ASME Journal of Engineering and Science in Medical Diagnostics and Therapy
Online journal at:
https://asmedigitalcollection.asme.org/medicaldiagnostics

Sheridan Perry
Department of Mechanical Engineering,
Embry-Riddle Aeronautical University,
1 Aerospace Blvd.,
Daytona Beach, FL 32114
e-mail: perrys8@my.erau.edu

Matthew Folkman
Pediatric Orthopedics,
Rainbow Babies and Children's Hospital,
2101 Adelbert Road,
Cleveland, OH 44106
e-mail: matthew.folkman@uhhospitals.org

Takara O'Brien
Department of Aerospace Physiology,
Embry-Riddle Aeronautical University,
1 Aerospace Blvd.,
Daytona Beach, FL 32114
e-mail: obrient8@my.erau.edu

Lauren A. Wilson
Department of Aerospace Physiology,
Embry-Riddle Aeronautical University,
1 Aerospace Blvd.,
Daytona Beach, FL 32114
e-mail: wilsol29@my.erau.edu

Eric Coyle
Department of Mechanical Engineering,
Embry-Riddle Aeronautical University,
1 Aerospace Blvd.,
Daytona Beach, FL 32114
e-mail: coylee1@erau.edu

Raymond W. Liu
Pediatric Orthopedics,
Rainbow Babies and Children's Hospital,
2101 Adelbert Road,
Cleveland, OH 44106
e-mail: raymond.liu@uhhospitals.org

Charles T. Price
International Hip Dysplasia Institute,
Orlando Health,
3160 Southgate Commerce Blvd.,
Orlando, FL 32806
e-mail: charles.price@orlandohealth.com

Unaligned Hip Radiograph Assessment Utilizing Convolutional Neural Networks for the Assessment of Developmental Dysplasia of the Hip[1]

Developmental dysplasia of the hip (DDH) is a condition in which the acetabular socket inadequately contains the femoral head (FH). If left untreated, DDH can result in degenerative changes in the hip joint. Several imaging techniques are used for DDH assessment. In radiographs, the acetabular index (ACIN), center-edge angle, Sharp's angle (SA), and migration percentage (MP) metrics are used to assess DDH. Determining these metrics is time-consuming and repetitive. This study uses a convolutional neural network (CNN) to identify radiographic measurements and improve traditional methods of identifying DDH. The dataset consisted of 60 subject radiographs rotated along the craniocaudal and mediolateral axes 25 times, generating 1500 images. A CNN detection algorithm was used to identify key radiographic metrics for the diagnosis of DDH. The algorithm was able to detect the metrics with reasonable accuracy in comparison to the manually computed metrics. The CNN performed well on images with high contrast margins between bone and soft tissues. In comparison, the CNN was not able to identify some critical points for metric calculation on a few images that had poor definition due to low contrast between bone and soft tissues. This study shows that CNNs can efficiently measure clinical parameters to assess DDH on radiographs with high contrast margins between bone and soft tissues with purposeful rotation away from an ideal image. Results from this study could help inform and broaden the existing bank of information on using CNNs for radiographic measurement and medical condition prediction. [DOI: 10.1115/1.4064988]

[1]Paper was presented at the International Mechanical Engineering Congress & Exposition®, New Orleans, LA, Ernest N. Morial Convention Center, October 29–November 2, 2023. IMECE2023.

[2]Corresponding author.

Contributed by the Applied Mechanics Division Technical Committee on Dynamics & Control of Structures & Systems (AMD-DCSS) of ASME for publication in the Journal of Engineering and Science in Medical Diagnostics and Therapy. Manuscript received December 11, 2023; final manuscript received February 5, 2024; published online April 2, 2024. Editor: Ahmed Al-Jumaily.

Victor A. Huayamave[2]
Department of Mechanical Engineering,
Embry-Riddle Aeronautical University,
1 Aerospace Blvd.,
Daytona Beach, FL 32114
e-mail: huayamav@erau.edu

1 Introduction

Developmental dysplasia of the hip (DDH) is widely known to be the most common etiology for the development of osteoarthritis of the hip. DDH occurs when the ball-and-socket hip joint is underdeveloped, in which the acetabulum (socket) is too shallow for the ball (femoral head) to be secure in the joint. This can lead to subluxation and, in more severe cases, complete hip joint dislocation [1]. Additionally, extraneous tension of connective tissue and tendons surrounding the joint can lead to long-term overcompensation during dislocation [2]. Current diagnostic imaging options include radiographs, ultrasound, computed tomography (CT), and magnetic resonance imaging [3,4]. The use of imaging can be limited according to factors such as interobserver variability, false positives and negatives, and limited reproducibility in follow-up examinations [2–4].

Machine learning has become an increasingly viable means to reduce human error in DDH diagnosis. Machine learning detection algorithms such as You Only Look Once (YOLO) train a neural network and teach the algorithm to process data to make predictions inspired by human input [5]. This has the potential to limit user subjectivity, predict developmental gaps between age progression, and reduce false positives and negatives [6]. Several studies have used neural networks for DDH assessment in radiography [7–11]. One approach utilizes probability predictions through classic machine learning to identify DDH from two- and three-dimensional ultrasounds [12,13]. Other studies have shown improvements in efficiency by implementing a machine-learning network to assist in diagnosis [6,14,15]. However, these studies have not investigated the prediction accuracy on radiographs with poor definition due to low contrast between bone and soft tissues, nor have they used rotated pelvic images that were not perfectly aligned with anatomical planes.

Misaligned radiographs are not designed for the standard assessment metrics, as the metrics are designed to be computed on aligned images. Additionally, the computation is inherently limited because the pelvic structure is a complex three-dimensional shape being represented using a two-dimensional image slice. Reference lines that are meant to be approximately horizontal on an aligned image can be significantly altered or distorted in misaligned images. This can adversely affect the standard deviation and variance of the computed assessment metrics. Few studies analyze misaligned images, but those that do show significant increases in measurement variation in angles, such as lateral center edge angle and Sharp's angle [16]. This increase in variation can be in part attributed to the change of the reference lines. Additionally, these metrics can be further influenced by the obfuscation of overlapping features. Compounding these limiting factors by computing these metrics by hand is not ideal as it adds significant additional risk of intra- and interobserver variability factors. This makes automating the assessment metrics an ideal solution.

Radiographs are typically used for children older than 6 months, and the ACIN, MP, lateral center-edge angle (CEA), and SA are used to assess DDH. The ACIN measures the lateral coverage of the FH by the acetabulum [17]. The MP measures the displacement of the femoral head to the center of the acetabulum [18]. Conversely, both the SA and CEA represent the acetabulum's depth and capacity to cover the FH [19].

This study prioritized the accuracy of the CEA due to its ability to account for variations in the shape and size of the FH and acetabulum compared to the SA. The CEA has a higher rate of reproducibility and is less affected by variations in patient positioning compared to the SA [20]. CEA also possesses the capability to monitor hip plasticity in adaptive changes in the shape and position of the acetabulum with respect to the FH over time [20]. The CEA is a strong indicator for the assessment of DDH and thus is critical for a neural network to identify accurately. The other goal of the study was to determine how the neural network would handle rotated pelvic images that were not perfectly aligned with anatomical planes.

The detection of DDH using medical metrics can be straightforward for an experienced radiographer. However, DDH diagnosis is a time-consuming and repetitive process, which can be detrimental. The purpose of this study was to use a neural network to predict DDH metrics in radiographic images and address the limitations in DDH assessment, providing tools for practitioners by increasing the accuracy of DDH diagnosis.

2 Material and Methods

A set of de-identified CT scans were collected from 60 subjects for use in this Institutional Review Board (IRB) approved study by Rainbow Babies and Children's Hospital and Case Western Reserve University under study number 20211382. The subject set included 30 males and 30 females, with three subjects at each year of age ranging from 8 to 17 years. CT scans with fractures, hip dysplasia, retained hardware, and known intravenous or oral contrast studies were excluded. All CT scans contained healthy osseous structures without apparent deformity.

The 60-subject CT scans were first converted into three-dimensional (3D) models using the hospital's Picture Archiving and Communication System (PACS). These 3D models were subsequently converted to two-dimensional simulated radiographs by reducing the observable slab length to reflect the natural opacity and viewing dimensions of radiographic imaging. This process is displayed in Fig. 1.

