Translational Animal Science, 2024, **8**, txad033
https://doi.org/10.1093/tas/txad033
Advance access publication 21 March 2023
Technology in Animal Science

Deep learning based landmark detection for measuring hock and knee angles in sows

Ryan L. Jeon[†], Joshua M. Peschel[†,1], Brett C. Ramirez[†], Joseph D. Stock[‡], and Kenneth J. Stalder[‡]

[†]Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA 50011, US
[‡]Department of Animal Science, Iowa State University, Ames, IA 50011, US

[1]Corresponding author: peschel@iastate.edu

Abstract

This paper presents a visual deep learning approach to automatically determine hock and knee angles from sow images. Lameness is the second largest reason for culling of breeding herd females and relies on human observers to provide visual scoring for detection which can be slow, subjective, and inconsistent. A deep learning model classified and detected ten and two key body landmarks from the side and rear profile images, respectively (mean average precision = 0.94). Trigonometric-based formulae were derived to calculate hock and knee angles using the features extracted from the imagery. Automated angle measurements were compared with manual results from each image (average root mean square error [RMSE] = 4.13°), where all correlation slopes (average R^2 = 0.84) were statistically different from zero ($P < 0.05$); all automated measurements were in statistical agreement with manually collected measurements using the Bland–Altman procedure. This approach will be of interest to animal geneticists, scientists, and practitioners for obtaining objective angle measurements that can be factored into gilt replacement criteria to optimize sow breeding units.

Key words: algorithm, computer vision, key point detection, swine

INTRODUCTION

Leg weakness in sows is a pertinent issue in swine breeding units that results in economic loss due to longer wean to service intervals and higher culling rates (Stalder et al., 2000). Signs of leg weakness manifest early in a sow's lifetime and can be detected by observing the conformation, stance, frame, and gait of a sow (Grindflek and Sehested, 1996; Koning, 1996; Van Steenbergen, 1989). Culling rates for gilts range up to 40% due to physical lameness, and in adult sows, a quarter of sow culling's are attributed to problems with the sow's feet and leg conformation (Stalder et al., 2003). Furthermore, Stalder et al. (2003) determined that sows reach positive net present value after their third parity with peak reproductive performance in the third through the sixth parity, but on average, are culled before their fourth parity. These decisions to selectively reduce female breeding stock are made by human observers, yet studies (Van Steenbergen, 1989; Main et al., 2000) have indicated that visual appraisal is highly subjective and dependent on many factors, including the evaluator's years of experience. Therefore, a more objective technique is needed for evaluating feet and leg conformation in gilts and sows due to subjectivity with human evaluators.

The hock and knee joint angles of sows are common conformation traits that are visually assessed by human evaluators. Draper et al. (1988) examined the effects of divergent selection for leg weakness in sows to find that the joint angle of the hock and knee are significantly different between the high and low genetic lines, suggesting that these angles can be potential indicators for risk of leg weakness. Additional genetic studies indicate that other feet and leg conformation traits range from low to moderate heritability (Reiland et al., 1978; Bereskin, 1979; Rothschild and Christian, 1988; Morrow et al., 1991; Serenius et al., 2001; Serenius and Stalder, 2004; Fan et al., 2009). Implementing an objective technique for quantifying feet and leg conformation traits can provide a more accurate estimate of heritability due to the higher repeatability of these measurements (Stock et al., 2017). Together, these studies indicate that with improved heritability estimates from an objective measurement technique, these traits can be included in genetic selection programs to improve hock and knee conformation, which is associated with sow longevity and productivity. This study demonstrates the application and repeatability of using a trained object detection deep learning model called You Only Look Once (YOLO) (Redmond et al., 2016) to objectively measure joint angles for the knee and hock in the side and rear stance position of sows.

MATERIALS AND METHODS

This section begins with data collection of sow images and describes the process of manually annotating biologically significant body landmarks. First, all images will be manually annotated for body landmarks. These images will be used to train an object detection model for these annotated body landmarks. The output of the deployed model is predicted detections on test images which can be used to objectively measure the angle between the landmarks (Fig. 1).

Received June 7, 2022 Accepted March 17, 2023.

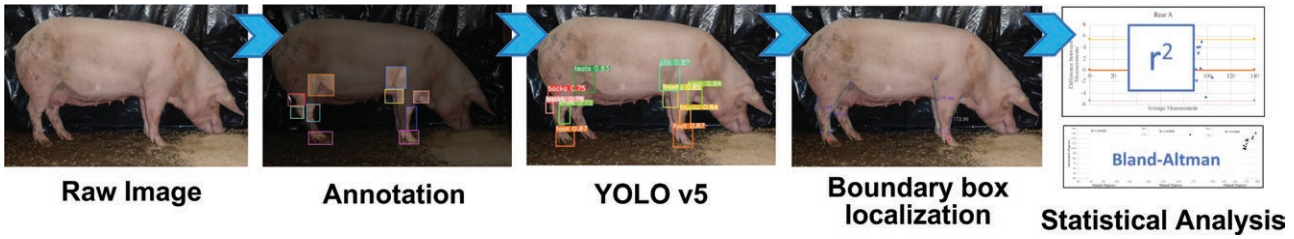

Figure 1. A figure summarizing the main components for determining hock and leg angles using the YOLO object detection algorithm for the side view. The automated measurement procedure requires raw images to be manually annotated. These images will be the training dataset that YOLO will use for training. The trained model localizes the location of each detected body landmark within a boundary box. From provided centroid and boundary box coordinates, a geometric algorithm determined hock and knee angles between body landmarks. These automated measurements are compared to those collected manually on the same image. Statistical tests determined the statistical significance of the slope (different from 0 and 1) between the automated and manual measurements and the statistical agreement between the two measures using the Bland–Altman tests.

Figure 2. Example annotation for the side and rear view images. Numbers correspond to class labels described in **Table 1**. There are two cases of feet in the side image, and two cases of hock and feet in the rear image. The geometric algorithm will identify which are the left and right based on their x and y coordinate.

Data Collection

Two raw image datasets consisting of side and rear profile images from 45 multiparous sows were collected using the procedure described by Stock et al. (2017). The raw image dataset comprised of 150 side view images and 100 rear view images. All images of the sows were captured using a digital camera (PL20, Samsung Electronics Co., Ltd. Yongin-City, Gyeonggi-Do, South Korea) in default portrait mode without the use of the zoom feature to maintain consistency across sows. The technicians held the camera approximately 2.4 m from the sow and 1.0 m above the ground.

Manual Annotation

Each image dataset required all images to be manually annotated for 10 body landmarks in the side view and 2 body landmarks in the rear view. An example of a manually annotated image is displayed in Fig. 2. These body landmarks are described in Table 1. Four augmentation methods artificially expanded the raw image dataset for increased variation throughout the training images: (1) horizontal flip, (2) saturation (between –25% and +25%), (3) hue (between –25 degrees and +25 degrees), and (4) exposure (between –25% and 25%). Augmentation assists with overfitting of the model and provides a more expansive set of possible scenarios to train from. Table 2 contains the final numbers of swine images in each of the three image datasets: train, validate, and test for the side and rear image dataset (total of six image datasets across the rear and side image datasets). The final dataset for the side portrait model uses 363 images while the rear view model uses 211 images after the artificial expansion of the image dataset through augmentation.

Deep Learning Approach

Introduction to the "You Only Look Once" algorithm.

The deep learning model YOLO is a one stage object detection algorithm that uses convolutional neural networks (CNNs) to detect objects in images or video. The YOLO

model can be used in real time applications, unlike traditional object detection models that use time expensive iterations of the sliding box windows. YOLO models implement a single pass (you only look once) of an image for faster predictions through a CNN to output a label and confidence probabilities (Fig. 3).

