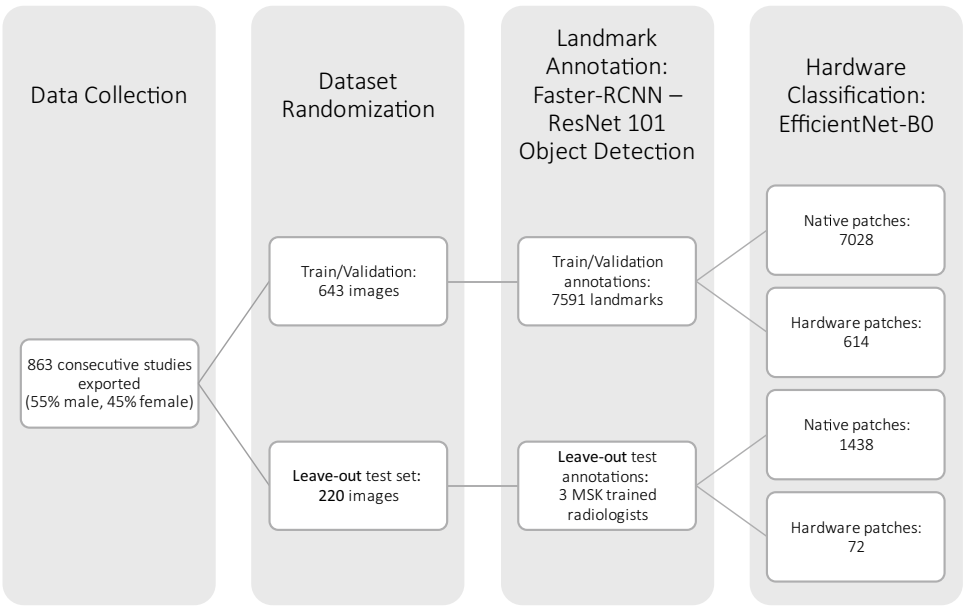



**Fig. 1** A flowchart summarizing the dataset usage throughout the various components

the two AI components. Figure 1 summarizes the dataset usage through the AI components. Demographics for each set can be found in Table 1.

## Definitions

We utilize accepted definitions for anatomic landmarks to calculate each of the desired metrics. The femoral head apex point was defined as the most superior point along the contour of the femoral head. The center of the femoral head point was defined as the center of a circle that would best fit within the femoral head. The ischium point was defined as the most inferior point along the contour of the ischial tuberosity. The intercondylar notch point was defined as the most superior and central point along the intercondylar notch of the distal femur. The tibial spine point was defined as a point immediately subjacent to the tibial spines at the level of the tibial plateaus. The ankle mortise point was defined as the central midpoint of the tibial plafond.

The total length of the leg is defined as the distance from the femoral head apex to the ankle mortise, the femur length is defined as the distance from the femoral head apex to the intercondylar notch, and the tibia length is defined as the distance from the tibial spine to the ankle mortise [10]. The mechanical angle is defined as the angle between the femoral head apex, tibial spine, and ankle mortise [11]. The pelvic tilt is defined by the angle of deviation from horizontal between the bilateral ischium points [12, 13]. All lengths are reported in pixels, as the pixel spacing metadata was discarded as a part of the JPEG conversion. It should be noted, however, that the standard radiography systems used to acquire these images utilized pixel spacings of approximately 0.14 to 0.19 mm/pixel. In a production system, ready access to this data would render it trivial to translate pixel length into units of physical measurement. All angles are reported in degrees. Figure 2a-c shows sample annotations and how the angles are calculated based on the annotations.

## Machine Design

Machine design consists of 2 stages of AI-based systems. The goal of the first stage of the AI application is to derive the following features from an image: individual lengths of each femur, tibia, and each leg, the mechanical axis (MA) angle, leg discrepancy, and pelvic tilt. The goal of the second stage is to detect the presence of surgical hardware at each of the anatomic landmarks. The overall architecture of the system is diagrammed in Fig. 3.

## Data Annotation

For the localization system in the first stage of the machine, a system of point annotations or keypoints is used. The keypoints identified included: femoral head apex, femoral head center, ischium, intercondylar notch, tibial spine, and ankle mortise. Up to 12 total keypoints are defined on each image, as each lower extremity will have up to these 6 points, depending on the field of view of the image and the present anatomy. Figure 2a shows an example of annotation image. The annotation software used for training and expert validation was the VGG Image Annotator (VIA) [14]. Each image was annotated once. Three student annotators (N.L., C.N., A.K.) performed preliminary annotations on the training and validation set, each of which were reviewed by a fellowship-trained musculoskeletal radiologist with over 10 years of experience (B.D). Point annotations were then converted to bounding box annotations by defining a square bounding box with center at the annotated point and size scaled to a proportion of the mean image dimension. This scaling coefficient was manually chosen for the entire dataset to scale the box to approximately the size of the femoral head. Within the training and internal validation set, 1270 femoral apex, 1271 femoral head center, 1263 ischium, 1269 intercondylar notch, 1260 tibial spine, and 1258 ankle mortise annotations were performed for a total of 7591 total annotations.