The base position of each CT scan was set by aligning the superior aspect of the femoral heads, rotating the pelvis to display symmetric obturator foramina, and centralizing the tip of the coccyx between the pubic tubercles. The simulated radiographs were then manipulated in 3D space to predetermined set points. Each subject's pelvis was rotated along the mediolateral and craniocaudal axes in set increments. There were five specific increments along each axis, leading to a total of 25 pelvic images per subject, as shown in Fig. 2.

Around the craniocaudal axis, the five positions were the coccyx centered between the pubic tubercles, the coccyx rotated to the medial edge of the obturator foramen (each side), and the coccyx rotated midway between these two points (each side). Around the mediolateral axis, the five positions were the tip of the coccyx centered between the pubic tubercles, the superior and inferior pubic rami superimposed, the distance midway between these two points, the pubic tubercles in line with the sacrococcygeal line, and midway between this point and the centered pubic tubercle point.

Images were subsequently saved in each position. Upon collecting the 25 radiographs, each image was adjusted for uniform brightness and contrast, and a standard sharpness of 35% was applied to allow for visualization of radiographic landmarks used in measurements. Therefore, a total of 1500 images were available for training. Two hundred and fifty of the 1500 images were split off the

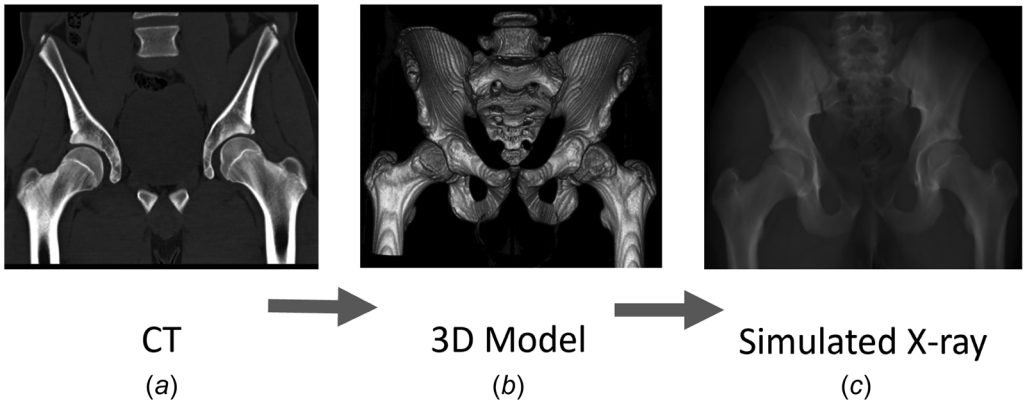


Fig. 1 Visualization of CT conversion process showing the 2-step process to convert CT scans to radiographs with (*a*) the original CT scan, (*b*) the converted 3D model, and (*c*) the final simulated radiograph

Fig. 2 A collage of images depicting the variation along craniocaudal and mediolateral axes for a single subject

set to function as testing samples to assess the trained algorithm. The remaining 1250 images were used to teach, validate, and test the algorithm. Before training the network, the images were preprocessed to optimize the data.

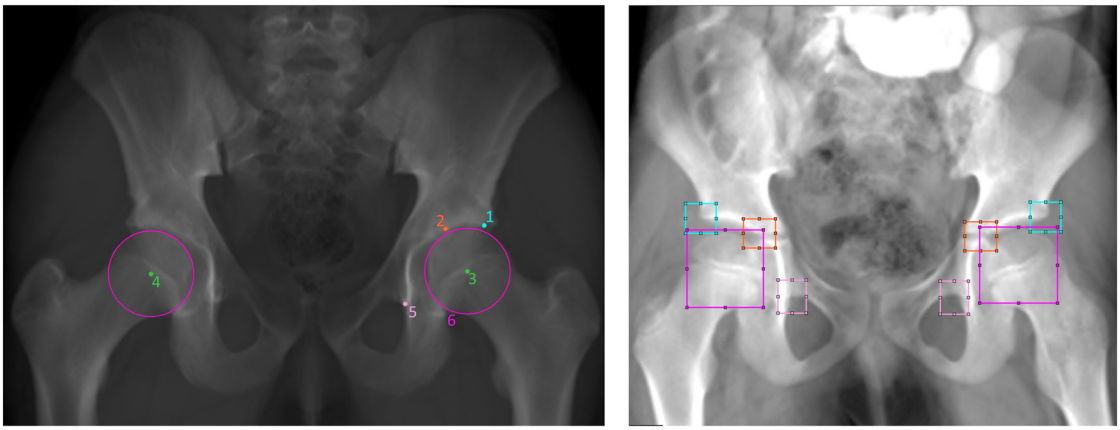
2.1 Preprocessing Images. The first step in this process was to resize the images since many of them were not uniformly sized. Note that this study used MATLAB (MathWorks, Natick, MA) to perform all programming and implement a neural network. The images were resized to JPEG images that were 1564 pixels wide by 940 pixels tall with a resolution of 96 dots per inch. Additionally, the photos were filled with extraneous data (e.g., patient information and camera zoom percentage) that was not a part of the hip and could, therefore, be removed based on the consistent location of this data within the image, as shown in Fig. 3. Although the removal process worked well for most of the images, a small number of processed images still had a nominal amount of text. This was due to initial inconsistent sizing, which, when resized, stretched the extraneous text to be inconsistent with the majority of the data samples. This nominal amount of text did not influence the algorithm prediction.

2.2 Labeling Procedure. The processed images were then labeled utilizing MATLAB R2022B's image labeling tool (image processing toolbox). The previously described metrics (ACIN, CEA, SA, MP) were analyzed, and unique locations were determined to be used to create the ground truth metrics [21]. These locations are the femoral head, the lateral acetabular roof, the triradiate cartilage, and the pelvic teardrop (Köhler teardrop) [22].

Example images of these points were created by medical specialists; the locations were labeled as 6, 1, 2, and 5, respectively, in Fig. 4(*a*). Note that 3 and 4 refer to the centers of the femoral head and will be computed from the label of 6. Fig. 4(*b*) describes the labels of the key locations replicated, and some of these locations are abbreviated as follows: Sourcil Sharps MP (SSMP), Sourcil Tönnis

(*a*) (*b*)

Fig. 3 Cropping of image artifacts (*a*) image prior to cropping and (*b*) image after cropping



(*a*) (*b*)

Fig. 4 Labeling process (*a*) depicts locations assessed by medical professionals and (*b*) depicts bounding boxes chosen by the engineering team (blue is SSMP, orange is STP2, purple is FH, pink is pelvic teardrop, and green is the center of the femoral head). (Color version online.)

P2 (STP2), and FH. The locations shown in Fig. 4(*b*) were labeled throughout the training set of 1250 images, with the labeling process being reviewed by medical experts. At the end of the procedure, the labeling was reviewed for consistency.

2.3 Algorithm Setup. The labels were exported and subsequently split into a training and testing set, further subdivided as shown in Fig. 5. Additionally, to reduce computational demand, the images and corresponding label coordinates were uniformly shrunk by a factor of 4 to 391×234. Tiny-yolov4-coco was the base network for the algorithm, which possesses 2 detection heads and is trained on the coco dataset. Csp-darknet53-coco was considered; however, despite being the generic base network for YOLOv4, the increase in computational cost coupled with marginal improvement in results prevented it from being used in this study [5]. Four anchor boxes were assigned per detection head. It is important to note that YOLOv4 requires images with pixel length and width that are multiples of 32. This necessitated a slight augmentation performed by a MATLAB transform function to increase the images and labels to the network input size of 416×256, which can increase errors in the network. This increase in image and label size was strictly for training, and the network fed the 391×234 images for assessment purposes. The data were augmented by flipping and randomly scaling the image to increase the available training data and improve algorithm accuracy. Color change augmentation was ignored due to the images being monochromatic. An example of augmented data is shown in Fig. 6.

