Computers and Electronics in Agriculture 234 (2025) 110206

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Redefining lameness assessment: Constructing lameness hierarchy using crowd-sourced data

Kehan Sheng [a], Borbala Foris [a], Marina A.G. von Keyserlingk [a], Tiffany-Anne Timbers [b], Varinia Cabrera [a], Daniel M. Weary [a,*]

[a] Animal Welfare Program, Faculty of Land and Food Systems, The University of British Columbia, 2357 Main Mall, Vancouver, British Columbia V6T 1Z6, Canada
[b] Department of Statistics, The University of British Columbia, 2207 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada

ARTICLE INFO

Keywords:
Prevision livestock farming
Early detection
Locomotion score
Animal welfare
Cattle
Computer vision

ABSTRACT

Lameness causes pain to dairy cows and economic losses to farmers, but can be difficult to detect and routinely monitor. Despite numerous attempts to develop automatic detection methods, few have been successfully applied on farms. The development of reliable automated methods is likely restricted by the lack of large, labeled training datasets that capture the diversity in lameness cases within and among farms. Additionally, conventional gait scoring methods employed for annotating training videos are subjective and unreliable, adding noise to training data and thus hindering model performance. We propose a novel approach to lameness assessment in dairy cows, leveraging crowd-sourced data to construct a lameness hierarchy using the Elo-rating method. In this pilot study using 30 cow videos, our proposed lameness hierarchy constructed from pairwise lameness assessments achieved high inter-observer reliability (intraclass correlation coefficient (ICC) = 0.81) among experienced assessors. In contrast, we found that the traditional, subjective gait scoring systems to be inconsistent, with intra- and inter-observer reliabilities of ICC = 0.62±0.09 and 0.44±0.02, respectively. We also demonstrated feasibility for the pairwise assessment to be executed by untrained assessors (in this case, crowd workers recruited via Amazon MTurk), evidenced by high agreement between hierarchies generated by crowd workers and experienced assessors (ICC = 0.85). We created a subsampling algorithm, and found that recruiting just 8 crowd workers per video pair was sufficient to construct a reliable lameness hierarchy. This method also decreased the number of pairwise comparisons required by 61 %, relative to evaluating all possible comparisons between every pair of cows. We conclude that our proposed lameness hierarchy method, using easily accessible crowd workers to facilitate quick and accurate labeling of lameness videos, enables a reliable and granular evaluation of lameness. We suggest that this approach be used to create large training datasets suitable for developing robust lameness detection models.

1. Introduction

Lameness in dairy cows is a global issue, with a mean prevalence of 23 % and ranging from 5 % to 45 % across studies (reviewed by Thomsen et al., 2023). Lameness is also underestimated on farms (Denis-Robichaud et al., 2020), due in part to infrequent manual assessments that miss lameness cases and delay detection and treatment. Sahar et al. (2022) found that 69 % of cows were lame when assessed every 2 weeks for a 12-week period, highlighting the value of more frequent assessments, but implementing more frequent assessments using the current gait scoring methods is impractical as these require a trained expert (Sprecher et al., 1997; Brenninkmeyer et al., 2007). The subjective, time-consuming, and labor-intensive nature of these methods underscores the value of developing automated assessment methods (Alsaaod et al., 2019; Kang et al., 2021). Progress on automation shows promise, with some studies reporting high accuracy (e.g., 97 % accuracy: Jiang et al., 2022; 98 % mAP: Kang et al., 2022; 99 % accuracy: Li et al., 2023). However, commercial application has been limited (Alsaaod et al., 2019; Kang et al., 2021), likely due to the lack of affordable systems that can achieve high accuracy in real-farm environments (Kang et al., 2021; Swartz et al., 2024).