**Table 1** Demographic data of both training and testing sets

| | Training/validation set | Test set |
|---|---|---|
| N | 643 | 220 |
| Mean age (SD, range) | 61.4 years (15.8 years, 16–96) | 60.7 years (16.2 years, 22–95) |
| Sex | 345 M (54%) / 298 (46%) F | 125 M (57%) / 95 F (43%) |
| Hip arthroplasty | 41 (6.4%) | 17 (7.7%) |
| Other hip hardware | 31 (4.7%) | 10 (4.5%) |
| Knee arthroplasty | 104 (16%) | 67 (30%) |
| Other knee hardware | 91 (14%) | 23 (10%) |
| Ankle hardware | 24 (3.7%) | 11 (5.0%) |
| Amputation/incomplete anatomy | 39 (6.1%) | 12 (5.4%) |

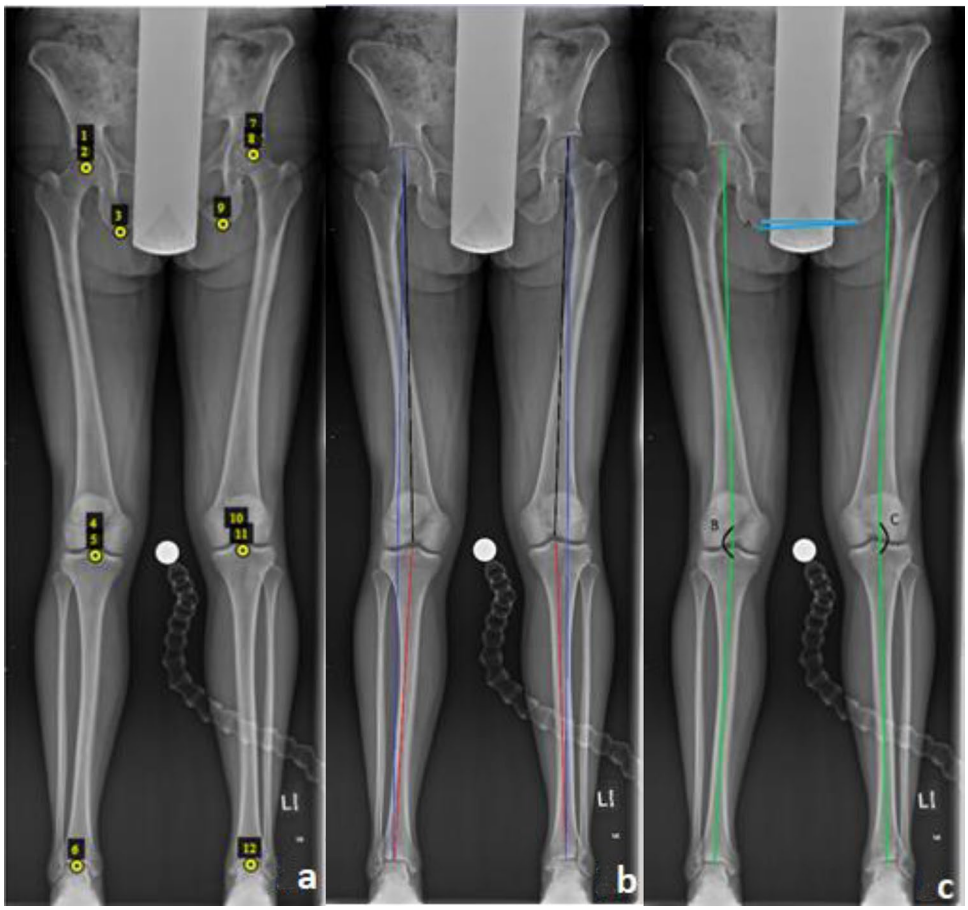



**Fig. 2** Defined landmarks and calculation of each metric. **a** depicts the twelve anatomic landmarks used for the measurements. **b** depicts the measurement of distances. Black represents femur length, blue represents leg length, and red represents tibia length. **c** depicts the measure of the angles. Angle A (blue) represents the measure of the pelvic tilt compared to a horizontal line (black). Angles B and C (green) represent the mechanical axis angle for each lower extremity

To validate the machine, 3 fellowship-trained musculoskeletal radiologists were asked to individually annotate the test set of 220 images. Individual annotations were aggregated to form a consensus gold standard set of annotations. For each point labeled by at least two radiologists, coordinates for each annotation were averaged and taken to represent the gold standard. Distance and angular measurements, which require data from multiple points, were considered to be defined if coordinates for all the requisite points were present. If one or more of the required points