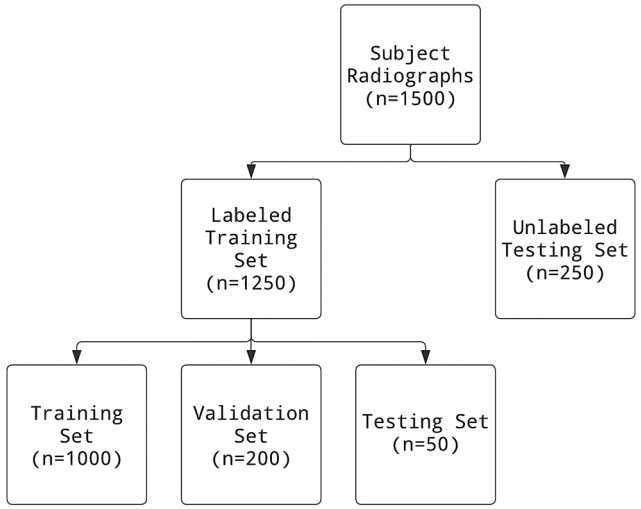


Fig. 5 Flowchart depicting the usage of subject data

Table 1 depicts the pertinent training options used on the network. Three network optimizer algorithms were considered, which were stochastic gradient descent with momentum (SGDM), root-mean-squared propagation (RMSProp), and adaptive moment estimation (Adam). The optimizer algorithm utilized is Adam due to it utilizing adaptive learning benefits from optimizers like RMSProp as well as the benefits of a gradient descent from optimizers such as SGDM [23]. Adam converged to excellent results without overtraining within a period of 25 training epochs; note that the learning rate was

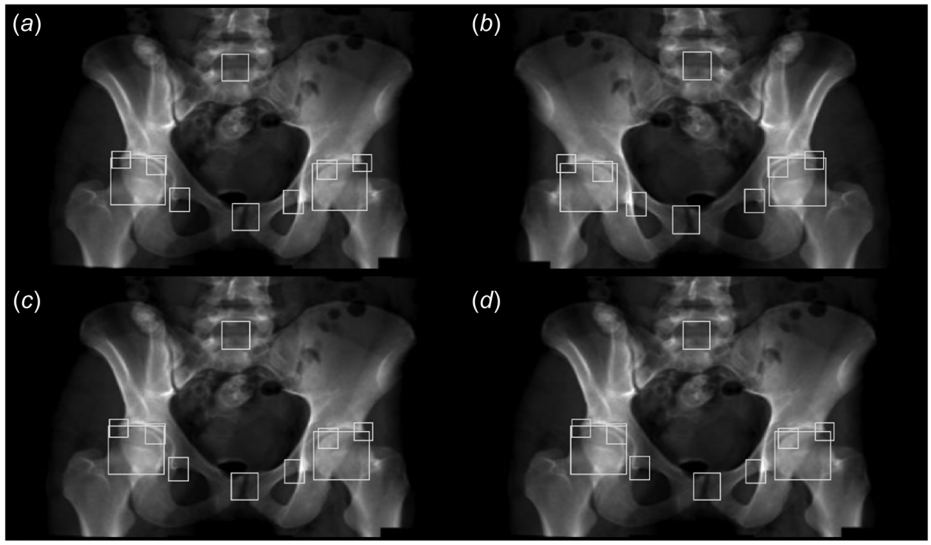


Fig. 6 Augmented training data for use in increasing algorithm accuracy ((*a*) is the original figure, (*b*) is the horizontal mirror of *a*, (*c*) is *a* with warped boundary boxes, and (*d*) is *a* with a different warp on the boundary boxes)

Table 1 Optimal training settings for network training

Training setting	Setting chosen
Maximum epochs	25
Learning rate	0.001
Mini-batch size	25
Batch normalization statistics	Moving
Network output	Best validation loss

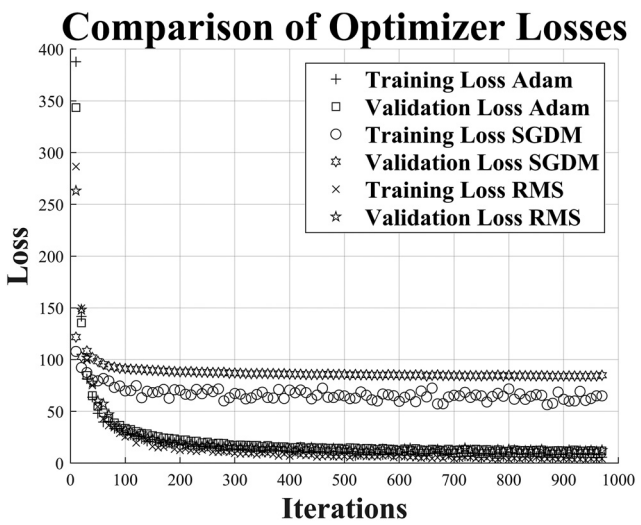


Fig. 7 Plot depicting the training and validation losses of the Adam, SGDM, and RMSprop

unchanged for each epoch as the network converged appropriately, as illustrated in Fig. 7. The bounding box loss was trained using a mean square error loss function, and cross-entropy was used to calculate the classification loss. The network was trained using a multi-GPU setup in parallel on an RTX A4000 and RTX4000 (NVIDIA Corp., Santa Clara, CA).

2.4 Postprocessing Setup. The network outputs had to be filtered through a multistage process to provide a similar comparison to the ground truth values. The process is shown in flowchart form in Fig. 8, with the input being 250 sets of predicted bounding boxes from the unlabeled testing set. The first step to postprocessing was to check if enough labels were found for each label category. After this stage, the predictions were converted from boxes into center point locations. The center point locations were run through a K-means clustering algorithm that was set using squared Euclidean distance as defined in the following equation:

$$d(x,c) = (x-c)(x-c)' \tag{1}$$

where x is the specific observation and c is the centroid for Eq. (1).

The clustering algorithm was used to separate the left and right hips to facilitate the correct identification for the computation of the metrics. Occasionally, multiple predictions were made on each side, which required the use of a weighted average shown in the following equation:

$$W = \frac{\sum_{i=1}^{n} \omega_i X_i}{\sum_{i=1}^{n} \omega_i} \tag{2}$$

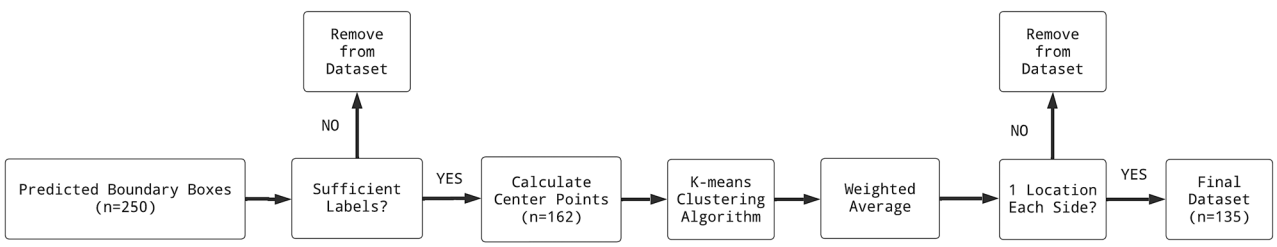


Fig. 8 Postprocessing flowchart showing filtering setup used on the network predictions

where n is the number of observations, ω_i is the weight for the corresponding observation, and X_i is the corresponding observation. The weighted average used the algorithm-computed scores as the weights and the center point coordinates for the observations. Finally, the last step of postprocessing is to check and make sure a single observation exists on each side of the hip.

3 Results and Discussion

The network efficacy can be analyzed using a combination of machine learning- and result-based metrics. The machine learning metrics provide insight into the network's accuracy and show where training deficiencies relative to the labeling may occur. In contrast, analyzing the results compared to expected values from surrounding literature helps determine the effectiveness of the labeling and network. Statistical analysis was also performed to assist in diagnosing sources of error within the network. Using both metrics in tandem will provide an overview of the general effectiveness of the network chosen and help inform future work.

3.1 Machine Learning Metrics. Numerous machine learning metrics are used to determine a trained network's accuracy and precision. These include precision, recall, F-measure, and mean intersection over union (IoU). The mean IoU is the average of the IoU. The IoU statistically gauges the similarity between two data sets shown in the following equation:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

A and B are different sets, with the IoU being the distance between the sets. Precision is defined as a proportional quantity that determines what percentage of identifications are correct and is defined in the following equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

where TP is true positive, and FP is false positive. The recall is defined as the proportion of true positives that were identified correctly and is defined in the following equation:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

where TP is true positive, and FN is false negative. Combining the precision and recall into a single metric can be done by calculating a metric known as the F-score. The F-score or F-measure is a metric that utilizes both precision and recall. The F-score is able to verify the accuracy of a model on a particular dataset and is defined as the harmonic mean of the precision and recall, as shown in the following equation: [24].

$$F - \text{Score} = 2 * \frac{P * R}{P + R} \tag{6}$$

where P is precision, and R is recall. The F-score is a generally accepted metric for use in analyzing networks, but it does possess limitations. The F-score has the primary limiting factor that it does not use the true negative in the calculations when used in a confusion matrix. This makes it ideal for situations where distinguishing whether measurements are correctly identified in the appropriate class is not a priority. However, for situations such as facial recognition, predictive analytics, and medical diagnoses, it is less suitable to rely on the F-score, as accurately classifying the appropriate locations is a crucial component to ensuring that the algorithm functions properly [25].