One potential reason for reduced model performance is that the models have been trained using data labeled through subjective gait scoring methods with low reliability (Channon et al., 2009). Some

* Corresponding author.
E-mail address: dan.weary@ubc.ca (D.M. Weary).

https://doi.org/10.1016/j.compag.2025.110206
Received 2 April 2024; Received in revised form 20 February 2025; Accepted 26 February 2025
Available online 6 March 2025
0168-1699/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

studies have labeled training data using only one assessor providing a single assessment for each cow (Abdul Jabbar et al., 2017), but even when multiple assessors are used inter-observer reliability is often not reported (Jiang et al., 2022). Unreliable training data often results in degraded model performance. (Ying, 2019). Moreover, most models have been developed using a limited number of training cases from a single farm (e.g., just 5 cows in Shrestha et al., 2018; 23 cows in Abdul Jabbar et al., 2017), likely insufficient for developing robust models.

To address these limitations, collecting extensive training data from multiple farms with standardized labeling is essential. Advancement in computer vision in recent years rely heavily on crowd-sourcing (Deng et al., 2009). For example, ImageNet, with over 15 million labeled images across 22,000 categories, is one of the most used datasets for training computer vision models (Shin et al., 2016). ImageNet was labeled by crowd workers on Amazon Mechanical Turk (MTurk) identifying objects in images. Applying a similar approach to lameness video labeling poses unique challenges. Unlike simple object identification, lameness assessment traditionally relies upon trained experts (Brenninkmeyer et al., 2007). Accessing and compensating these experts can be difficult and costly.

We propose redefining lameness assessment to enable simple and rapid assessments by untrained crowd workers. Research shows that pairwise lameness assessment (i.e., assessing which cow is more lame when presented with 2 cows walking side by side) is more reliable than absolute scoring (Gardenier et al., 2021), and can be effectively performed by crowd workers (Sheng et al., 2023). Here we propose that results from pairwise comparisons can be integrated to develop a lameness hierarchy, ranking cows from the most healthy to the most lame, akin to the construction of social dominance hierarchy among animals (Neumann and Kulik, 2020).

The primary aims of this study were to develop the lameness hierarchy methodology to label training data for automatic lameness detection in a more granular, cheap and efficient way. Specifically, we begin by assessing the reliability of traditional lameness assessment method, followed by the development of a lameness hierarchy based upon pairwise lameness assessments by both experienced assessors and crowd workers recruited from MTurk. We also explored an algorithm aimed at minimizing the total number of pairwise comparisons required, as well as reducing the number of workers needed for each comparison, while still providing a reliable hierarchy.

2. Materials and methods

2.1. Video materials

We conducted this study from March to June 2021 at The University of British Columbia (UBC) Dairy Education and Research Centre in Agassiz, BC, Canada. The UBC Animal Ethics Committee approved the data collection and experimental procedure under Protocol A19-0299. Each week, we captured side-view videos of lactating Holstein dairy cows walking individually, filming once in the morning and once in the afternoon on the same day, to guarantee at least one high-quality video for each cow in the group. We amassed a total of 1046 videos from 85 cows, averaging 12.31±7.72 (mean ± SD) videos per cow. In this pilot study, we chose 30 high-quality videos featuring distinct cows to test the feasibility of constructing a lameness hierarchy and the efficiency of various hierarchy construction algorithms. Evaluating these algorithms required creating a fully connected graph by comparing each cow with every other cow. Increasing the number of cows would lead to quadratic growth in pairwise comparisons, significantly raising the cost for crowd workers at this stage. These videos were selected with the aim of having 37 % of the videos depicting lame cows, reflecting the typical rate found in commercial dairy herds in North America (von Keyserlingk et al., 2012). This selection was based on live lameness assessments conducted by an experienced assessor on the same day the videos were recorded, using a 5-level gait scoring system, where a gait score greater than 2 is considered lame (Flower and Weary, 2006). To facilitate smooth online accessibility under varying internet conditions, these videos were compressed to 30 % of their original size. Subsequently, they were uploaded to Amazon Simple Storage Services (S3). S3 then generated a unique URL for each video, which we incorporated into an HTML file to provide easy access to the videos. The entire workflow pipeline used for constructing the lameness hierarchy is shown in Fig. 1.