Overview of the proposed deep learning architecture.

The YOLOv5 object detection network (Fig. 4) can be understood in three parts: (1) a backbone, (2) a neck, and (3) a head.

Table 1. Labels and class names for the side and rear view

View	Label	Class
Side	0	Rump
	1	Back hock
	2	Foot
	3	Teat
	4	Front hock
	5	Shoulder
	6	Elbow
	7	Neck
	8	Front Elbow
Rear	0	Hock
	1	Foot

Table 2. Final number of images across the rear and side for the train, validate, and test image datasets

	Train	Validate	Test
Rear	211	16	16
Side	363	20	16

The backbone is where images are passed through CNNs, which in short, are pattern recognition algorithms that leverage deep learning techniques to predict boundary boxes for detected class objects and their probabilities. The YOLOv5's backbone is an open-source neural network framework called Cross Spatial Partial Darknet53, used for feature extraction and model training. Darknet-53 contains 53 convolutional layers, where the common spatial patterns connections find spatial patterns, splitting each layer into two. Through the neck, one layer passes through the remaining convolution layers, but the others do not. Like the original Darknet, all results are then aggregated for final predictions outputted by the head.

Statistical metrics are commonly used in deep learning.

Deep learning studies conventionally use three common statistical metrics: (1) Intersection of Union (IoU), (2) precision, and (3) recall. IoU is a popular metric to measure the localization accuracy of boundary boxes. This is the probability that the classifier found an anchor box containing a class object, determined by first calculating the IoU (Eq. 1) and illustrated in Fig. 5. Predicted boundary boxes that closely match the area of the annotated boundary box will receive a higher IOU than those with little overlap. Together, the IoU and confidence score determines if the classifier's prediction is a true positive or false negative.

$$\text{IoU} = \frac{\text{Area of Intersection (shaded)}}{\text{Total area of union}} \quad (1)$$

Precision and recall are two statistical metrics that are used to evaluate potential models during the training phase. Precision is defined as the number of true positives divided by the sum of the true positives and false positives. Recall is defined as the ratio of true positives, and sum of true positives and false negatives. Figure 6a displays the

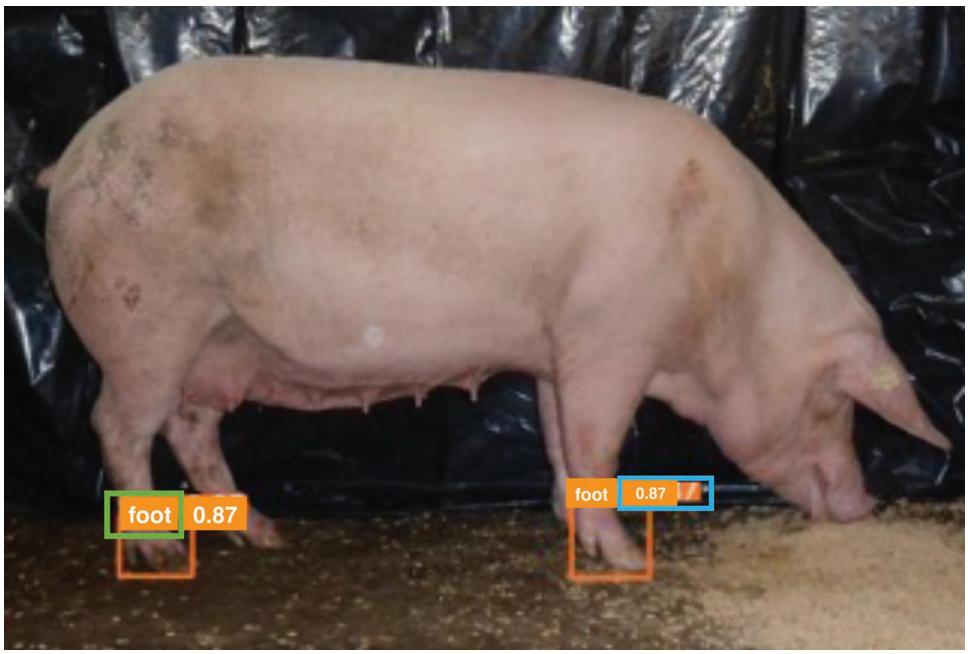

Figure 3. Example output from YOLO v5 model. Through a single pass of the CNN, the YOLO model predicts the labels (green box), boundary boxes (orange box), and the confidence probabilities (blue box) for predicted objects detected in the image.

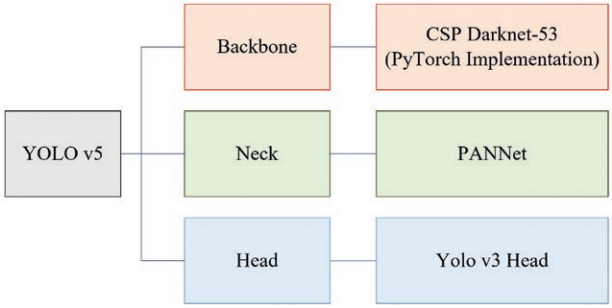

Figure 4. Parts to the YOLO v5 network. The neck features many roles, but primarily serves as an aggregation step. Features like the pyramid pooling and path aggregation are featured in the neck. One of the advantages of later versions of YOLO is increased accuracy due to improvements in the neck. Finally, the head is used to implement feature detection of the sow body landmarks through predicted annotations in each test image.

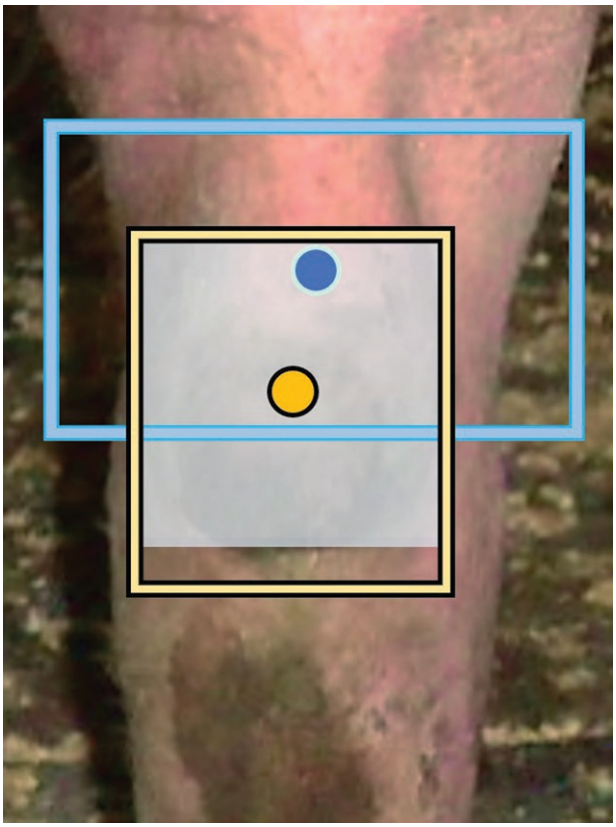

Figure 5. An example depicting IoU) calculation. The blue box is the predicted boundary box, and the orange is the ground truth label. The orange and blue dot is the centroid of the detected hock for the ground truth and predicted boundary box. The IoU metric is calculated as the ratio of the area of overlap between a predicted boundary box and the manual annotation, over the area of the union between the two boundary boxes (Eq. 1).

inverse trend between confidence score and recall in our side view model, and shows that the two metrics follow an inversely proportional relationship. This trend is the result of the critical values increasing as confidence increases, which widens the confidence interval, and results in less precision. Alternatively, Fig. 6b shows the proportional trends between confidence score and precision for our side view model, indicating that as the confidence score increases, recall will increase.