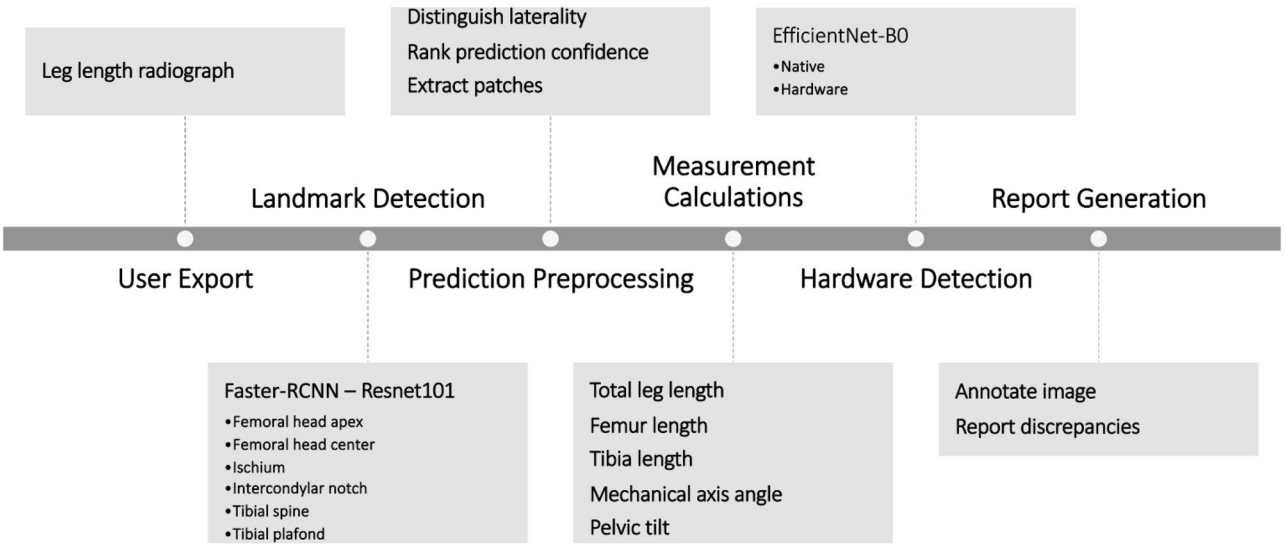



**Fig. 3** Data flow of the final AI application

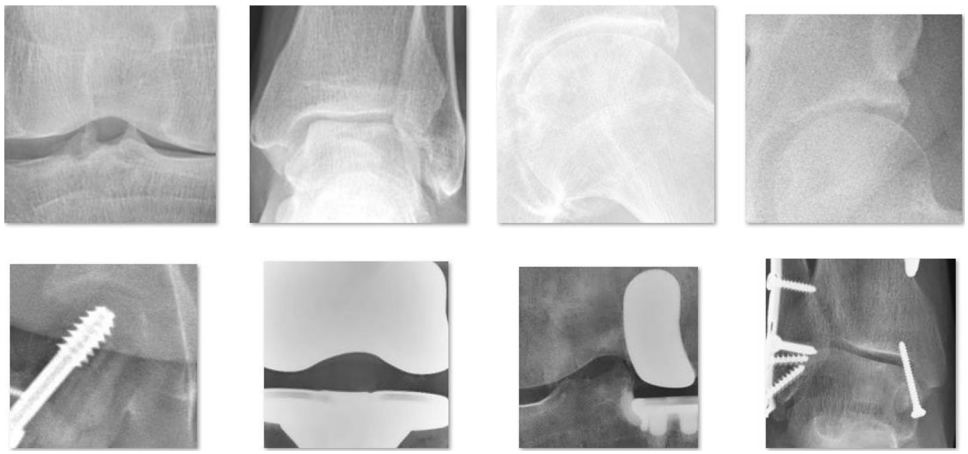


**Fig. 4** Example image patches used to train the hardware classifier model. Images were classified as native or hardware and used to train an EfficientNet-B0 model. The top row shows sample native image patches. The bottom row shows sample hardware image patches

was unlabeled in the aggregate set, such as in the case of an amputation, the measurement was considered not applicable.

For the second-stage classification step, point annotations in the first stage were used to extract a corresponding patch around the annotation for each image within the training and validation sets. Specifically, image patches corresponding to the femoral head apex, femoral head center, intercondylar notch, tibial spine, and ankle mortise annotations were extracted. Each patch was twice the size of the bounding box predicted by the first-stage system, which resulted in a small image that still generally contained the entire joint. Hardware descriptions are classified into two categories: native and metal. Figure 4 shows examples of each category.

Each image patch was categorized according to these labels, and annotations were reviewed by a fellowship trained musculoskeletal radiologist. The patches extracted from the images in the original training set were used for the classifier training set, and the patches extracted from the images in the original validation set were used for the

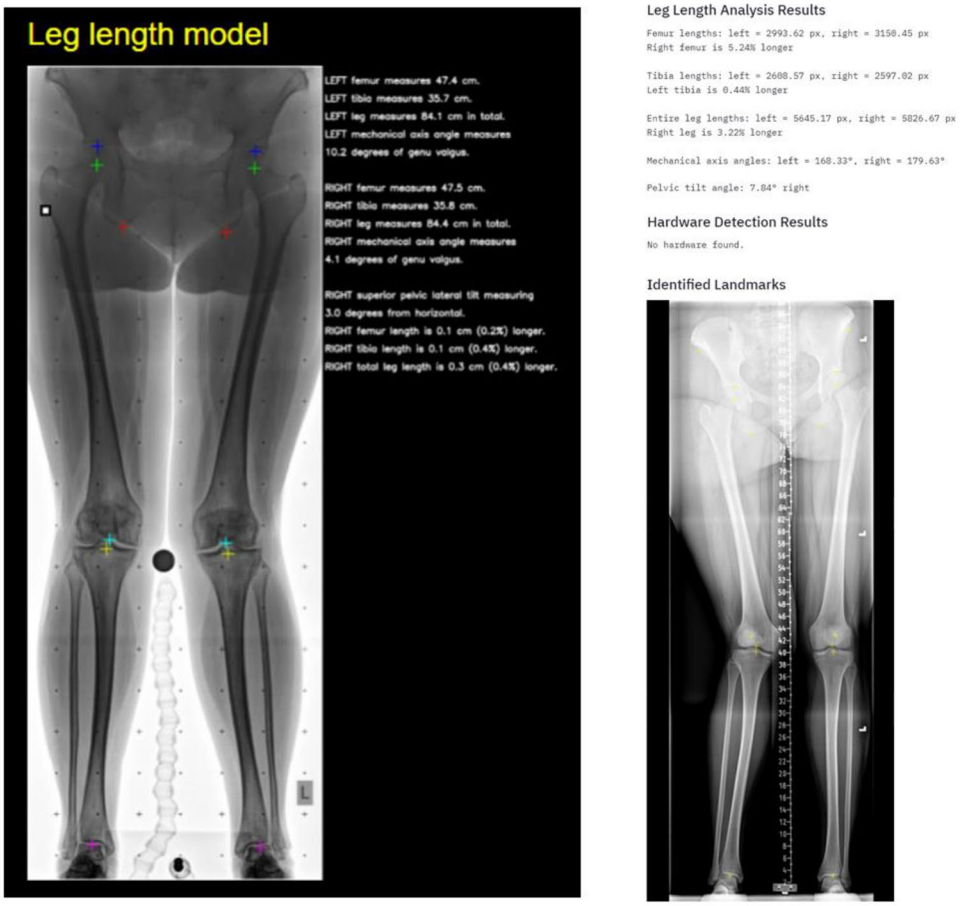



**Fig. 5** Current version of the clinical application (left). Current version of the webapp (right). Output format is subject to change