2.2. Gait scoring

We recruited 5 experienced assessors; each having received formal training in lameness assessment and having published at least one peer-reviewed research paper on cattle lameness. Each assessor scored the same set of 30 cows using a 5-level gait scoring system (Flower and Weary, 2006). Gait scoring was completed online using a custom-web interface (Fig. 2). To assess inter- and intra-observer reliability, each assessor evaluated each cow in 3 rounds of assessments, resulting in a total of 15 assessments for each video. Time between successive rounds varied from 3 to 54 days depending upon assessor availability.

The average gait score for each cow was determined based upon all 15 scores. We utilized the "irr" package R version 4.2.3 (R Core Team, 2023) to calculate the Intraclass Correlation Coefficient (ICC), as a means to assess both intra-observer and inter-observer reliability. This was done using a two-way random effect model that focuses on agreement (Hallgren, 2012). To gauge inter-observer reliability, calculations were performed 3 times, encompassing all 3 rounds of gait scoring among the 5 assessors.

In addition, to determine the optimal number of experienced assessors and rounds needed for a reliable evaluation of gait score, we incrementally sampled between 1 and 5 assessors and progressively sampled 1 to 3 rounds of scoring per assessor. Spearman correlation coefficients were then calculated between the average score of each cow derived from the subsampled set and that obtained from the complete set. Under the same number of assessors being subsampled, different combinations of individual assessors were possible (e.g., it could be assessor A, B; or assessor B, C when sampling 2 assessors). We calculated the mean and standard error of Spearman correlation coefficients across all unique combinations of assessors and rounds.

2.3. Pairwise lameness assessment

2.3.1. Assessment interface

We created an interface in HTML5 for conducting online pairwise lameness assessments, wherein videos were configured to auto-play in a loop, depicting 2 cows walking simultaneously in opposite directions (Fig. 3). Users were required to evaluate which cow was more lame within each pair, selecting 1 of the 7 available bubble buttons (Fig. 3). To streamline data analysis, we assigned integer values ranging from −3 to + 3 to represent the bubble buttons from left to right, with 0 indicating that the two cows were similar in lameness.

2.3.2. Pairwise lameness assessment by experienced assessors

Each of the 30 cows was compared with every other cow (following Harary, 1962), resulting in n = 435 (calculated as $\frac{30*29}{2}$) video pairs. To evaluate the quality of responses, we incorporated two attention checks: 1 positive and 1 negative. The positive attention check was derived from one of the 435 video pairs, selected where there was a clear lameness difference between the 2 cows (the absolute difference in the average gait scores between the two cows was 1.27). This left 434 video pairs as test pairs. The negative attention check was devised by creating a mirrored version of a randomly selected video. We created 54 tasks, each task was structured to have 1 positive and 1 negative attention check, along with 8 test pairs randomly selected from all 434 test pairs. An exception was made for the first task, which included 10 test pairs to accommodate all test videos within 54 tasks. The sequence of video pairs