Average precision (AP) is a statistical metric used to compare models with different parameters, such as different epoch sizes, test image sizes, or changes to model architecture. This metric is defined as the area under the precision-recall curve, which can be manually calculated using integration to find the area underneath the curve. The mean average precision (mAP) for a class, is the average of the AP for that class, across all test images. The overall mean average precision (overall mAP) is the average of all average APs of each class. Machine learning models benefit from both a larger training image dataset and a greater number of epochs, which is the number of passes a model undergoes during training. Models using different parameters can be compared by using the overall mAP. In Table 3, the overall mAP of each model increases with increased training image dataset size. The mean average precision increases with increased epoch size (Fig. 7) which follows a similar trend to training image dataset size. The default—*patience* function sets an early truncation, activated at a plateau when setting the number of epochs to 500, curtailing the final epoch count to 331 for the side images, and 221 epochs for the rear images.

Training, validation, and test image datasets.

The raw image dataset comprised of 150 side view images and 100 rear view images for training the model. Images that were unable to be manually annotated were removed. A preliminary calculation using the Bland–Altman test (Altman et al., 1983; Bland and Altman, 1986) indicated that for an alpha value of 0.05 at a 95% confidence interval, 11 and 16 test images were needed to measure the statistical difference between the two measures for the rear and side, respectively. For higher confidence, test image dataset for the rear increased to 16. This study used a ratio of 70-15-15 for rear view model, and 70-20-10 for the side view, model where there were 16 test images for each of the side and rear-view image datasets.

The training dataset consists of the largest proportion of the total images (70%) because the manual annotations labeled on the training images are specifically used to guide the parameters of the new model during the training phase. Using algorithms such as the tensors previously described, the model determines pixel elements of the test image anchor box that is common with elements within the annotated boundary box. It is possible, that if the model is trained purely on test images, then overfitting can occur resulting in a trained model that cannot accurately detect objects in images that are different from images found in the training dataset. For example, the model developed on a dataset trained purely on pigs in one barn may not work when using images from a different barn at a different time of day. Therefore, a validation test set is used to determine validation mAP of the model which provides an indicator of how well the model is when trained on images different than those in the training dataset. While the validation dataset traditionally comprises 15% of the total dataset, it influences the loss functions that are used to modify the hyperparameters for each model. The validation mAP after each epoch is used to determine which model is considered the best. Since these parameters are influenced by the metrics obtained by the validation dataset, it is important to create a test dataset separate from the other two

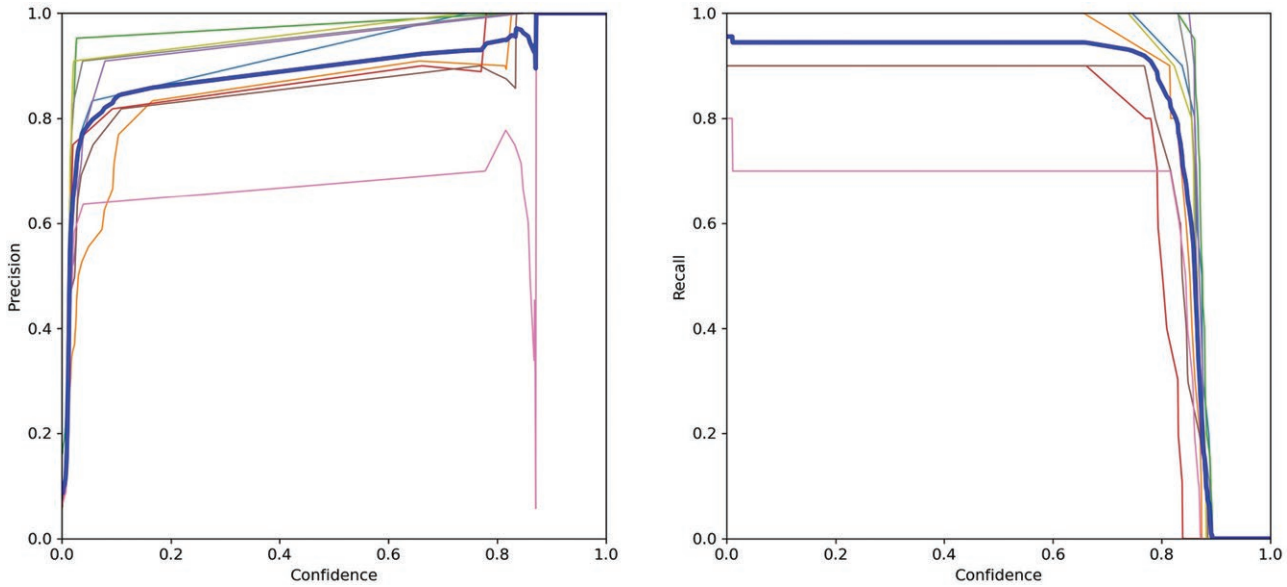

Figure 6. Trends between (a) recall and confidence and (b) precision and confidence for each class in the side view. The thicker blue line represents the average across all classes. Recall for all class objects decreases as confidence increases due to changes in confidence intervals. The inverse occurs where precision increases as confidence decreases. Pink (the neck) followed erratic patterns in later epochs due to possible overfitting in the training and validation dataset.

Table 3. Evaluating model performance over varying numbers of training images (only side view model shown). The mAP values increase with an increasing number of training images. Maximum number of epochs is dependent on the number of training images. The trained model using 363 training images over 331 epochs received an overall mAP of 0.94

# Training images	Overall mAP	Patience enabled final epoch count
10	0.48	72
121	0.87	278
242	0.88	310
363	0.94	331

datasets. The test dataset traditionally comprises 10% of the dataset but provides an overall mAP metric that is used to measure how well the model works overall, that is not biased by the validation test dataset.

Determining hock and knee angles from training model output.

The trained YOLO v5 model outputs a.txt file that consists of localization data for each detected object in every test image. The.txt file is organized where each row is an instance of a detected class object. These labels follow the same numerical labeling in annotation and in Table 2. There are by default, five columns that represent the label, x and y coordinate for the centroid and the height and width of the boundary box. It is challenging to set a custom IoU and non-maximal suppression threshold without knowing the IoU for each boundary box. Therefore, this model enabled the *–IOU* parameter during output, so each.txt file output would include the IoU as well. Therefore, six parameters for each output are utilized and listed below:

1. Label (Refer to Table 2)
2. x Centroid Coordinate
3. y Centroid Coordinate
4. Height of Boundary Box
5. Width of Boundary Box
6. IoU of Boundary Box

Hock and Knee Angles

This paper uses a geometric algorithm to obtain two hock angles from the rear profile and four hock and knee angles from the side profile using data outputted from a YOLOv5 model trained on images of sows collected from two barns. The hock angles from the rear view are Left (denoted Hock-Left; H-L) and Right (denoted Hock-Right; H-R). The four hock and knee angles from the side view are denoted Side-Back-1 (back side of the back leg, abbreviated S-B-1), Side-Back-2 (front side of the back leg, abbreviated S-B-2), Side-Front-1 (back side of the front leg, abbreviated S-F-1), and Side-Front-2 (front side of the front leg, abbreviated S-F-2). These angles are derived based on previous literature from Stock et al. (2017), Nakano et al. (1987), and Van Steenbergen (1989), as well as by the PIH 101 Feet and Leg Soundness in Swine (Wood and Rothschild, 2001) guide from the National Swine Improvement Federation.

Rear hock angles.

In the rear view, the two angle measurements are determined using the feet and hocks. In quadruped mammals, the hock is a backward protrusion between the tarsal bones and the tibia of the posterior side of the animal. The H-L angle is determined by calculating the angle between the left foot and the right hock, where the vertex of H-L is the left hock. The H-R angle is determined by calculating the angle between the right foot and left hock, where the vertex of H-R is the right hock.