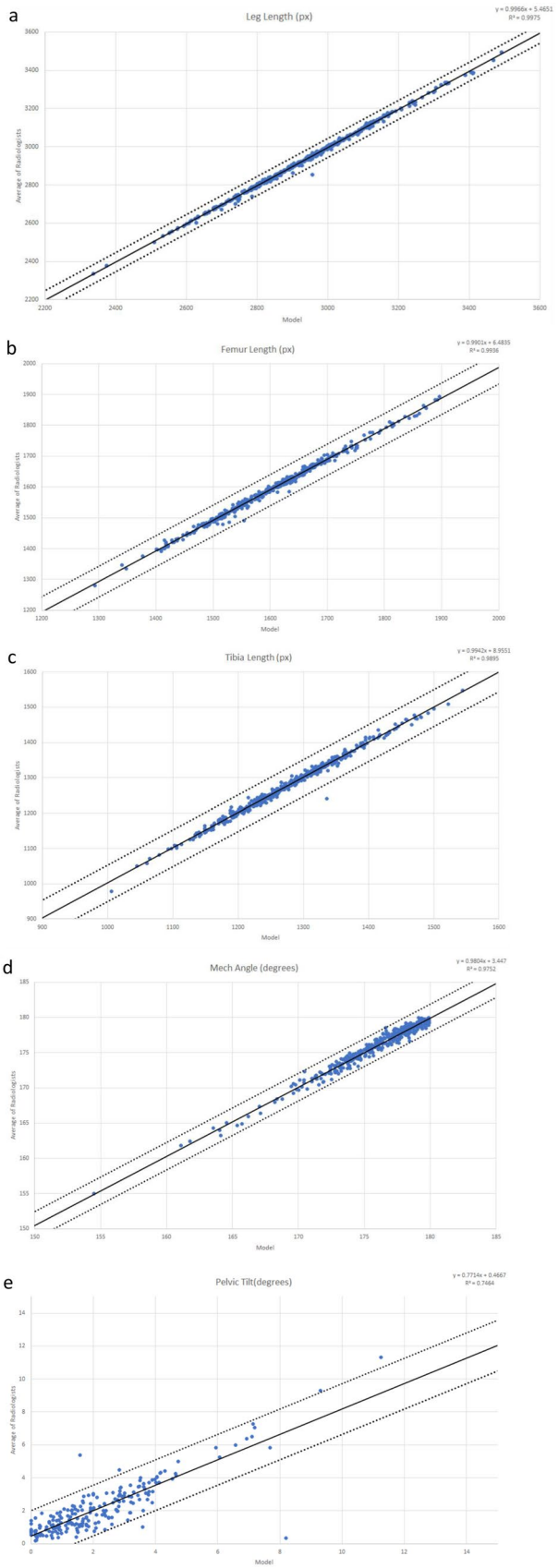



**Fig. 6** Scatterplot of leg (**a**), femur (**b**), and tibia lengths (**c**) with parallel lines of +—50 pixels (about 1 cm) Scatterplot of mechanical angle (**d**) and pelvic tilt (**e**) with parallel lines of +—2 degrees

classifier leave-out test set. The training set consisted of 7007 native and 635 metal image patches. The leave-out test set consisted of 1438 native and 72 metal image patches.

## Model Selection and Training

We selected a Faster-RCNN—ResNet 101 model from the open-source TensorFlow Object detection library with initial weights from pretraining on the Common Objects in Context (COCO) dataset [15]. Faster-RCNN—ResNet 101 has been previously used in the literature on radiological images to characterize knee osteoarthritis on radiographs [16]. Training was performed on a workstation utilizing a NVIDIA P100 graphics processing unit, using a TensorFlow 1.15 framework in Python 3.8.5. Image augmentation at the time of training was performed with random horizontal flips. No other pre-processing transformations or normalizations performed on the JPEG images before training. Training was performed with a fixed learning rate of 0.0003, until the intersection over union validation loss visually plateaued, at 64,982 steps.

A pretrained EfficientNet-B0 model was selected to perform the hardware image classification task [17]. EfficientNets are a state-of-the-art network that have been previously utilized in the medical imaging literature to diagnose COVID-19 on radiographs [18], identify diabetic retinopathy [19], and identify osteoporosis on hip radiographs [20]. Pretrained weights from the ImageNet dataset classification task, further optimized using the NoisyStudent training algorithm, were utilized to initiate model weights and improve convergence [21]. Model weights and architecture were obtained from an open-source library [22] based on the PyTorch 1.8 framework. This was implemented into a custom training and inference routine utilizing the FastAI package in Python 3.8.5 [23]. Training was performed on a separate workstation utilizing a NVIDIA P100 graphics processing unit. Staggered training was performed, initially only on the final layer and then on the entire model, for 5 and 30 epochs, respectively, for a total of 35 epochs, at which

**Table 2** Object detection confusion matrix

| | Actual | | |
|---|---|---|---|
| | | Landmarks | Other |
| Predicted | Detected | 2446 | 101 |
| | Missed | 62 | - |

point validation loss plateaued. A cyclically variable learning rate peaking at 0.001 was used for each stage of training.

### System Integration

A minimum confidence score threshold for detection of 0.9 was selected. In the case of multiple model predictions for landmark localization, an automated logic system processing the raw detections was implemented, where the landmark with the highest confidence was selected. The system split the image into left and right halves, and, for each half, the highest confidence bounding box was selected for each of the labeled landmarks. The center of each bounding box was selected as the point coordinate from which further measurements were taken, as per the definitions previously outlined.

The intent of the study is not only to train a deep learning model to measure these angles and lengths, but to create a usable system for deployment. A simple-to-use web application has been designed and deployed to process provided images and outputs the AI models as a complete preliminary dictation with annotated image. Figure 5 shows a sample screen capture of the web application.