3.2 Medical Metrics. The results from the algorithm must be compared to known health metric values. The ranges for healthy radiographic metrics are not universally agreed upon as the frame for each metric changes with respect to numerous biological factors. In general, the range for dysplasia grows narrower as the age of the subject increases [26]. This general framework for analysis is tabulated in Table 2. Since the simulated radiographs are for healthy patients, ideally, the predicted metrics should fall within these expected values with a very small tolerance. However, accounting for the rotated images requires a degree of additional tolerance to compensate for the degree of misalignment. In this case, recall that the images were postprocessed utilizing the tolerances outlined in Sec. 2.4 and Table 3.

Table 2 General metric ranges for healthy, borderline, and dysplastic hips

Metric	Healthy	Borderline	Dysplastic	Sources
SA	33°–38°	39°–42°	>42°	[22]
CEA	25°–42°	20°–25°	<20°	[21,22]
ACIN	<= 13°	N/A	>13°	[21,22]
MP	<33%	N/A	>33%	[27]

Table 3 Outlier limits developed to allow for realistic values given the inherent limitations of misaligned radiographs

Measurement	Limit	References
Sharps angle	60°	[22,28]
Center-edge angle	60°	[21,22,29]
Acetabular index	25°	[21,22,30]
Migration percentage	40%	[27,31]

3.3 Statistical Analysis. A method to analyze the results is to compare the standard deviation of the predictions to the standard deviation of the ground truth data. If the network predictions possess similar standard deviations to the results, then it can be concluded that the outputs are not gaining a significant increase in variance error from the network. The formula for standard deviation is shown in the following equation:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \tag{7}$$

where σ is the standard deviation, x_i is each value in the set, μ is the arithmetic mean, and N is the size of the set. The size of N for this analysis is the number of subject images in each corresponding set (e.g., 25 for ground truth). While the standard deviation is a powerful statistical indicator, another statistic will be useful in comparing the sets of data, namely, the Z-score. The z-score determines the number of standard deviations a measurement in a particular set is from the arithmetic mean, and its formula is defined in the following equation:

$$Z = \frac{x - \mu}{\sigma} \tag{8}$$

where Z is the z-score, x is the observed value, μ is the arithmetic mean, and σ is the standard deviation related to the set that the observed value is a member. Utilizing these statistical metrics in tandem will allow for a greater understanding of the limitations of the network. Note that the metrics will be normalized using a probability density function that follows the formula defined in the following equation:

$$v_i = \frac{c_i}{N * w_i} \tag{9}$$

where c_i is the number of elements in the bin, N is the number of elements of the input data, and w_i is the width of the bin.

Table 4 Machine learning metrics are broken down by class showing deficiencies in two classes and high accuracy in two classes

Class	Avg. precision	Avg. F-measure
FH	0.9743	0.6899
SSMP	0.9291	0.6651
STP2	0.6105	0.5117
SO	0.5422	0.4913

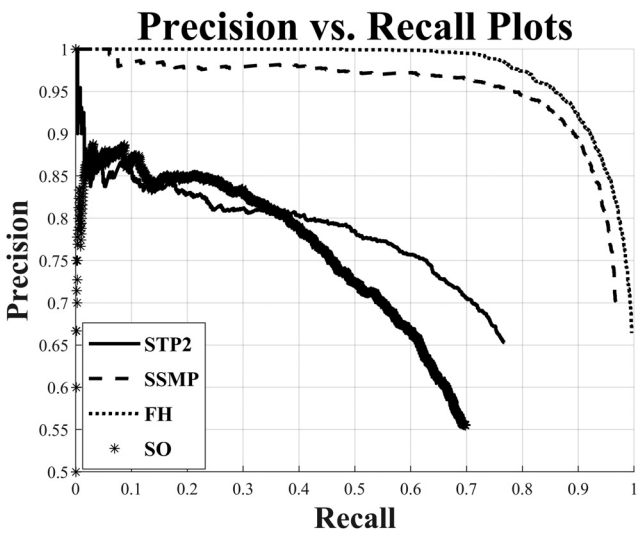


Fig. 9 Precision and recall plots broken down by class showing that the femoral head and lateral acetabular roof were located most accurately

3.4 Machine Learning Metric Outputs. The YOLOv4 network training outputs were compared to ground truths using the equations discussed in Sec. 3.1 and subsequently tabulated in Table 4. For reference, the mean IoU of the network was 0.80. The precision and recall were calculated using the computer vision toolbox function "evaluateDetectionPrecision." The IoU threshold for the function was set to the value of 0.5, which required half of the box output from the network to overlap with the bounding boxes from the ground truth data to count as an overlap. As seen in Fig. 9, while most of the labels were acceptably accurate, the two most accurate labels were the femoral head and lateral acetabular roof. This is excellent as those two locations are critical to measuring the center edge angle. The average precision and F-measure for those classes were also significantly higher than the remaining classes, as seen in Table 4.

This discrepancy can be attributed to the fact that the other 4 labels have a large degree of variability in shape, size, and location, and thus, due to the wide range, the algorithm struggles to identify points correctly. To rectify this, additional data could be used to improve training for these classes. Alternatively, these locations may necessitate a different algorithm type, such as a segmentation algorithm, to correctly identify these specific locations.

3.5 Machine Learning Image Evaluation. The raw ground truth values were used to judge the algorithm's accuracy. This was done by feeding the label locations into a MATLAB program and coding functions to calculate the angles for both hips. Additionally, a code was developed to show the calculation line locations on the image. Figure 10 shows an example output calculating the migration percentage. Note that the left and right hips are from the subject perspective, not the viewing perspective. The associated metrics in Fig. 10 can be seen in Table 5.

The ground truth calculations from Table 5 are compared to the machine learning algorithm predictions to determine the degree of variation in outputs and assist with characterizing the effectiveness

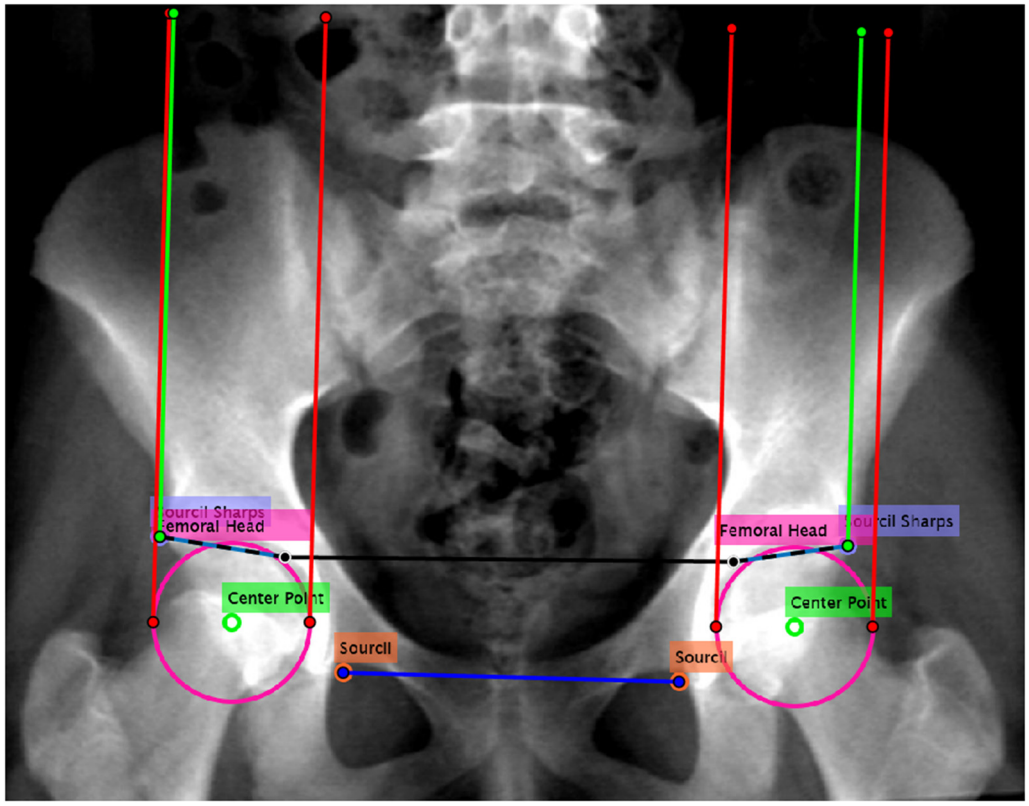


Fig. 10 Ground truth image used for migration percentage calculations (blue line is pelvic teardrop reference, pink represents FH, light blue SSMP, green dots are centers of the FH, red and green lines are distance markers for migration percentage calculations). (Color version online.)