Rear hock algorithm.

For each of the 16 rear view test images, four boundary boxes are detected in each image. These four boundary boxes

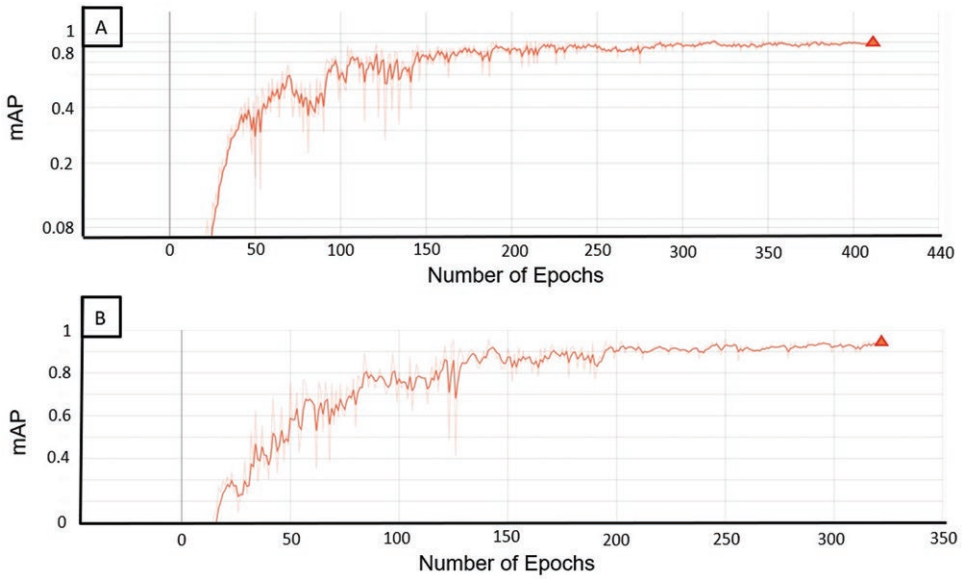

Figure 7. Relationship between mAP and the number of epochs (a: side, b: rear). The number of epochs plateaued at the start of 300 for the side images, and at the start of 200 for the rear images. Triangles in orange are the early truncation point automatically enabled by the patience function.

are the feet (left/right) and the hock (left/right). The trained model provided an output of centroid coordinates for each detected body landmark. For calculating the left hock angle, the centroids of the left foot, left hock, and right hock are used, where the vertex of the angle is the left hock. Euclidean distances between the centroid of the left foot, the left hock, and the right hock are calculated using a derivation of the Pythagorean theorem (Eq. 2):

$$\text{distance} = \sqrt{(x_2 - x_1)^2} + \sqrt{(y_2 - y_1)^2} \tag{2}$$

Angles are calculated using a derivation of the Law of Cosines (Eq. 3):

$$c^2 = a^2 + b^2 - 2ab \cdot \cos(C) \tag{3}$$

where a, b, and c are distances of a triangle, and C is the angle opposite of vertex c (Eq. 4). The inverse of the derivation can then be directly used to calculate the angle when provided distances of each side of the triangle (Eq. 5).

$$\cos(C) = \frac{a^2 + b^2 - c^2}{2ab} \tag{4}$$

$$\arccos(a) = \frac{a_{12}{}^2 + a_{13}{}^2 - a_{23}{}^2}{2(a_{12} \cdot a_{13})} \tag{5}$$

where a_{12} is the distance between the left foot and the left hock, a_{13} is the distance between the right hock and left hock, and a_{23} is the distance between the left foot and right hock.

Calculation of the right hock angle follows the same geometric approach used in Eq. 4 but uses the centroids of the left foot, left hock, and right hock, where the vertex of the angle is the right hock. An example is shown in Fig. 8.

Figure 8. Centroids detected by the trained YOLO model, and body angles calculated from the detected body landmarks, for the rear view image. Centroids used for the angles are in blue, where the vertex of the angles are highlighted.

Side hock and knee angles.

For each of the 16 test images in the side view, the trained model identified and localized 10 body landmarks to use in

hock and knee angle measurements. The side view angles are calculated using the same derivation of the Pythagorean theorem and Law of Cosines as described previously for the rear view. All four sets of hock and knee angles in the side view required three body landmark locations. The first set of body landmarks are used to detect the posterior hock angle. This angle required the centroid coordinates of the rump, the hock, and the back foot. The second set of body landmarks are used to detect the anterior hock angle. This angle required the centroid coordinates of the teats, the front hock, and the back foot. The third set of landmarks made the posterior knee angle, which required the centroid coordinates of the shoulder, the back knee, and the front foot. The fourth set of landmarks made the anterior knee angle, which required the centroid coordinates of the neck, the front knee, and the front foot. An example of a side view is shown in Fig. 9.

Side Hock and Knee Algorithm

In the side view, four angles are determined, starting from the posterior side of the animal. The S-B-1 and S-B-2 angles are the angle between the tarsal bone and the tibia/fibula. The knee of the sow is a hinge joint, made up of various muscles and ligaments that connects the humerus with the radius and carpals. The S-B-2 angle deviates slightly from the one used in Stock et al. (2017), where the upper vertex is the maximum of curvature that defines the conflux between the loin, the leg, and the side. The S-F-1 and S-F-2 angles are determined by the angle between the pig's humerus with the radius/carpal. The S-F-1 angle deviates from the one used in Stock et al. (2017), where the upper vertex is shifted to the maximum of the curvature that defines the shoulder, and the vertex is shifted directly above the shoulder joint.

Both models considered the body landmark "foot" as one class and later renamed front and back foot using the coordinates provided by the model output. Previously, the model trained using images annotated with both "front foot" and "back foot," had difficulty discerning between the front and back foot which resulted in multiple simultaneous predictions of the front and back foot, even with non-maximal suppression enabled. Therefore, a different iteration of the model found higher mAP when foot is annotated as one class, where each test image is annotated with two instances of foot. The foot coordinate is later preprocessed into front and back foot by determining which foot coordinate for the x centroid is closest to the rump and the neck.

The feet centroids underwent coordinate adjustment using the boundary box height and width to determine the southernmost vertices. The model accurately measured foot centroid in each image, yet the hock and knee angle did not directly use the centroid of the foot. However, the boundary box corners indicated the end vertices of each hock and knee angle. Therefore, the upper left corner of each foot boundary box became the foot coordinate of the posterior angle measurements. Subtracting half the width and height of each boundary box from the centroid coordinate provides the upper left corner of the boundary box. For the anterior angles, subtracting half the height of the boundary box from the foot centroid yielded the foot centroid. All angles are found using the same arccosine function described previously. An example of adjusting foot angles is displayed in Fig. 10.