### Analysis

The performance of each model was tested at multiple levels. For the initial localizing object detection model, performance metrics of precision, recall, and F1 score were assessed for each annotated point as compared to the gold standard radiologist aggregate. Intersection over union (IoU) was selected as the metric for evaluation of detection, as is conventional for the object detection nature machine learning model utilized. A detection was considered a true positive if the IoU of the predicted bounding box with the corresponding gold standard annotation was > 50%. A detection was considered a false positive if the IoU of the predicted box and the gold standard annotation was < 50%. A detection was considered a false negative if there was no predicted bounding box to correspond to the gold standard annotation.

Measurements of the total leg length, femoral length, tibial length, mechanical axis angle, and pelvic tilt angle were computed as previously defined from both the predicted landmarks and the radiologist annotations. Statistical measures of mean absolute difference, mean relative difference, standard deviation, Pearson correlation coefficient, and intraclass correlation coefficient between the predicted values and the gold standard values were computed based on a single rater, absolute agreement, 2-way random effects model. A significance level of 0.05 was chosen.

Finally, the performance of the hardware classification stage was characterized by sensitivity, specificity, accuracy, and F1 score.

### Deployment

The AI system incorporated into a web application allows a radiologist to input a single image and generates an annotated report of the leg with the following measurements, if applicable: bilateral leg length, bilateral femur length, bilateral tibial length, leg length discrepancy, bilateral mechanical axis angle, and pelvic tilt angle.

The application is built on Streamlit, a Python web framework designed specifically for data science and presentation. The framework handles routing and styling automatically, allowing for fast and simple implementation of data processing algorithms. The web application automatically detects when a valid image has been uploaded via the Upload Image button and then triggers the beginning of the data processing algorithm. After the image is uploaded, our system performs inference on the image to localize the landmarks of the lower extremity. The locations of the objects' centers are then used to calculate length and angle measurements for the analysis of the leg image. Image patches are extracted based on the detected bounding boxes, which are then analyzed using the hardware classifier model to report hardware at each defined landmark. The script creates two Python objects—one which represents length and angles, and one which represents detected hardware. This approach allows the data to be displayed in any manner and could even be input directly into clinical software if desired.

## Results

### Model Performance

A total of 2508 coordinates were considered labeled within the gold standard set after aggregation of the 3 radiologist annotations across the 220 image test set. At the basic object detection level, the machine correctly localized 2446 points (98%), with 62 points not detected and 101 points spuriously detected. Table 2 shows the confusion matrix at the object detection level. This corresponds to a recall of 0.975, precision of 0.960, and F1 score of 0.968.

**Table 3** Comparison of A.I. with ground truth (radiologist average)

| | Absolute Mean error | Absolute mean Error | Standard deviation |
|---|---|---|---|
| Leg length | 6.37 pixels | 0.22% | 8.93 pixels |
| Femur length | 9.97 pixels | 0.64% | 7.89 pixels |
| Tibia length | 5.94 pixels | 0.47% | 8.61 pixels |
| Mech angle | 0.42° | | 0.54° |
| Pelvic tilt | 0.67° | | 1.74° |

**Table 4** Leg length Pearson correlation coefficient (r value)

| | Radiologist 2 | Radiologist 3 | A.I |
|---|---|---|---|
| Radiologist 1 | 0.999 ($p < 0.00001$) | 0.999 ($p < 0.00001$) | 0.998 ($p < 0.00001$) |
| Radiologist 2 | NA | 0.999 ($p < 0.00001$) | 0.999 ($p < 0.00001$) |
| Radiologist 3 | 0.999 ($p < 0.00001$) | NA | 0.999 ($p < 0.00001$) |

The measurements computed from radiologists' annotations were compared against the system-predicted measurements for each of leg length, tibia length, femur length, M.A. angle, and pelvic tilt. In 0.1% of measurements, the A.I. model output a measured value, and none of the radiologists measured that same value. In 1.3% of measurements, the A.I. model did not output a value, and all 3 of the radiologists output said measurement. The remaining 98.6% of measurements are shown in Fig. 6a-e, scatterplots showing predicted measurements versus ground truth. Each predicted measurement was compared to the ground truth to obtain the absolute mean errors and standard deviations as detailed in Table 3, showing closely approximated length measurements on average < 1% difference from radiologist measurements and closely approximated angle measurements on average < 1° from radiologist measurements.

Pearson correlation coefficients of the leg length and mechanical axis angle computed among the AI model and of the radiologists demonstrate statistically significant correlations, as detailed in Tables 4 and 5, respectively. Table 6 details Pearson correlation coefficients between the AI system and the radiologist gold standard for each of total leg length, femur length, tibia length, mechanical axis angle, and pelvic tilt angle, which demonstrate statistically significant correlations.

Intraclass correlation coefficients comparing the AI system-generated measurements versus the gold standard aggregate measurements and individual radiologist measurements can be seen in Table 7. Excellent agreement is seen between the AI system and radiologists for leg length, femur length, tibia length, and mechanical axis. Good agreement is seen between the AI system and radiologists for pelvic tilt.

**Table 5** Mechanical axis angle Pearson correlation coefficient (r value)

| | Radiologist 2 | Radiologist 3 | A.I |
|---|---|---|---|
| Radiologist 1 | 0.983 ($p < 0.00001$) | 0.986 ($p < 0.00001$) | 0.978 ($p < 0.00001$) |
| Radiologist 2 | NA | 0.988 ($p < 0.00001$) | 0.989 ($p < 0.00001$) |
| Radiologist 3 | 0.988 ($p < 0.00001$) | NA | 0.983 ($p < 0.00001$) |

**Table 6** Machine vs gold standard Pearson correlation coefficients

| Measurement | R value (A.I. vs. radiologist average) |
|---|---|
| Leg length | 0.999 ($p < 0.00001$) |
| Femur length | 0.997 ($p < 0.00001$) |
| Tibia length | 0.995 ($p < 0.00001$) |
| Mech angle | 0.988 ($p < 0.00001$) |
| Pelvic tilt | 0.864 ($p < 0.00001$) |

Figure 7 shows sample output from the machine compared against the gold standard. Figure 7a is a sample image with misplaced objects. The model detects two different locations for the right ankle mortise. These cases are rare across our dataset and further reduced in number with our automated point selection algorithm but can cause the output to export a measurement that is inaccurate. Careful review of the output image for any errors such as these before submitting a generated report would be necessary in a clinical environment.