Table 5 Medical metrics comparison between the ground truth and high-quality image

Metric	Ground truth		High quality image	
Abbrev.	Left hip	Right hip	Left hip	Right hip
SA	40.01°	34.68°	34.28°	36.37°
CEA	32.38°	41.91°	37.76°	40.95°
ACIN	8.42°	8.62°	6.93°	10.87°
MP	17.17%	2.87%	14.61%	9.18%

of the network. Figure 11 depicts the same image as shown in Fig. 10 using the machine learning outputs, and Table 4 (highlighted in yellow) shows the new calculated values based on the machine learning outputs. As seen in both Figs. 11 and 10, the outputs are close to the ground truth data shown in Fig. 10. Note that the shrunken ground truth images (391 × 234) were used as inputs, and thus the boundary boxes had to be resized by a factor of 4 to calculate and display the metrics on the higher-quality images. For this study, radiographs with high contrast margins between bone and soft tissue were defined as high-quality while radiographs with low contrast margins were defined as low-quality images. While the network produced reasonably accurate results for high-quality radiographs, it struggled to find certain points on low-quality radiographs. This is reflected in Fig. 12 and supported by the metric calculations in Table 6. These differences are unacceptably high and are heavily influenced by the pelvic teardrop locations being misidentified as well as the small shifts in the locations of the femoral head of the left hip. This shows that while the algorithm can locate the general points for analysis, it will need refinement to identify the metrics on rotated images properly.

In some cases, the network has difficulty distinguishing critical locations from background noise. This results in it not being able to find some of the requisite prediction locations, such as the acetabular teardrop and the medial head of the acetabulum, as illustrated in Fig. 13. This, in turn, prevents most of the metrics from being calculated since the teardrop reference line locations are not calculated. The teardrop reference line is used as the orthogonal reference to draw the vertical lines for the CEA and MP calculations.

Overall, the network can consistently identify key locations on high-quality, nonrotated grayscale images. This is of key importance as it proves that it is indeed possible to automate metric analysis using a traditional computer vision detection approach on misaligned radiographs through the automation of this process. While limitations exist, such as when the radiograph is rotated significantly, features are obscured, or contrast is poor, the network will still converge to a solution for most of the requisite network outputs. Steps can be taken to help overcome the limitations which are outlined in Sec. 3.7.

3.6 Statistical Analysis Results. In order to provide the best assessment of the trends, it is necessary to remove outliers from both the ground truth and fully processed network outputs. However, using the recommended values for healthy metrics from Table 2 does not give regard to the fact that rotated images can be accurately calculated to have larger angles due to an increase in variation from the misalignment [29]. A rule for calculating most of the metrics is to compute them with respect to a teardrop reference line drawn between both Kohler teardrops. The vertical lines to compute factors such as CEA and MP are defined to be orthogonal to the teardrop reference line. As such, if the Kohler teardrops are rotated in a misaligned image, the variation in the results can significantly increase. Taking these factors into account, along with referencing relevant articles for physically measured values, yields an outlier limit table, as shown in Table 3. All values computed to be greater than these were ignored.

The standard deviation and Z-scores for the four metrics between the ground truth and network predictions were compared. For the standard deviation, as shown in Fig. 14, the network prediction variance for each metric stayed within the boundaries of the

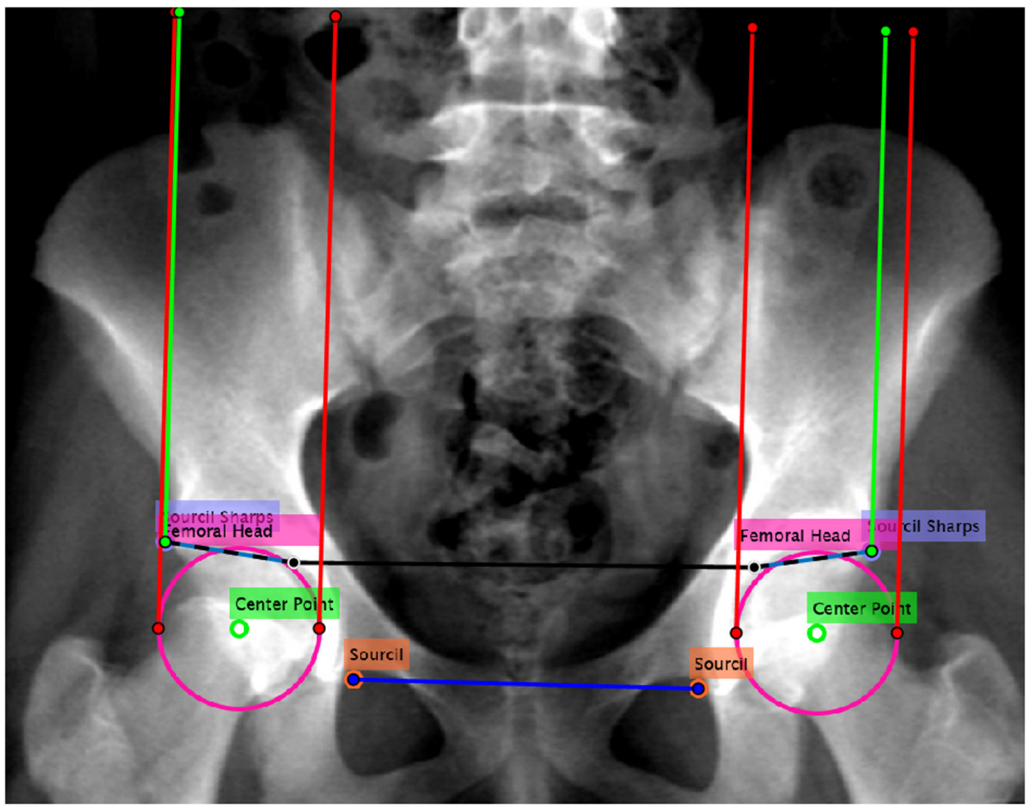


Fig. 11 Preliminary output results showing correlation to ground truth approximations (blue line is pelvic teardrop reference, pink represents FH, light blue SSMP, green dots are centers of the FH, red and green lines are distance markers for migration percentage calculations). (Color version online.)