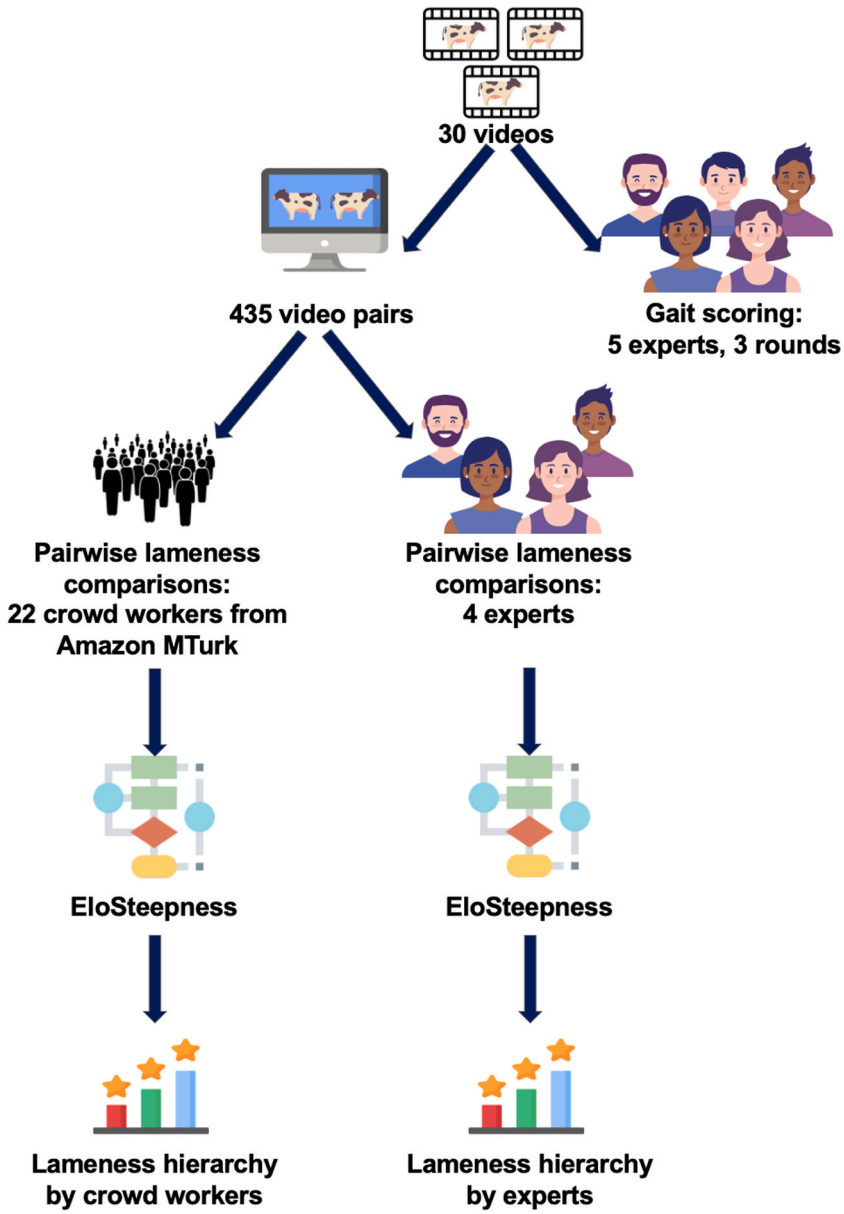


Fig. 1. A workflow pipeline for constructing the lameness hierarchy. First, 5 experts were recruited to conduct 5-level gait scoring on 30 videos, with each expert reassessing the same videos 3 times on different days. We created 435 video pairs for pairwise lameness comparisons. We recruited 22 crowd workers from Amazon Mechanical Turk and 4 lameness experts to assess which cow was more lame in these pairs. Using the EloSteepness algorithm, we constructed two separate lameness hierarchies: one based on crowd workers' assessment results and the other on experts' evaluations.

was randomized within each task.

Building on a script developed in our prior research (Sheng et al., 2023), we employed the Amazon MTurk Application Programming Interface (MTurk, API; Amazon Web Services Inc., 2020) to distribute tasks for pairwise lameness assessment and to collect and manage responses from crowd workers. Amazon MTurk is a widely used online platform that facilitates the acquisition of responses to straightforward tasks, referred to as Human Intelligence Tasks (HITs). This platform allows individuals to register as crowd workers and earn compensation for completing these HITs.

We invited 4 experienced lameness assessors to complete 54 HITs within the MTurk sandbox; this included 1 assessor from the 5 assessors who performed gait scoring as mentioned above, and 3 newly recruited individuals. All assessors had received formal training in lameness assessment, and 3 of them had at least 1 peer-reviewed publication on cattle lameness. The MTurk sandbox functions as a test space, replicating the actual MTurk interface, and allows developers to test their HITs without incurring costs. For each HIT, we set a time limit of 30 min for completion. It took each expert approximately 3 h to finish all HITs. No monetary compensation was provided to these experienced assessors for their participation.