Manual Evaluation

Three human evaluators manually calculated all six hock and knee angles independently using J-image (Rasband, 2018), a

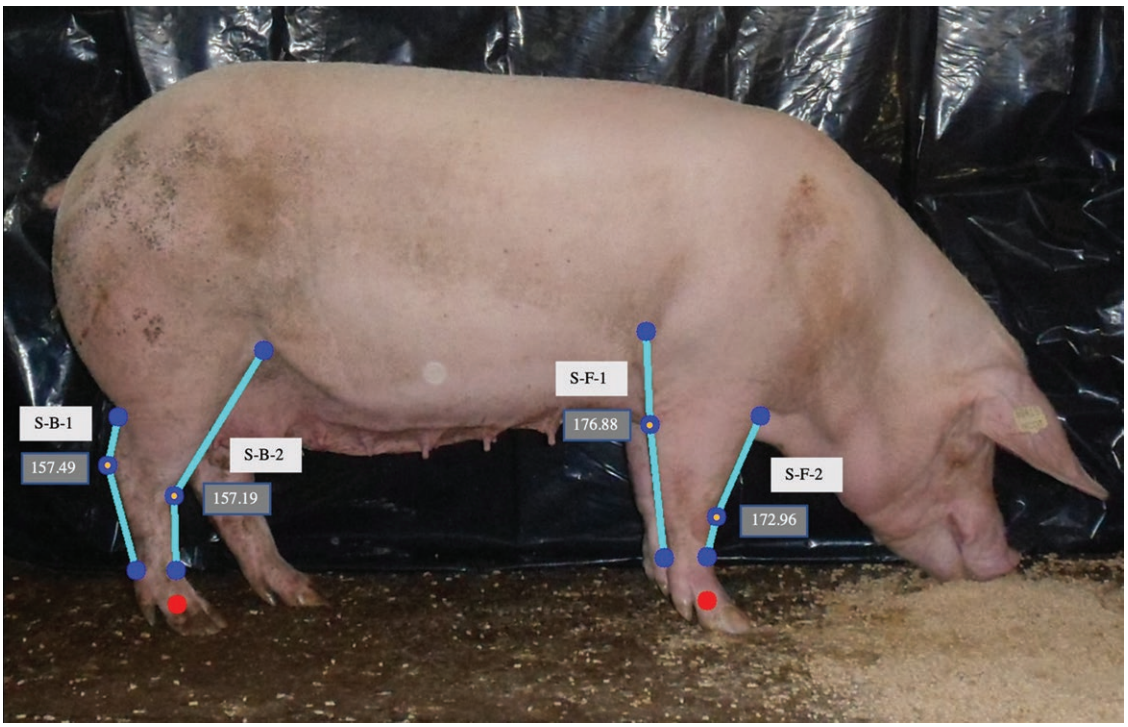

Figure 9. Centroids detected by the trained YOLO model, and body angles calculated from the detected body landmarks, for the side view image. Centroids used for the angles are in blue and the red centroids represent the centroid of the class foot. Angle vertexes are highlighted in yellow.

Figure 10. Calculation of adjusted foot coordinates (blue) from origin (red). Boundary box in gray. Where x_{cb}, y_{cb} are the centroid coordinates of the back foot, and x_{cf}, y_{cf} are the centroid coordinates of the front foot.

software used for simple image analysis that calculates angles by manually clicking three points. Instructions were provided so evaluators knew where the ideal vertex and two endpoints were located on each image. For quality control, all datasets contained three duplicate images, which were used to check that labelers were consistent across the same image. The three duplicate images were found to be effective in determining the quality of the human evaluator, as the observed measurements would either be similar to each other or different. These angle measurements were averaged across the three evaluators to obtain a total average for each manually obtained angle measurement.

Manual evaluation methodology

1. Using J-Image, click the angle button at the top to access the angle measurement tool.
2. The mouse will now change to indicate the angle tool has been enabled.
3. For each body part, please observe the image provided.
4. Then, click the three points that best match the three points displayed in the example image.
5. On a new version of the provided excel template, write this value down under the correct file name.

Statistical Analysis for Comparing Automated and Manual Methods

Evaluating the reliability of the manual methodology.

Interclass correlation coefficient (ICC) measures the strength of the evaluation methodology by measuring how strongly the angles obtained from each evaluator resembled each other. The ICC indicates if the hock and knee angles can be reliably determined by different evaluators. The ICC was determined by performing a two-way ANOVA test. Values of ICC that approach 1 indicate high reliability of an experimental method, while values that approach 0 indicate poor reliability across different raters. Using the values obtained from the test, the ICC was determined using Eq. 6 (Fisher et al., 1950).

$$\text{ICC} = \frac{\text{Variance of Interest}}{\text{Total Variance}} \tag{6}$$

Hypothesis testing for the regression slope.

Pearson correlation coefficient quantified the relationship between the manual and automated measurements and was used to test if the regression slope is statistically significant from zero and one. A regression slope hypothesis test where, the null hypothesis assumes that the slope is equal to zero, tests if there is a significant linear relationship between the automated measurement X and the manual measurement Y. The result of the test is a confidence interval that may or may not include 0. When testing for difference from zero, a confidence interval that contains 0 indicates that 0 is a likely candidate for the true value of the difference. While the *P*-value is unknown, it can be inferred that the *P*-value will be less than alpha, 0.05. A significance level of 0.05 was used for the linear regression *t*-test. Standard error of the slope was calculated using Eq. 7.

$$SE_{slope} = \frac{\sqrt{\sum \frac{(y_i - \bar{y}_i)^2}{n-2}}}{\sqrt{\sum (x_i - \bar{x}_i)^2}} \tag{7}$$

where n is the number of observations, y_i is the automated measurement for observation i, $\bar{y}_i$ is the estimated automated measurement for observation i, x_i is the manual measurement for observation i, $\bar{x}_i$ is the estimated manual measurement for observation i.

Degrees of freedom for a simple linear regression is equal to n–2. The t statistic is determined using the following formula:

$$t = \frac{b_1}{\text{standard error of slope}} \tag{8}$$

The *P*-value to test for difference from zero is calculated using the t statistic (Eq. 8) and the degrees of freedom.

Root mean squared error is a statistical metric used to compare the distance of individual measurements from the regression line. This is calculated using Eq. 9.

$$\text{RMSE} = \sqrt{\frac{\Sigma (P_i - O_i)^2}{n}} \tag{9}$$

where n is the number of observations, P_i is the automated measurement for observation i, O_i is the manual measurement for observation i.

Bland–Altman test for determining statistical agreement.
The Bland–Altman test is considered the standard for comparing two quantitative measurement techniques (Altman and Bland, 1983; Bland and Altman, 1986) and is used in this study to test for statistical agreement between automated and manual measurements on the same sow image. This procedure quantifies the agreement between two measures on the same, individual unit and is conducted by plotting the difference between the automated and manual measurements on the y axis, regressed on the average of the values on the x axis. The technique recommends limits of agreement set at 1.96 ±SD of the mean difference. Evaluation of the two methods is based on how many points are outside the borders set by the limits of agreement.

RESULTS

Interclass Correlation Coefficient

Table 4 presents the ICC across all 16 images, between the three evaluators. This study found an interclass correlation of 0.92 for the side profile and 0.93 for the rear profile. Further breakdown of these two groups indicates that most of the angle measurements are moderately to highly repeatable (ICC = 0.82 to 0.94), with the exception of S-B-1 (ICC = 0.69).

Correlation Between Manually and Automated Measurements

Table 4 presents the coefficient of determination (R^2), slope (±SE), intercept (±SE), RMSE, the *P*-value to test if the slope is significantly different from zero, and the confidence interval for alpha = 0.05. Correlations between the automated and manual measurements ranged from 0.67 to 0.95 across all angles. The greatest correlation is 0.95 for the H-L angle, while the lowest correlation is 0.69 for the S-B-1 angle. All *P*-values were found to be less than 0.05, and therefore interpreted as significantly different from zero. The confidence interval for alpha = 0.05 is used to test if the slope is statistically different from zero. If the confidence interval includes 0, it suggests that the true value can reasonably be 0. However, in this study, all confidence intervals did not include zero, indicating that the slope for all angles were statistically different from zero. The resulting *P*-values are less than the alpha value (0.05), which is considered the conventional threshold for statistical significance.

Bland–Altman Plots

Table 5 presents the results obtained from the Bland–Altman analysis between the manual and automated angle measurements. The average difference between automated and manual measurements ranged from –0.04 to 4.51. H-L, and S-F-1 had the smallest mean of difference, at –0.04 and –0.01, respectively. S-B-1 and S-F-2 had the largest difference, at 4.51 and 3.65, respectively. All means are positive or close to zero, indicating that on average, automated measurements generally are larger than manually collected measurements. Figure 11 displays the Bland–Altman plots, with bounds described in Table 5. The Bland–Altman plots indicate that all angle measurements are statistically similar to each other. S-B-1 had three points outside the threshold, which is five points below the 50% threshold required. S-B-2, S-F-1, and S-F-2 had two points that are in close proximity to the threshold border.