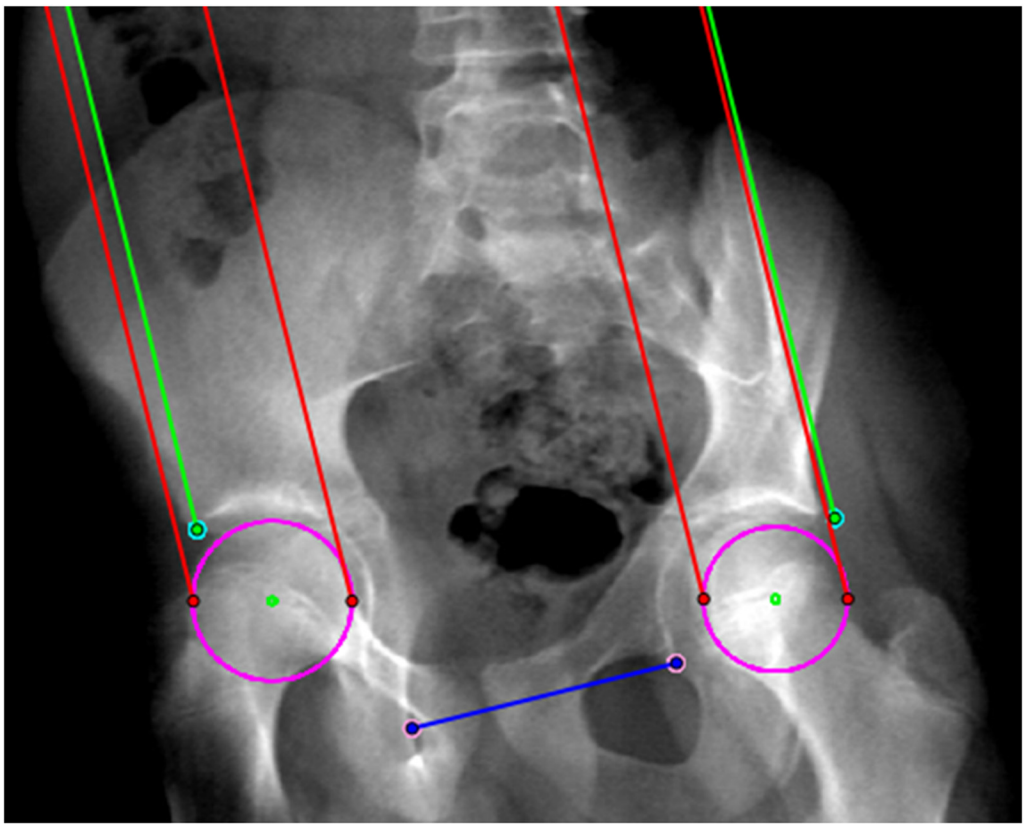


Fig. 12 Misaligned pelvic image showing networks struggle with locating certain locations such as the pelvic teardrop when rotated images are shown (blue line is pelvic teardrop reference, pink represents FH, light blue SSMP, green dots are centers of the FH, red and green lines are distance markers for migration percentage calculations). (Color version online.)

Table 6 Medical metrics predicted from nonaligned images show deficiencies in some of the calculations, specifically the ACIN and CEA this is corroborated by the percentage error between nonaligned predicted medical metrics and ground truth data

Metric	Ground truth		Low quality image	
Abbrev.	Left hip	Right hip	Left hip	Right hip
SA	38.78°	51.92°	28.55°	56.68°
CEA	45.33°	37.40°	50.12°	32.76°
ACIN	4.59°	1.28°	5°	2.38°
MP	6.94%	9.86%	4.90%	13.52%

manually labeled ground truth data. In particular, the network predictions possessed less variance for the acetabular index than the ground truth. The network did not increase the variance of the predictions in comparison to the given ground truth data. This shows that the network has converged to a solution and eliminates concerns of the network being overtrained.

The Z-scores can be seen in Fig. 15, which shows an excellent correlation between the true measures and the estimated ones. It is key to note that the standard deviations used to compute the Z-scores are the manual hand-labeled standard deviations for the corresponding image sets. This is key to understanding as it relates the network observations to the ground truth. The arithmetic mean values used were computed from the prediction sets. The Z-scores were plotted up to five standard deviations off of the mean; however, the majority remained within 3 standard deviations. Additionally, most of the observations lay within 1 standard deviation of the arithmetical mean. This is excellent, given the high precision shown in Fig. 14; this means that a majority of the labeling was consistent from both the hand and network predictions.

3.7 Limitations. The current limitations of the network stem from its inability to accurately locate some of the labels. Additionally, while 1500 images appear to be a large number in terms of machine learning, it is relatively small. When broken down, it only permits 60 images of each rotation, which may be insufficient to train the network to locate the less distinct points, such as the pelvic teardrop. There are also concerns regarding the fact that most detection algorithms were designed to identify color photographs rather than monochrome images. The addition of color assists in delineating key features of images. Overcoming these limitations will require modification of the approach for calculating the metrics and potentially require replacing and shifting the network layers and learning scheme. This network utilized a transfer learning scheme as it loaded a pretrained network that was subsequently retrained. Replacing the pretrained network with one more suited for medical imagery may improve results. Alternatively, switching the training method to change the weights in all layers of the network fully could potentially improve results. Additionally, it may become necessary to redefine the method of computing the assessment metrics if the orthogonality condition yields inaccurate results.

4 Conclusion

The goal of this study was to use a neural network to predict DDH metrics in radiographic images and address the limitations of DDH assessment. One of the key results noted is that the network responses are precise and statistically correlate with the ground truth labeling using standard deviation and Z-score analysis. Additionally, it was found that image quality plays a key role in whether the network will be capable of predicting the required locations to

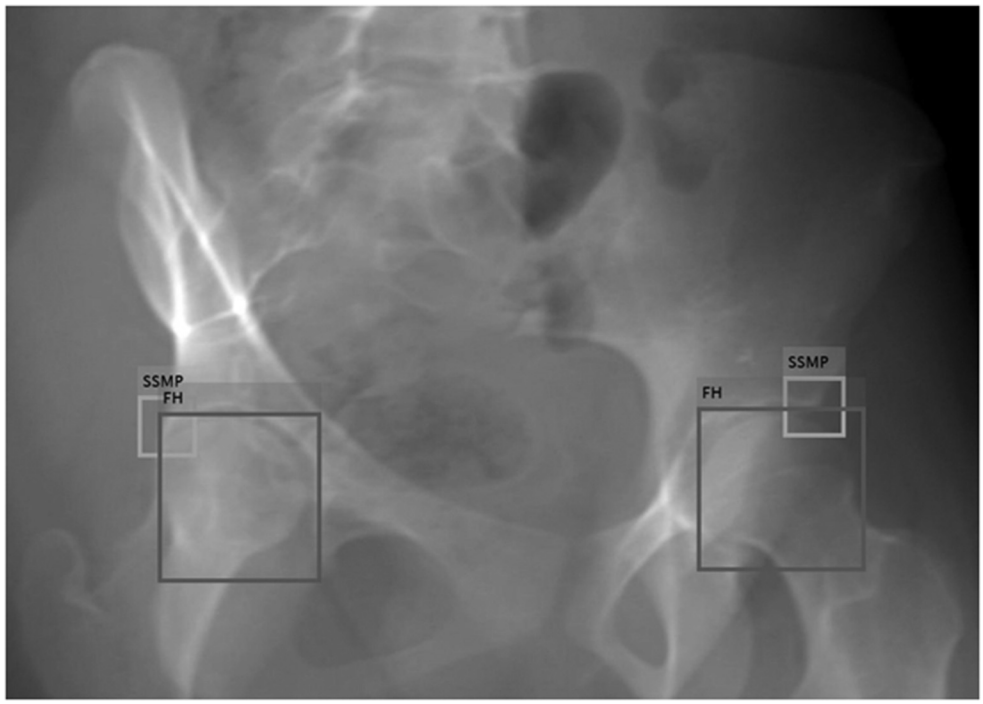


Fig. 13 Image depicting network limitations, particularly with not being able to locate the critical points for the teardrop and acetabular index reference lines, which are critical for calculating metrics