2.3.3. Pairwise lameness assessment by crowd workers

To select crowd workers capable of providing high-quality responses, we first launched a preliminary HIT on MTurk, involving 12 video pairs, and hired 100 workers who held a MTurk Master's qualification. This qualification indicates that the workers have demonstrated high accuracy, a substantial approval rate, and have completed a significant number of HITs in the past. Each worker received $1.2 USD for completing this test HIT, which typically took about 12 min. The choice of these 12 video pairs was informed by the input from experienced lameness assessors after they had finished all 435 pairwise lameness assessments. The selected pairs included 6 that were easily distinguishable in terms of lameness (with one cow being, on average, at least

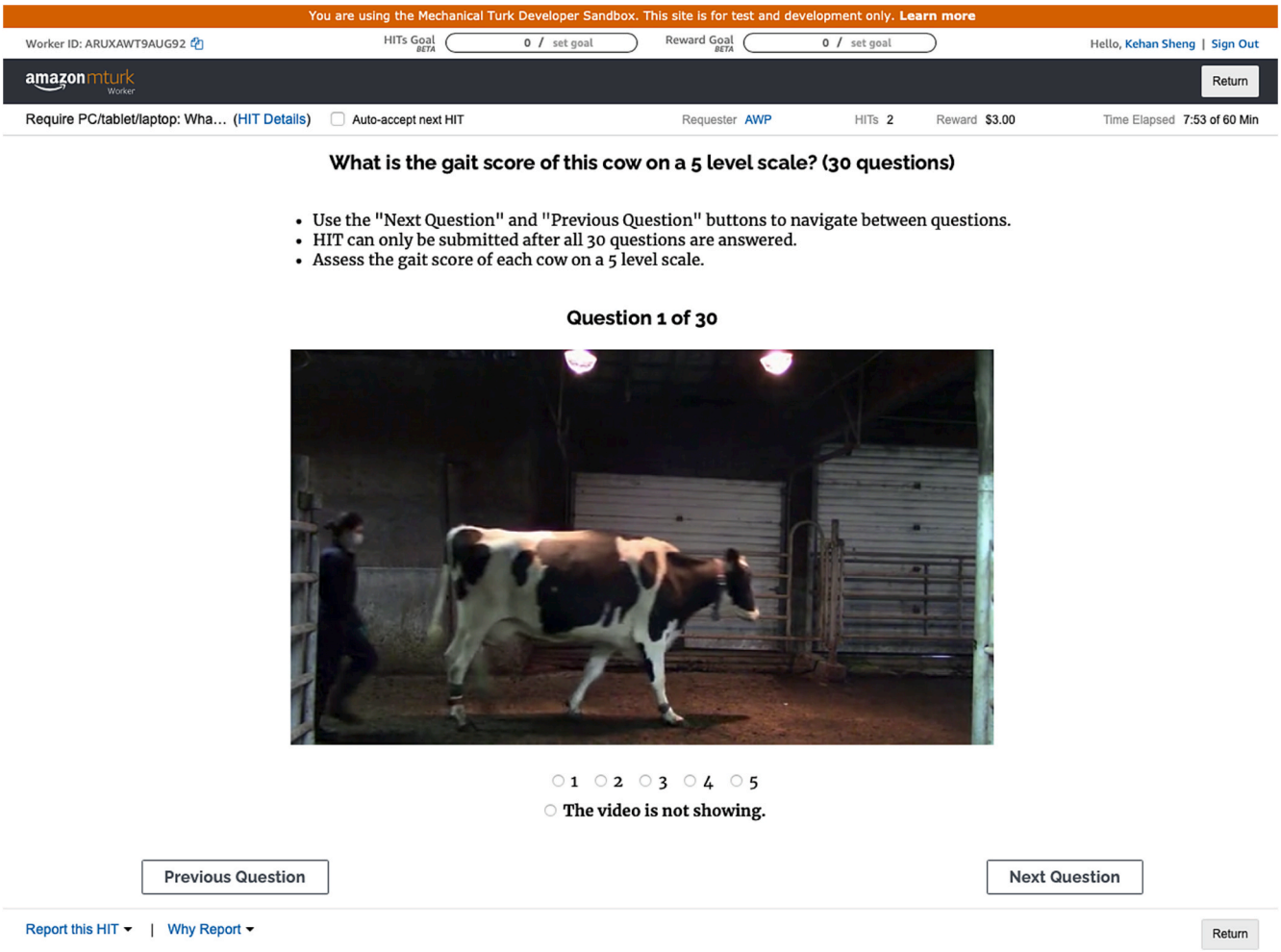


Fig. 2. The online interface used by experienced assessors to conduct 5-level gait scoring for 30 cows. The videos were configured to auto-play continuously in a loop.