### Classification Model

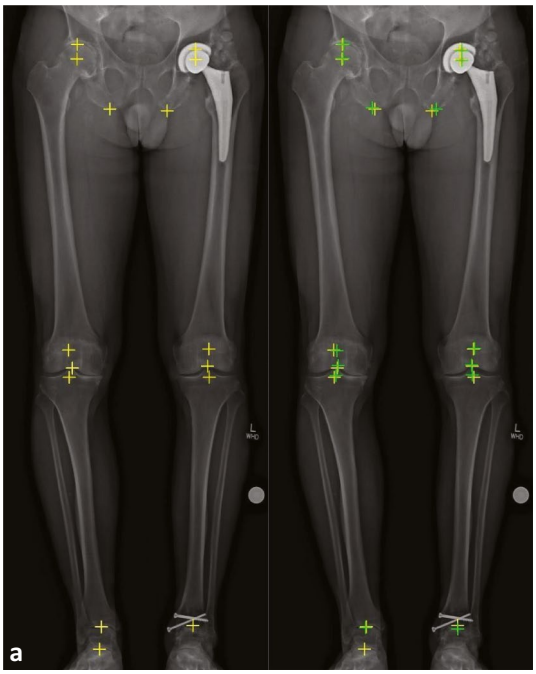
The hardware detection stage of the machine demonstrated a sensitivity of 95.8%, specificity of 100%, accuracy of 99.8%, and a F1 score of 97.9% as detailed in Table 8. The 3 misclassified patches are shown in Fig. 8.

## Discussion

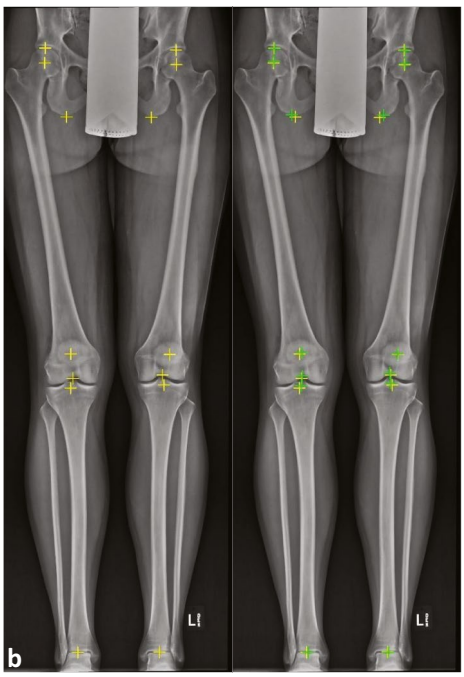
We have built an AI system that automatically identifies anatomic landmarks of a bilateral lower extremity radiograph, quantifies bilateral leg lengths and associated discrepancy, and characterizes the length of the femur and tibia along with computing mechanical axis angles and pelvic tilt.

**Table 7** Intraclass correlation coefficients (ICC): high ICC among the measurements of leg length, femur length, tibia length, mechanical axis angle, and pelvic tilt indicates a high level of agreement between radiologist measurements and machine predicted measurements. For reference, values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively. [24]

| | VS aggregate (CI; p value) | VS individual radiologists (CI; p value) |
|---|---|---|
| Leg length | 1.00 (1.00–1.00; 0) | 1.00 (1.00–1.00; 0) |
| Femur length | 0.99 (0.94–1.00; 0) | 0.98 (0.95–0.99; 0) |
| Tibia length | 1.00 (0.99–1.00; 0) | 0.99 (0.98–0.99; 0) |
| Mech axis | 0.98 (0.98–0.99; 0) | 0.99 (0.99–0.99; 0) |
| Pelvic tilt | 0.86 (0.83–0.89; < 0.00001) | 0.8 (0.77–0.83; < 0.00001) |




| | Radiologist Average (R/L) | A.I. Model (R/L) |
|---|---|---|
| Leg Length (Pixels) | 2853/2826 | 2956/2816 |
| Femur Length (Pixels) | 1570/1537 | 1581/1550 |
| Tibia Length (Pixels) | 1240/1243 | 1336/1216 |
| Mechanical Angle (Degrees) | 172.1/174.9 | 172.3/173.9 |
| Pelvic Tilt (Degrees) | 1.05 | 1.85 |




| | Radiologist Average (R/L) | A.I. Model (R/L) |
|---|---|---|
| Leg Length (Pixels) | 3159/3146 | 3161/3146 |
| Femur Length (Pixels) | 1717/1687 | 1731/1699 |
| Tibia Length (Pixels) | 1392/1407 | 1382/1395 |
| Mechanical Angle (Degrees) | 177.0/178.0 | 177.3/178.9 |
| Pelvic Tilt (Degrees) | 0.52 | 0.39 |

**Fig. 7** **a** A case where the object detection model detected two ankle mortises on the same leg, which leads to an error in the leg length and tibia length measurements. Points identified by our model are labeled in yellow; the average point identified by radiologists is labeled in green. **b** A sample properly detected image with points very similar to that of the radiologists. Points identified by our model are labeled in yellow; the average point identified by radiologists is labeled in green

**Table 8** Confusion matrix for machine predictions of hardware presence versus native anatomy

| | | Actual | |
|---|---|---|---|
| | | Hardware | Native |
| Predicted | Hardware | 69 | 0 |
| | Native | 3 | 1438 |

Validation results show robust performance at the detection level, demonstrating exceptional agreement with the gold standard of aggregated radiologist measurements. Performance on pelvic tilt prediction is noted to demonstrate slightly less agreement, and it is suspected that this is due to the fact that the pelvic tilt measurement is consistently very close to zero in the test set. Therefore, small variations of a few pixels in the points that the system selects will result in a large percent relative error and weaker correlation. Furthermore, we note greater subjectivity between radiologists in measuring pelvic tilt, as can be seen in Table 7, where the ICC drops substantially when radiologist measurements as individual points are used to compute correlation, relative to when the single aggregate measurement is used. Nevertheless, system predicted pelvic tilt remains largely within 1 degree of radiologist measured values on average, which remains impressive.