Table 4. Summary statistics with ICCs, regression coefficient (R^2), slope (± standard error [SE]), intercept (±SE), root mean squared error (RMSE), the *P*-value to test if the slope is significantly different from zero, and confidence interval of the slope

	ICC[1]	ICC[2]	R^2	Slope (±SE)	Intercept (±SE)	RMSE	*P* value	CI
H-L	0.93	0.93	0.95	1.35 (0.08)	–32.80° (7.77°)	5.84°	<0.01	(0.50, 0.91)
H-R		0.94	0.86	1.08 (0.11)	–7.18° (9.52°)	3.07°	<0.01	(0.76, 1.24)
S-B-1	0.92	0.69	0.74	0.97 (0.16)	8.71° (24.00°)	5.66°	<0.01	(0.40, 1.21)
S-B-2		0.82	0.67	0.70 (0.14)	42.41° (19.46°)	2.66°	<0.01	(0.41, 1.59)
S-F-1		0.87	0.75	0.94 (0.15)	10.05° (25.90°)	2.82°	<0.01	(0.58, 1.19)
S-F-2		0.84	0.86	0.85 (0.09)	27.55° (15.23°)	4.73°	<0.01	(0.57, 1.12)

[1]ICC by group (rear profile and side profile) calculated using Eq. 6.
[2]ICC by individual angle calculated using Eq. 6.

Table 5. Statistical parameters used to create the Bland–Altman plots for each knee and hock angle (Hock-Left, Hock-Right, Side-Back-1, Side-Back-2, Side-Front-1, Side-Front-2). Means of difference indicate average offset, where negatives indicate that the automated measured less on average, and positives indicate the automated measured higher on average. Upper and lower bound thresholds are calculated using a 95% confidence interval. Points in close proximity to the threshold border are also noted. All manual measures are averaged across three evaluators

	Mean of difference (degrees)	Upper bound (degrees)	Lower bound (Degrees)	Points outside Threshold	Points on border	Are methods in statistical agreement?
H-L	–0.04	–12.27	12.18	0	0	Y
H-R	0.88	–6.74	8.50	1	1	Y
S-B-1	4.51	–2.41	11.43	3	1	Y
S-B-2	0.27	–5.11	5.65	0	2	Y
S-F-1	–0.01	–5.73	5.70	0	2	Y
S-F-2	3.65	–2.43	9.74	1	2	Y

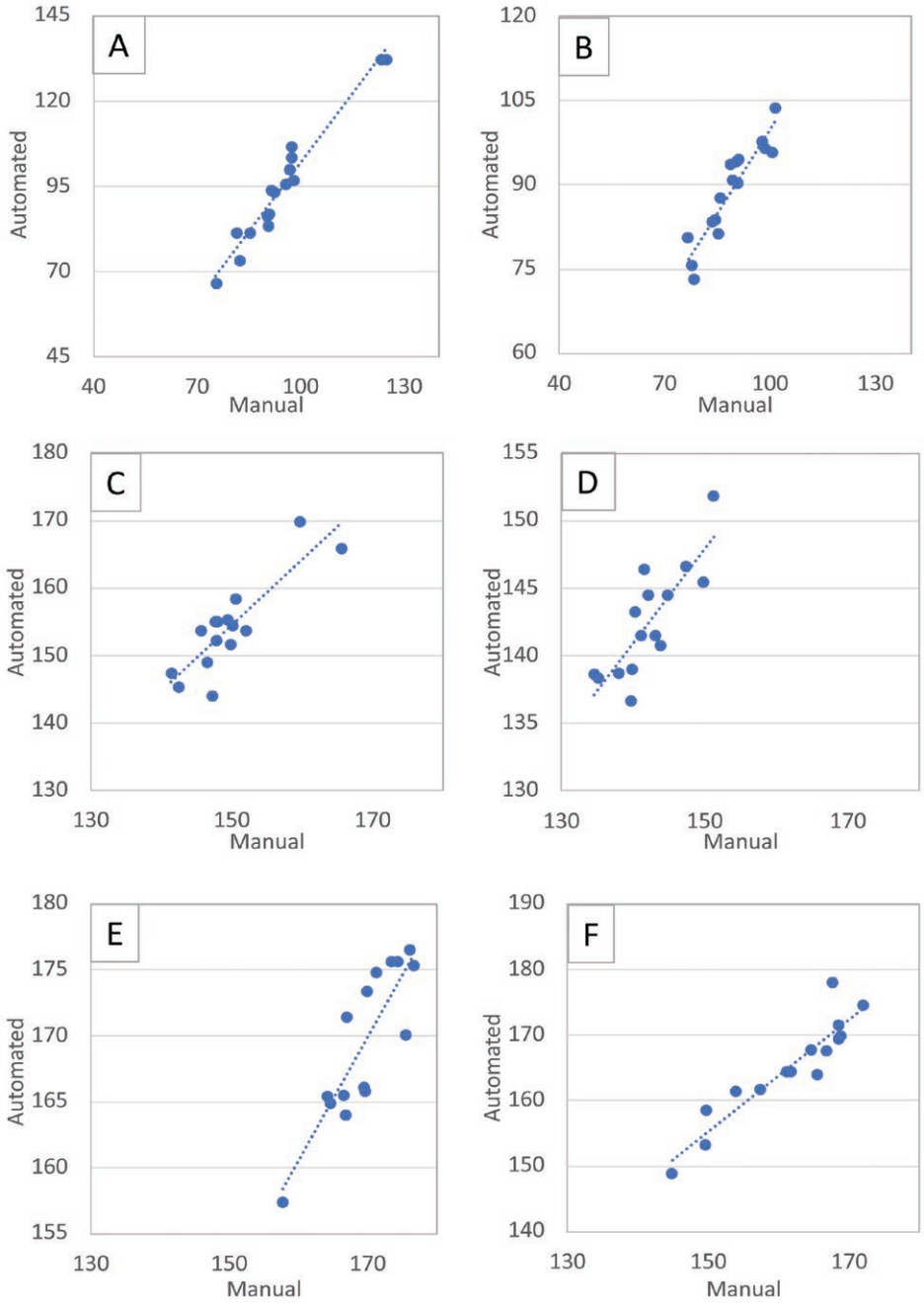

Figure 11. Bland–Altman plots for each knee and hock angle (Hock-Left, Hock-Right, Side-Back-1, Side-Back-2, Side-Front-1, Side-Front-2). The plots visualize the differences between pair of automated and manual points against the mean of each pair. The middle bold line represents the overall mean. The top and bottom dashed lines represent the limits of agreement, set at 1.96 standard deviations above and below the mean.

DISCUSSION

A statistical analysis of the angle measurements obtained through manual evaluation and the trained object detection model found four major findings that indicate that an automated approach can be used to obtain reliable angle measurements of the hock and knee that can be useful in gilt replacement selection procedures to replace less reliable subjective methods currently used to evaluate breeding herd replacement females and ultimately be used to evaluate both elite breeding boars and sows.

Finding 1: The automated approach resulted in feature extractions that were as accurate as human evaluators. As seen in Table 5 and Fig. 11, according to the Bland–Altman plots, all automated measures are in statistical agreement with those obtained manually. In five of the six angles, one or two points approached the border, however, decreasing the threshold to one standard deviation did not make significant changes to the number of points outside the threshold, indicating that these results are robust.