2 degrees more lame than the other based on the 7 bubble scale shown in Fig. 3), and another 6 pairs where the lameness differences was moderately clear (with one cow being 1.5 degrees more lame than the other).

Within 7 h, 100 crowd workers had accepted and finished the preliminary HIT. From the responses received, we identified 22 out of 100 crowd workers who achieved 100 % accuracy across 12 video pairs in discerning which cow was more lame. These 22 workers were subsequently awarded a qualification ("lameness hero"), authorizing them to proceed with the 54 test HITs.

In a manner analogous to the formulation of pairwise lameness assessment tasks for experienced assessors, we created 54 HITs. However, unlike before, the positive and negative attention checks were not uniform across all 54 HITs. Instead, we selected 3 easily distinguishable video pairs from the pool of 435 to serve as positive attention check questions. This left 432 video pairs as test pairs. Each HIT randomly included 1 of these 3 positive checks, 1 negative check created by mirroring a randomly selected video from the 30 individual cow walking videos, and 8 test video pairs chosen randomly from all 432 test pairs. The order of video pairs within each task was randomized.

We launched the 54 HITs on MTurk and made them accessible exclusively to workers holding the "lameness hero" qualification. All submitted HITs were automatically approved, and we compensated workers $1 USD for each HIT, which typically took about 10 min to finish. These HITs were listed on the market for 30 d; by the end of this period, the number of workers who completed each HIT varied, ranging from 14 to 20 workers per HIT. 17.9±1.25 (average ± SD) workers completed each HIT.

2.3.4. Pairwise lameness assessment analysis

After obtaining pairwise responses from 4 experienced assessors and the 22 crowd workers, we conducted data cleaning and analysis in Python 3.11 (van Rossum and Drake, 2009) and R version 4.3.0 (R Core Team, 2023).

Our previous research indicates that while inter-observer reliability among crowd workers for pairwise assessments can be low, their collective average responses tend to closely match those of expert assessors (Sheng et al., 2023). Consequently, our analysis focused on comparing the mean responses of the crowd workers with those of experienced assessors using the Intraclass Correlation Coefficient (ICC). We did not calculate inter-observer reliability in pairwise lameness assessments among different crowd workers or among various experienced assessors.

We divided the 435 video pairs into two categories: pairs with pronounced lameness differences (where one cow was deemed ≥ 1 degree more lame than the other based on the 7 bubble scale shown in Fig. 3, as per expert consensus in pairwise lameness assessment) and pairs with more subtle differences (where the perceived lameness difference was $<$ 1 degree according to expert consensus). For each of these categories, we calculated the agreement between the average responses of the crowd workers and those of the experienced assessors using ICC.

We implemented the data filtering and clustering techniques outlined in our previous study (Sheng et al., 2023), reaffirming our earlier