Prior work has attempted to use AI to characterize leg length radiographs, and we improve upon this by reducing our exclusion criteria, capturing a larger training set, and using a novel annotation and machine learning architecture to improve efficiency. Zheng et al. [25] used a segmentation-based approach that generated strong correlation with radiologist measurements. However, this study was limited to children and radiographs without hardware, had a total training/validation dataset of 179 patients, and only attempted to quantify lengths. Our training set included radiographs with arthroplasty present, allowing the system to identify measurements even when encountering hip, knee, or ankle replacements. Our system achieves a comparable level of performance on patients greater than 16 years without such limitations. Schock et al. [8] also use a segmentation-based approach to quantify mechanical axis angles but exclude single-leg view examinations and trained on a combined training/validation set of 149 patients. By virtue of the object detection-based algorithm, our system is able to handle incomplete anatomy and amputations, as the model utilized independently detects all visible landmarks. There was no discernable drop in accuracy for images with missing anatomy included in our test set. By detecting the same landmarks a radiologist recognizes, our system is able to achieve a higher level of performance and radiologist agreement—for instance, we achieve ICC for mechanical axis angles of 0.99, compared to reported ICC of 0.86–0.89.

The landmark detection approach also has significant advantages over the segmentation-based approach utilized in these prior works. As landmarks are used by radiologists to perform measurements, this is an intuitive approach that is easy to troubleshoot and does not rely on complex geometric mapping and manipulation to retroactively estimate landmark coordinates. This also results in a more robust system, as our calculations are not sensitive to subtle perturbations in the contour or alignment of the anatomy, particularly in the cases of more extensive deformity. For instance, no subjectively apparent discrepancy was noticed with the performance of the model on patients with severe osteoarthrosis. Importantly, landmark annotation is a simple annotation task, requiring an order of magnitude or two less time to annotate compared to the laborious task of segmentation. This allows for expansion of the size of our training set to be several times the size of pre-existing work, which theoretically allows for more robust machine training and better generalizability.

The addition of automatic hardware detection and localization adds a final layer to the system that improves its clinical usefulness. Although it may seem straightforward for a radiologist to identify the presence of hardware, automation makes initial reports generated by the system more useful by identifying additional pertinent information. Additionally, these data help to identify situations when the accuracy may be limited, as the array of orthopedic hardware is vast and diverse in function and appearance which can potentially affect detection of landmarks if the hardware is extensive or unique. This is particularly useful in light of the web application, which allows not only radiologists

**Fig. 8** Misclassified patches from the hardware detection stage of the system. These patches were mistakenly classified as native anatomy by the machine. No patches with native anatomy were misclassified as having hardware

but providers from other specialties to upload radiographs into the web application for a fast preliminary assessment to determine on-the-spot functional or operative implications. This automated data analysis is also much faster than the laborious, time-consuming process of radiologists analyzing each data point. Finally, by reducing the manual effort needed to include these metrics in reporting, radiologists are more likely to include this information, thereby improving the quality and usefulness of radiology reports while simultaneously reducing a source of intra- and inter-reader variability.

### Limitations & Future Work

Several limitations are acknowledged and enhancements for future work with the system are presented. As with other deep learning-based AI projects, additional data from other sites would serve to improve generalizability of the system and potentially improve accuracy. As only radiographs of patients 16 or older were used for training, the application has not been validated on and may not be reliable for radiographs of children. The efficiency of producing landmark keypoint annotations utilized in this project somewhat mitigates this limitation, as it becomes substantially less time-intensive to annotate new data to retrain model weights for a new application site in the case of a production model.

Second, as the field of computer vision continues to evolve at a rapid pace, there may be new different models that are more suited to the task of keypoint detection compared to the object detection models that are used in the current system. Although performance with the selected models is highly accurate, future work should evaluate multiple models of various architectures to select the optimal performing model to improve machine accuracy.

Third, our web application can could be refined to achieve faster performance using GPU inference, utilize session states to save program progress, accept multiple images, and provide easy translation from the output into a reporting application. Ultimately, however, the web application is designed to showcase the potential of the model rather than provide a production-level service, as such services will be of most use with integration into the clinical workflow. Further improvements that could improve workflow would include decision systems to convert pixels to actual units of measurement, either by accessing the pixel spacing DICOM metadata field or, perhaps more accurately, identifying and adjusting based on technologist placed markers, such as rulers or magnification marker balls.

As such, a translational facet of future work would be to implement this system in a clinical workflow with results tracking and feedback to assess the net impact such a system might have on metrics such as radiologist productivity and need for report revision. Application of our system within a clinical environment could enable retrospective analysis of archived examinations and produce population-level analyses, yielding nuanced demographic data and population norms on a scale that has never before been practical to achieve. This could help us better identify subtle pathologies, more precise normal ranges, and ultimately achieve a level of personalized medicine never before possible.

The principles presented in this study are not only applicable to leg length and angle measurements, but also to nearly any measurement of the human skeleton. Ours is a flexible approach that can easily and rapidly be adapted to specific clinical questions compared to the published standard of segmentation-based annotation. We plan to continue our work building applications that assist radiologists with automated image measurements and with the aim of using AI to increase efficiency and productivity in the radiology workflow.