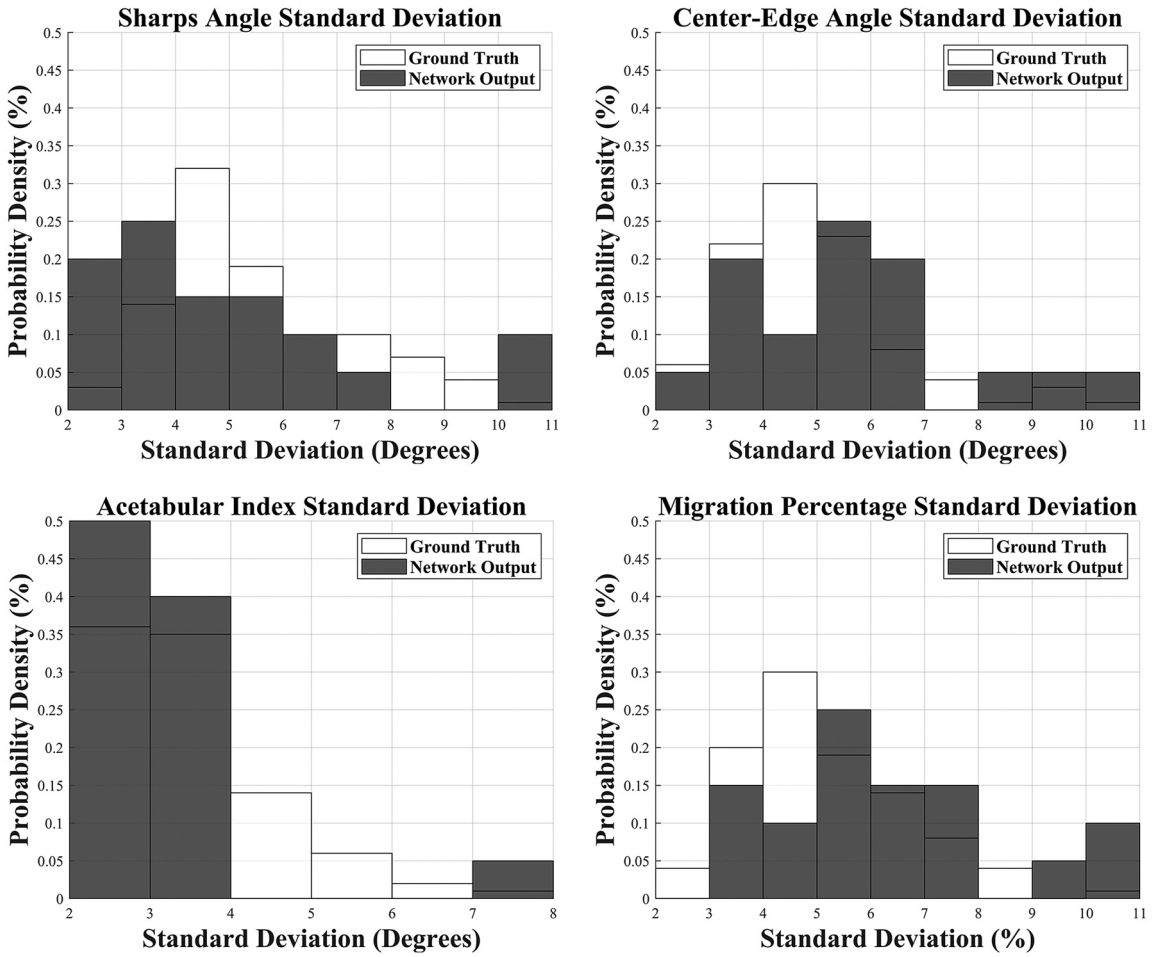


Fig. 14 Standard deviation comparison for the four medical metrics comparing ground truth and network outputs showing a correlation between input standard deviation variance and the output variance

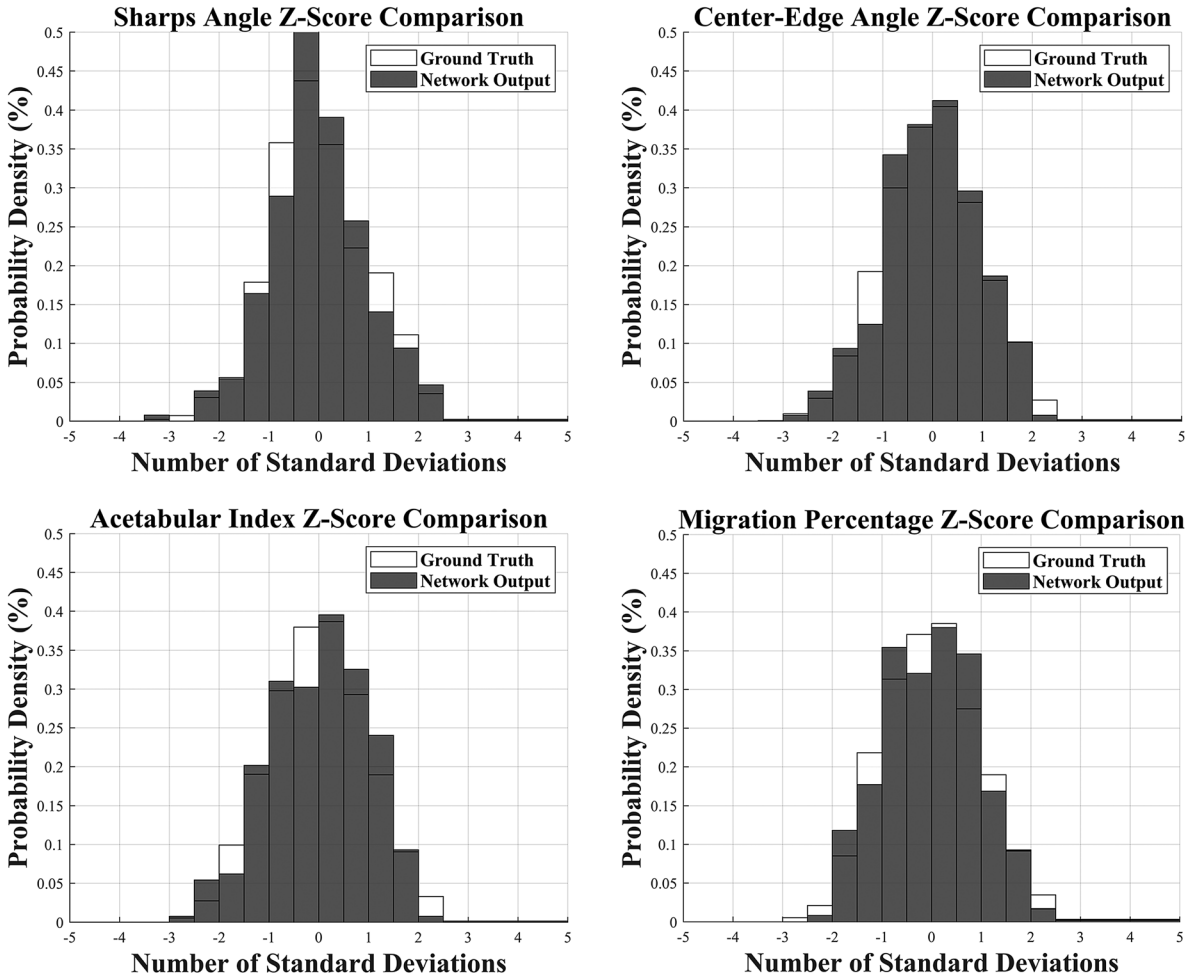


Fig. 15 Z-score comparison for the 4 medical metrics comparing ground truth and network outputs showing higher correlation between input variance and output variance than standard deviation

compute the medical diagnostic metrics. High-contrast images that were rotated along a single axis produced accurate results with predictions that converged. Lower-contrast images rotated along multiple axes either produced unsuitable results or were not capable of producing predictions. The network has significant difficulties with locating the pelvic teardrop and medial head of the acetabulum, which affects the measurements that rely on those locations. Overcoming the limitations will be required to proceed to the next phase of this study, which can be done by performing refinement as outlined in Sec. 3.7.

The applications of this neural network, once refined, can be extended to investigating radiographs where information is missing or corrupted, such as hemipelvic radiographs. Analyzing and correctly quantifying the metrics on datasets with nonideal or omitted information is of significant value. A key point to note is that the radiographic measurements of the values in this initial investigation are directly defined. This is to say that there were no attempts to correct the metrics to a value on an aligned radiograph. A future goal would be to have a machine learning network learn how to identify the bias in the values based on rotations according to the aligned image frame. This type of task is not easy for a human to perform, but a machine could potentially do so. The ability to automatically identify the radiographic metrics for use on longitudinal data is of extreme interest in understanding conditions that affect hip morphology and growth, which is of vital importance to treating conditions such as DDH.

Acknowledgment

This study was supported by the International Hip Dysplasia Institute (IHDI). The opinions, findings, and conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Funding Data

- US National Science Foundation CAREER (Award ID: CMMI-2238859; Funder ID: 10.13039/100000001).

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

Nomenclature

ACIN = acetabular index
CEA = lateral center-edge angle
CT = computed tomography
CNN = convolutional neural network
DDH = developmental dysplasia of the hip
FH = femoral head
FN = false negative
FP = false positive
IoU = intersection over union
MP = migration percentage
P = precision
R = recall
RMSProp = root-mean-squared propagation

SA = sharp's angle
SGDM = stochastic gradient descent with momentum
SSMP = sourcil sharps migration percentage
STP2 = sourcil Tönnis P2
TN = true negative
TP = true positive
YOLO = you only look once

References

[1] Agostiniani, R., Atti, G., Bonforte, S., Casini, C., Cirillo, M., De Pellegrin, M., Di Bello, D., et al., 2020, "Recommendations for Early Diagnosis of Developmental Dysplasia of the Hip (DDH): Working Group Intersociety Consensus Document," Ital. J. Pediatr., **46**(1), pp. 1–150.

[2] Loder, R. T., and Skopelja, E. N., 2011, "The Epidemiology and Demographics of Hip Dysplasia," ISRN Orthop., **2011**, pp. 1–46.