Finding 2: The Bland–Altman plot is a more reliable indicator of statistical agreement than the correlation coefficient. Correlations between the automated and manual measures are displayed in Table 4 and Fig. 12a–f. Many studies of this nature use the simple Pearson correlation coefficient to

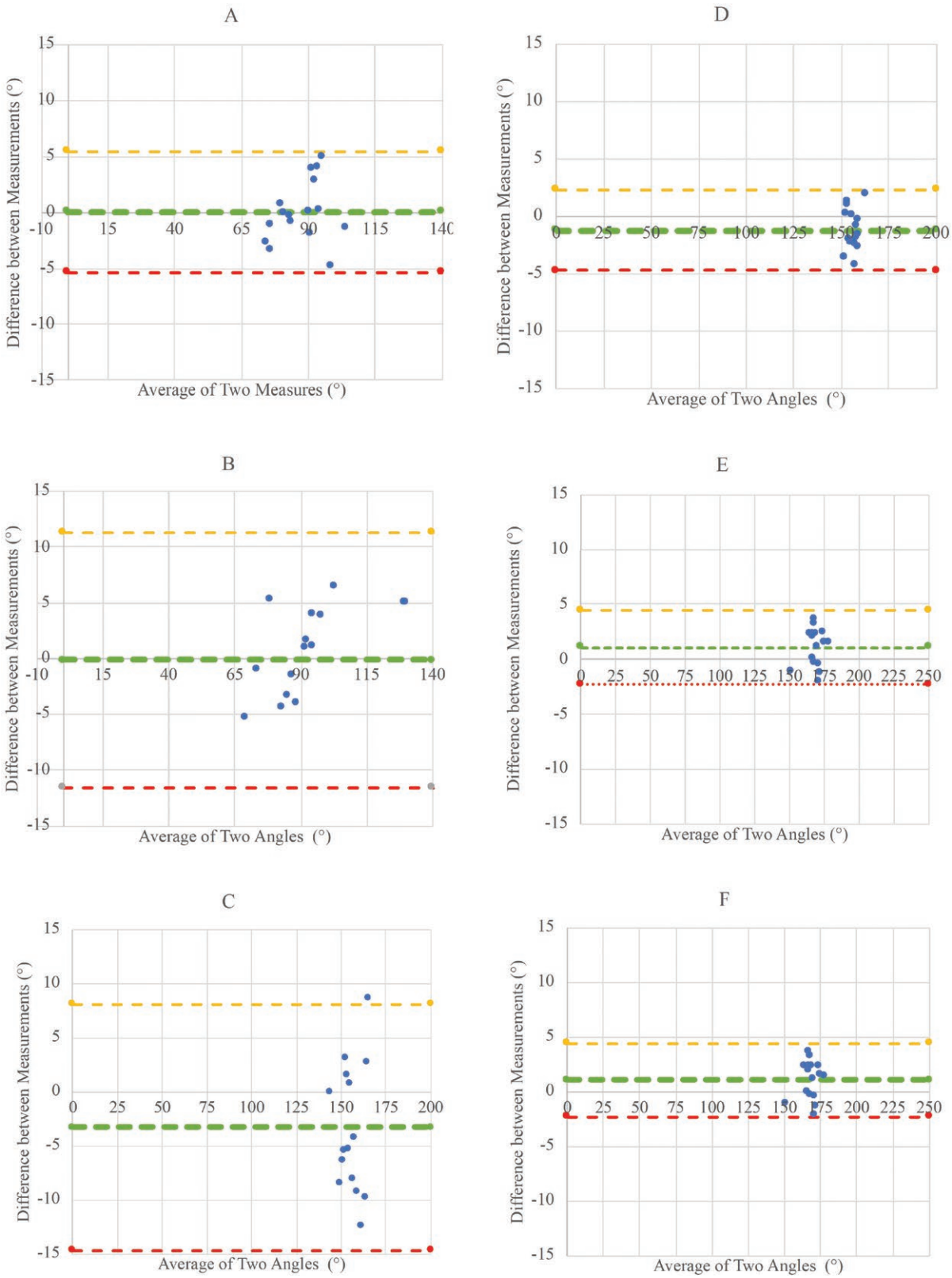

Figure 12. Correlations between automated and manual measurements (in degrees) by knee and hock angle (Hock-Left, Hock-Right, Side-Back-1, Side-Back-2, Side-Front-1, Side-Front-2). Correlations and the corresponding angle are depicted in the upper right corner.

determine statistical agreement, however this poses two main disadvantages. First, the correlation coefficient measures the strength of a relationship between the automated and manual measures but does not imply statistical agreement or cause and effect relationships. One observation concluded that dividing the obtained automated measures by a factor of 0.5 yields the same correlation as if it was multiplied by a factor of two. Regardless of the same correlation, both conversions

of the automated results are not the same as the original, and therefore should not be used as an alternative to manual methods. The Bland–Altman plot uses the difference between each pair of measurements to independently assess agreement between two measures. Therefore, the Bland–Altman plots are able to identify large quantitative differences between the automated and manual measurements, unlike the correlation coefficient. Second, the simple Pearson correlation coefficient is inherently affected by the range of data. This was discovered early in the study, when the back and front legs were originally averaged to yield only two angles from the side view. This reduced the range of side angles from 85° to 95°. As a result, all resulting correlations were found to average less than 0.05, however, the Bland–Altman plots from the averaged angle data suggested that all measurements were still in statistical agreement. These results indicate that correlation can conceal statistical agreement between two measures.

Finding 3: Statistical metrics (ICC and mAP) indicate that the manual evaluation and trained model is reliable for feature extraction in side and rear view images of sows. The ICC of each sow angle is displayed in Table 4 and the mAP's of the two trained models for the side and rear view are displayed in Table 3. These statistical metrics are important in assessing the system used to collect data. The ICC measures how reliable the in-person evaluation is for collecting manual angle measurements, while the mAP measures how accurate the trained object detection model is for collecting automated angle measurements. The ICC across the side and rear images are 0.96 and 0.93 which indicates that the manual measurement system is reliable. The mAP across all detected body parts is 0.94 indicating that the automated measurement system is reliable.

Finding 4: The range of joint angle measurements of the hock and knee are similar to those found in previous studies. The hock and knee joints in sows have a normal biological range of motion. One genetic study compared the range of knee and hock angles from divergent lines for leg weakness (Draper et al., 1988). This study found that across the divergent lines, the knee joint angles range from 167.8° to 174.5°, while the hock angles range from 142.7° to 151.0°. Another study by Stock et al. (2017) found that across different parities, the least square means of the knee and hock joint angle of standing crossbred, multiparous sows, ranged from 157.6° to 161.9° and 142.3° to 151.5°, respectively. The range of hock and knee joint angles from both Draper et al. (1988) and Stock et al. (2017) are similar to the range of angles found in this study (Fig. 12).

The performance and robustness of the trained model can be improved by increasing the variation of unique images of pigs. In this study, only 45 multiparous sows were photographed due to resource constraints and challenges associated with live data collection. All pigs used in this study were cross bred from two genetic suppliers and were consistently a similar color (off white) and surface texture (light colored pelage). Therefore, this model would have similar performance when deployed on images of pigs of different breeds that had similar observable visual characteristics as the pigs used in the training dataset, such as the Landrace, Yorkshire, or Large White. However, this model would perform poorly when deployed on images of pigs that are of different color and skin texture, such as the Duroc, Hampshire, Berkshire, or Large Black. This is because these pigs have not been introduced to the model in the training dataset and will likely be unable to identify the pig when deployed on images of these breeds. Increasing the frequency of pigs with different colors can help train the model to recognize new colors and textures of pigs which will expand the model's ability to identify key body landmarks on pigs of differently colored breeds.