## Conclusion

In summary, we have constructed a novel, efficient, and accurate means of automatically quantifying leg lengths/angles and detecting hardware from bilateral lower extremity radiographs using a keypoint-based artificial intelligence system. Automatic quantitative analysis and reporting of radiographic leg length examinations using our system is feasible and not only has great potential to enhance radiologist workflow/efficiency, but may also have profound implications on the field of orthopedics as this conceptual approach can be extrapolated to nearly any measurement of the human skeleton.

**Author contributions** Authors N.L., C.N., B.D., R.B., M.F., J.P., A.H., L.D., and C.F. discussed study design. Authors B.D. and J.P. performed data collection. Authors N.L., C.N., A.K., A.L., E.B., K.S., J.P., and C.F. performed data annotation. Authors N.L., B.D., A.K., A.L., and C.F. devised and implemented the artificial intelligence model. Authors N.L. and C.F. performed data analysis. Authors B.D., R.B., M.F., E.B., and K.S. provided expert clinical guidance and support. All authors contributed to the manuscript.

**Funding** This study was not funded by a research grant.

**Availability of Data and Material** Sample data available upon request.

**Code Availability** Our end-to-end AI system to generate radiology reports from X-ray images is freely available at: http://xrayhead.com:8090/

### Declarations

**Ethics Approval** This research study (IRB-57973) was conducted retrospectively from data obtained for clinical purposes. An IRB official waiver of ethical approval was granted by the Stanford University IRB.

**Consent to Participate** Waived, as only images were obtained for this HIPAA compliant retrospective review.

**Consent for Publication** Waived, as only images were obtained for this HIPAA compliant retrospective review.

**Conflict of Interests/Competing Interests** The authors declare that they have no conflicts of interest.

## References


1. Khamis, S. and E. Carmeli, *Relationship and significance of gait deviations associated with limb length discrepancy: a systematic review.* Gait & posture, 2017. **57**: p. 115-123.
2. Raczkowski, J.W., B. Daniszewska, and K. Zolynski, *Functional scoliosis caused by leg length discrepancy.* Archives of medical science: AMS, 2010. **6**(3): p. 393.
3. Röder, C., et al., *Total hip arthroplasty: leg length inequality impairs functional outcomes and patient satisfaction.* BMC musculoskeletal disorders, 2012. **13**(1): p. 1-8.
4. Vaidya, S.V., et al., *Total knee arthroplasty: limb length discrepancy and functional outcome.* Indian journal of orthopaedics, 2010. **44**(3): p. 300-307.
5. Moschella, D., et al., *Wear patterns on tibial plateau from varus osteoarthritic knees.* Clinical Biomechanics, 2006. **21**(2): p. 152-158.
6. Sykes, A., et al., *Patients' perception of leg length discrepancy post total hip arthroplasty.* Hip Int, 2015. **25**(5): p. 452-6.
7. Kim, S.H., et al., *Reliability and validity of the femorotibial mechanical axis angle in primary total knee arthroplasty: navigation versus weight bearing or supine whole leg radiographs.* Knee surgery & related research, 2018. **30**(4): p. 326.
8. Schock, J., et al., *Automated Analysis of Alignment in Long-Leg Radiographs by Using a Fully Automated Support System Based on Artificial Intelligence.* Radiology: Artificial Intelligence, 2020. **3**(2): p. e200198.
9. Hau, M.Y.T., et al., *Two-dimensional/three-dimensional EOS™ imaging is reliable and comparable to traditional X-ray imaging assessment of knee osteoarthritis aiding surgical management.* The Knee, 2020. **27**(3): p. 970-979.
10. Sabharwal, S. and A. Kumar, *Methods for assessing leg length discrepancy.* Clinical orthopaedics and related research, 2008. **466**(12): p. 2910-2922.
11. Moreland, J.R., L. Bassett, and G. Hanker, *Radiographic analysis of the axial alignment of the lower extremity.* The Journal of bone and joint surgery. American volume, 1987. **69**(5): p. 745–749.
12. Tyrakowski, M., H. Yu, and K. Siemionow, *Pelvic incidence and pelvic tilt measurements using femoral heads or acetabular domes to identify centers of the hips: comparison of two methods.* European Spine Journal, 2015. **24**(6): p. 1259-1264.
13. Tannast, M., et al., *Estimation of pelvic tilt on anteroposterior X-rays—a comparison of six parameters.* Skeletal radiology, 2006. **35**(3): p. 149-155.
14. Dutta, A., A. Gupta, and A. Zissermann, *VGG image annotator (VIA).* URL: http://www.robots. ox. ac. uk/~ vgg/software/via, 2016.
15. *Faster-RCNN ResNet 101 Coco Config*. 2018, GitHub.
16. Liu, B., J. Luo, and H. Huang, *Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN.* International journal of computer assisted radiology and surgery, 2020. **15**(3): p. 457-466.
17. Tan, M. and Q.V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* arXiv e-prints, 2019: p. arXiv:1905.11946.
18. Marques, G., D. Agarwal, and I. de la Torre Díez, *Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network.* Applied soft computing, 2020. **96**: p. 106691.
19. Chetoui, M. and M.A. Akhloufi. *Explainable Diabetic Retinopathy using EfficientNET*. in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2020. IEEE.
20. Yamamoto, N., et al., *Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates.* Biomolecules, 2020. **10**(11): p. 1534.
21. Xie, Q., et al., *Self-training with Noisy Student improves ImageNet classification. arXiv e-prints, art.* arXiv preprint arXiv: 1911.04252, 2019.
22. Wightman, R., *PyTorch Image Models*. 2019, GitHub.
23. Howard, J. and R. Thomas, *fast. ai-Making neural networks uncool again*.
24. Koo, T.K. and M.Y. Li, *A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research.* Journal of Chiropractic Medicine, 2016. **15**(2): p. 155-163.
25. Zheng, Q., et al., *Deep Learning Measurement of Leg Length Discrepancy in Children Based on Radiographs.* Radiology, 2020. **296**(1): p. 152-158.