[3] Joiner, E. R. A., Andras, L. M., and Skaggs, D. L., 2014, "Screening for Hip Dysplasia in Congenital Muscular Torticollis: Is Physical Exam Enough?," J. Child. Orthop., **8**(2), pp. 115–119.

[4] Jejurikar, N., Moscona-Mishy, L., Rubio, M., Cavallaro, R., and Castañeda, P., 2021, "What is the Interobserver Reliability of an Ultrasound-Enhanced Physical Examination of the Hip in Infants? A Prospective Study on the Ease of Acquiring Skills to Diagnose Hip Dysplasia," Clin. Orthop. Relat. Res., **479**(9), pp. 1889–1896.

[5] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M., 2020, Optimal Speed and Accuracy of Object Detection, YOLOv4, arXiv.2004.10934.

[6] Lalehzarian, S. P., Gowd, A. K., and Liu, J. N., 2021, "Machine Learning in Orthopaedic Surgery," World J. Orthop., **12**(9), pp. 685–699.

[7] Stotter, C., Klestil, T., Röder, C., Reuter, P., Chen, K., Emprechtinger, R., Hummer, A., Salzlechner, C., DiFranco, M., and Nehrer, S., 2023, "Deep Learning for Fully Automated Radiographic Measurements of the Pelvis and Hip," Diagnostics (Basel), **13**(3), p. 497.

[8] Jensen, J., Graumann, O., Overgaard, S., Gerke, O., Lundemann, M., Haubro, M. H., Varnum, C., et al., 2022, "A Deep Learning Algorithm for Radiographic Measurements of the Hip in Adults—A Reliability and Agreement Study," Diagnostics (Basel), **12**(11), p. 2597.

[9] Holden, A., Seth, R., Ahmed, A., Joel, W., Ajay, K., Louis, V., Allan, H., et al., 2022, "Artificial Intelligence-Generated Hip Radiological Measurements Are Fast and Adequate for Reliable Assessment of Hip Dysplasia an External Validation Study," Bone Jt. Open, **3**(11), pp. 877–884.

[10] Zhang, S.-C., Sun, J., Liu, C.-B., Fang, J.-H., Xie, H.-T., and Ning, B., 2020, "Clinical Application of Artificial Intelligence-Assisted Diagnosis Using Anteroposterior Pelvic Radiographs in Children With Developmental Dysplasia of the Hip," Bone Jt. J., **102-B**(11), pp. 1574–1581.

[11] Park, H. S., Jeon, K., Cho, Y. J., Kim, S. W., Lee, S. B., Choi, G., Lee, S., et al., 2021, "Diagnostic Performance of a New Convolutional Neural Network Algorithm for Detecting Developmental Dysplasia of the Hip on Anteroposterior Radiographs," Korean J. Radiol., **22**(4), pp. 612–623.

[12] Zonoobi, D., Hareendranathan, A., Mostofi, E., Mabee, M., Pasha, S., Cobzas, D., Rao, P., Dulai, S. K., Kapur, J., and Jaremko, J. L., 2018, "Developmental Hip Dysplasia Diagnosis at Three-Dimensional US: A Multicenter Study," Radiology, **287**(3), pp. 1003–1015.

[13] Hareendranathan, A. R., Zonoobi, D., Mabee, M., Cobzas, D., Punithakumar, K., Noga, M., and Jaremko, J. L., *Toward Automatic Diagnosis of Hip Dysplasia From 2D Ultrasound*, IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, Australia, Apr. 18–21, pp. 982–985.

[14] Hawnaur, J., 1999, "Diagnostic Radiology," BMJ, **319**(7203), pp. 168–171.

[15] van Lent, W. A. M., Deetman, J. W., Teertstra, H. J., Muller, S. H., Hans, E. W., and van Harten, W. H., 2012, "Reducing the Throughput Time of the Diagnostic Track Involving CT Scanning With Computer Simulation," Eur. J. Radiol., **81**(11), pp. 3131–3140.

[16] Jacobsen, S., Sonne-Holm, S., Lund, B., Søballe, K., Kiær, T., Rovsing, H., and Monrad, H., 2004, "Pelvic Orientation and Assessment of Hip Dysplasia in Adults," Acta Orthop. Scand., **75**(6), pp. 721–729.

[17] van der Bom, M. J., Groote, M. E., Vincken, K. L., Beek, F. J., and Bartels, L. W., 2011, "Pelvic Rotation and Tilt Can Cause Misinterpretation of the Acetabular Index Measured on Radiographs," Clin. Orthop. Relat. Res., **469**(6), pp. 1743–1749.

[18] Faraj, S., Atherton, W. G., and Stott, N. S., 2005, "Inter- and Intra-Measurer Error in the Measurement of Reimers' Hip Migration Percentage," J. Bone Jt. Surg. Am., **87**(2), p. ADV42.

[19] Li, Q., Zhong, L., Huang, H., Liu, H., Qin, Y., Wang, Y., Zhou, Z., et al., 2019, "Auxiliary Diagnosis of Developmental Dysplasia of the Hip by Automated Detection of Sharp's Angle on Standardized Anteroposterior Pelvic Radiographs," Medicine (Baltimore), **98**(52), p. e18500.

[20] Fischer, C. S., Kühn, J. P., Ittermann, T., Schmidt, C. O., Gümbel, D., Kasch, R., Frank, M., Laqua, R., Hinz, P., and Lange, J., 2018, "What Are the Reference Values and Associated Factors for Center-Edge Angle and Alpha Angle? A Population-Based Study," Clin. Orthop. Relat. Res., **476**(11), pp. 2249–2259.

[21] Beltran, L. S., Rosenberg, Z. S., Mayo, J. D., De Tuesta, M. D., Martin, O., Neto, L. P., and Bencardino, J. T., 2013, "Imaging Evaluation of Developmental Hip Dysplasia in the Young Adult," AJR Am. J. Roentgenol., **200**(5), pp. 1077–1088.

[22] Mannava, S. M. D. P. D., Geeslin, A. G. M. D., Frangiamore, S. J. M. D. M. S., Cinque, M. E. M. S., Geeslin, M. G. M. D., Chahla, J. M. D. P. D., and Philippon, M. J. M. D., 2017, "Comprehensive Clinical Evaluation of Femoroacetabular Impingement: Part 2, Plain Radiography," Arthrosc. Tech., **6**(5), pp. e2003–e2009.

[23] Kingma, D. P., and Ba, J., 2014, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980.

[24] Cheng, F., Zhou, Y., Gao, J., and Zheng, S., 2016, "Efficient Optimization of F-Measure With Cost-Sensitive SVM," Math. Probl. Eng., **2016**, pp. 1–11.

[25] Hand, D. J., Christen, P., and Kirielle, N., 2021, "F: An Interpretable Transformation of the F-Measure," Mach. Learn., **110**(3), pp. 451–456.

[26] Önal, V., Metin Tellioğlu, A., and Durum Polat, Y., 2023, "Morphometric Assessment of the Hip Joint in Children Aged 2–13 Years," Clin. Anat., **36**(6), pp. 926–936.

[27] Terjesen, T., 2012, "The Natural History of Hip Development in Cerebral Palsy," Dev. Med. Child Neurol., **54**(10), pp. 951–957.

[28] Yang, G. Y., Li, Y. Y., Luo, D. Z., Hui, C., Xiao, K., and Zhang, H., 2019, "Differences of Anteroposterior Pelvic Radiographs Between Supine Position and Standing Position in Patients With Developmental Dysplasia of the Hip," Orthop. Surg., **11**(6), pp. 1142–1148.

[29] Hong, K.-B., Lee, W.-S., Kang, K., Kang, K. T., and Cho, B. W., 2023, "Evaluation of Lateral and Anterior Center-Edge Angles According to Sex and Anterior Pelvic Plane Tilt Angle: A Three-Dimensional Quantitative Analysis," J. Orthop. Surg. Res., **18**(1), pp. 280–280.

[30] Powell, J., Gibly, R., Faulk, W., Mayer, S. W., and O'Donnell, C. M., 2020, "Can EOS Imaging Substitute Traditional AP Pelvis Radiographs? A Comparative Study," Orthop. J. Sports Med., **8**(4 suppl3), p. 2325967120S00260.

[31] Cliffe, L., Sharkey, D., Charlesworth, G., Minford, J., Elliott, S., and Morton, R. E., 2011, "Correct Positioning for Hip Radiographs Allows Reliable Measurement of Hip Displacement in Cerebral Palsy," Dev. Med. Child Neurol., **53**(6), pp. 549–552.