Another observation found that while overall mAP of the trained model was moderately high, the neck consistently received the lowest mAP out of the detected body landmarks, likely due to overfitting of neck. This is likely due to ambiguity and variability of the neck shape and is reflected in the precision recall curves previously discussed in the methods (Fig. 6). These curves reveal that the neck followed an inconsistent pattern with the other landmarks, which may have affected the results of the correlation and Bland–Altman tests for S-F-2. It is possible that this error is a result of overtraining a dataset limited in variety of sow feeding stances. In the training and validation datasets, nearly all pigs are in similar feeding positions, where the neck is scrunched together as the pig consumed feed. This was done to keep the sow still for raw image data collection. In the test images where the neck was incorrectly labeled, it appeared that the heads of the sows were pointed upwards. Since most images in the training and test data set are in a downward feeding stance, the error associated with neck landmark identification is likely due to overfitting and may be resolved by expanding the dataset to include images of sows in different stances, such as sows with their heads facing up.

Future studies would benefit from increasing the volume of images by introducing images of pigs of different sizes, breeds, and of varying hock and knee angles. Furthermore, all images used in this study were from two barn locations during two photoperiods. Increasing the variation of the background, such as lighting or background complexity can increase the robustness and performance of the model. While there was no record of lameness in the pigs used in the current study, the tool described in this paper can be implemented in future research to measure the hock and knee angles of pigs with different degrees of lameness quickly, objectively, and automatically.

CONCLUSIONS

To summarize, this study shows that a deep learning approach to identifying key body landmarks on rear and side images of sows can be leveraged to objectively obtain hock and knee angle measurements. The hock angles from the rear view are Left (denoted Hock-Left; H-L) and Right (denoted Hock-Right; H-R). The four hock and knee angles from the side view are denoted Side-Back-1 (back side of the back leg, abbreviated S-B-1), Side-Back-2 (front side of the back leg, abbreviated S-B-2), Side-Front-1 (back side of the front leg, abbreviated S-F-1), and Side-Front-2 (front side of the front leg, abbreviated S-F-2). Statistical tests compared the angles derived from the localization output from the YOLO algorithm with the angles obtained from manual evaluators on J-Image.

The trained rear and side view models obtained an overall 0.94 mAP across each body landmarks from the rear and side view images. From the coordinate data obtained from the trained model output, four angle measures derived from the deep learning process were found to be with manually obtained measurements on the same image (H-L, H-R, S-F-1, S-F-2). The slope of H-L is found to be statistically similar to 1, indicating a

strong correlation between the manual and automated H-L angles. A Bland–Altman test compared the angles obtained from the deep learning process and angles obtained from manual evaluators for each individual sow and found agreement across all six hock and knee angles (H-L, H-R, S-B-1, S-B-2, S-F-1, S-F-2). This comparison test has two key advantages over using correlations, (1) correlation does not change when scaling one set of measures, and (2) correlation is affected by narrowing the range of values. In summary, interpreting the Bland–Altman test indicates that the automated angle measures agree with manually derived angles and is recommended for comparing future automated measurement techniques with subjective manual measurement systems. Overall, these results will be of interest to swine breeders and veterinarians interested in integrating hock and leg angles with replacement gilt selection and for use in genetic breeding programs where there is a focus on improved feet and leg soundness.

LITERATURE CITED

Altman, D. G., and J. M. Bland. 1983. Measurement in medicine: the analysis of method comparison studies. The Statistician 32:307–317. doi:10.2307/2987937.

Bereskin, B. 1979. Genetic aspects of feet and leg soundness in swine. J. Anim. Sci. 48:1322–1328. doi:10.2527/jas1979.4861322x.

Bland, J. M., and D. G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1:307–310.

Draper, D., M. F. Rothschild, and S. Goedegebuure. 1988. Effects of divergent selection for leg weakness on angularity of joints in Duroc Swine. J. Anim. Sci. 66:1636–1642. doi:10.2527/jas1988.6671636x.

Fan, B., S. K. Onteru, B. E. Mote, T. Serenius, K. J. Stalder, and M. F. Rothschild. 2009. Large-scale association study for structural soundness and leg locomotion traits in the pig. Genet. Sel. Evol. 41. doi:10.1186/1297-9686-41-14

Fisher, R. A. 1950. Contributions to mathematical statistics. Wiley.

Grindflek, E., and E. Sehested. 1996. Conformation and longevity in Norwegian pigs. Proceedings of the Nordiska Jordbruksforskares Ferning Seminar 265:28–33.

Koning, G. 1996. Selection in breeding programs against leg problems. Proceedings of the Nordiska Jordbruksforskares Ferning Seminar 265:85–87.

Main, D. C., J. Clegg, A. Spatz, and L. E. Green. 2000. Repeatability of a lameness scoring system for finishing pigs. Vet. Rec. 147:574–576. doi:10.1136/vr.147.20.574.

Morrow, C., M. F. Rothschild, and D. Draper. 1991. Analysis of gait parameters in Duroc swine genetically divergant for front-leg structure1. J. Anim. Breed. Genet. 108:280–289. doi:10.1111/j.1439-0388.1991.tb00186.x.

Nakano, T., J. J. Brennan, and F. X. Aherne. 1987. Leg weakness and osteochrondrosis in swine: a review. Can. J. Anim. Sci. 67:883–901. doi:10.4141/cjas87-094.

Rasband, W. S. 2018. ImageJ. U. S. Nat. Inst. Health. 9:671–675. https://imagej.nih.gov/ij/.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: unified, real-time object detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. p. 779–788. doi:10.1109/CVPR.2016.91.

Reiland, S., N. Ordell, N. Lundeheim, and S. Olsson. 1978. Heredity of osteochondrosis, body constitution and leg weakness in the pig. A correlative investigation using progeny testing. Acta Radiol. Suppl. 358:123–137.

Rothschild, M. F., and L. L. Christian. 1988. Genetic control of front-leg weakness in Duroc swine. I. Direct response to five generations of divergent selection. Livest. Prod. Sci. 19:459–471. doi:10.1016/0301-6226(88)90012-7.

Serenius, T., M. L. Sevón-Aimonen, and E. A. Mäntysaari. 2001. The genetics of leg weakness in Finnish Large White and Landrace populations. Livest. Prod. Sci. 69:101–111. doi:10.1016/s0301-6226(00)00260-8.

Serenius, T., and K. J. Stalder. 2004. Genetics of length of productive life and lifetime prolificacy in the Finnish Landrace and Large White pig populations. J. Anim. Sci. 82:3111–3117. doi:10.2527/2004.82113111x.

Stalder, K. J., R. C. Lacy, T. L. Cross, and G. E. Conatser. 2003. Financial impact of average parity of culled females in a breed-to-wean swine operation using replacement gilt net present value analysis. J. Swine Health Prod. 11:69–74.

Stalder, K. J., R. C. Lacy, T. L. Cross, G. E. Conatser, and C. S. Darroch. 2000. Net present value analysis of sow longevity and the economic sensitivity of net present value to changes in production, market price, feed cost, and replacement gilt costs in a farrow-to-finish operation. Prof. Anim. Sci. 16:33–40. doi:10.15232/s1080-7446(15)31658-2.

Stock J. D., J. A. Calderón Díaz, C. E. Abell, T. J. Baas, M. F. Rothschild, B. E. Mote and K. J. Stalder. 2017. Development of an objective feet and leg conformation evaluation method using digital imagery in swine. J. Anim. Sci. Livest. Prod. 1:2.

Van Steenbergen, E. J. 1989. Description and evaluation of a linear scoring system for exterior traits in pigs. Livest. Prod. Sci. 23:163–181. doi:10.1016/0301-6226(89)90012-2.

Wood, C. M. and M. F. Rothschild. 2001. Feet and leg soundness in swine: Pork Ind. Handbook. PIH–101.