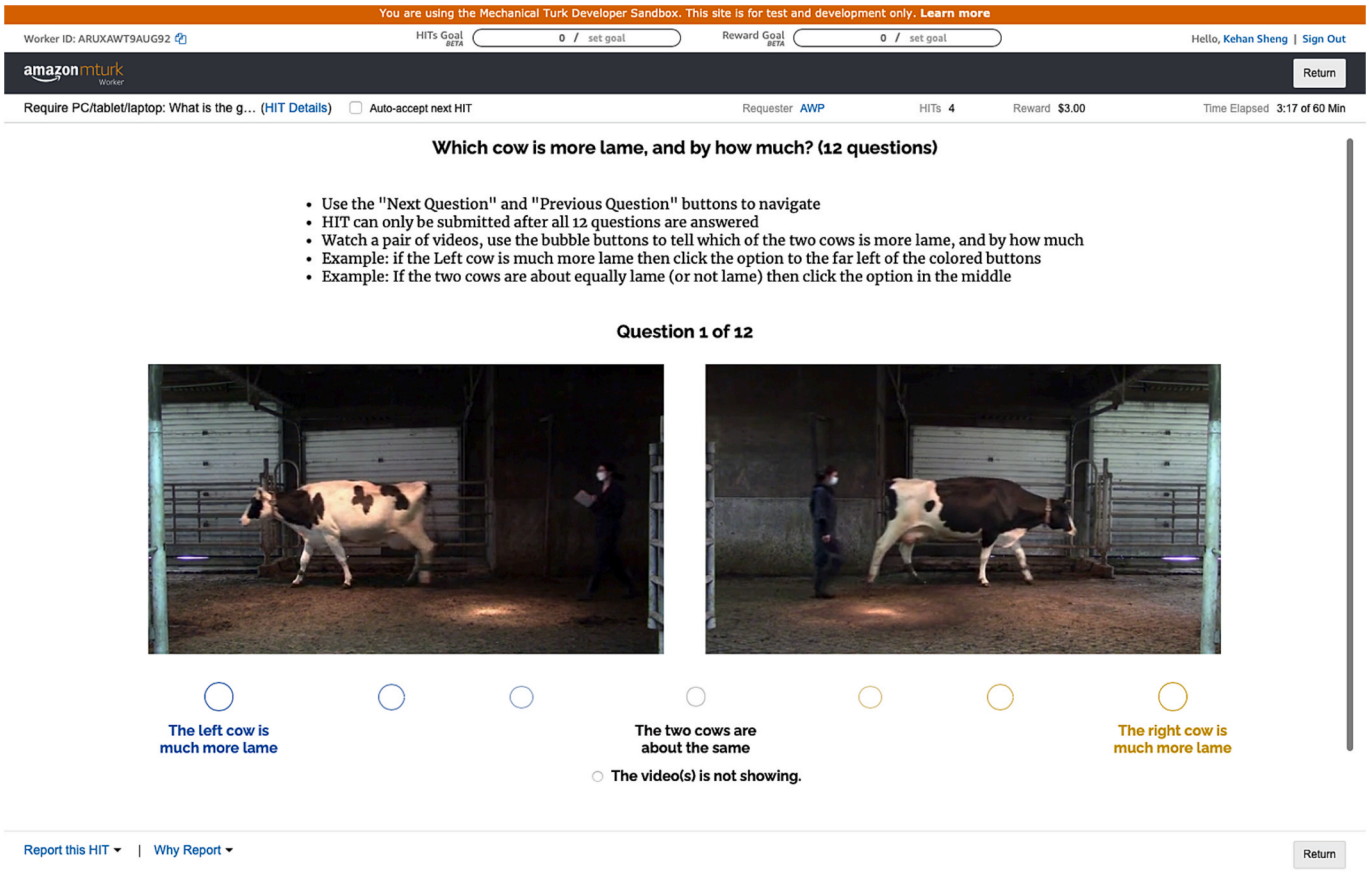


Fig. 3. The online interface used for pairwise lameness assessments. For each pair, assessors were asked to select which cow they perceived as more lame and by how much, selecting one of the 7 bubbles positioned beneath the videos (e.g., the largest blue bubble on the left signified that the left cow was much more lame). Users also had the option to report video not playing properly due to connectivity issues. The videos were configured to auto-play concurrently in a loop, with the two cows walking in opposite directions.

results. Specifically, we found that data filtering methods, including the use of responses only from workers who passed both negative and positive attention checks, along with clustering, did not improve the agreement between the average responses of crowd workers and experienced assessors. Therefore, all further analyses were conducted without the application of additional data filtering or clustering.

2.4. Lameness hierarchy generation

2.4.1. Lameness hierarchy from experienced assessors

To derive a lameness hierarchy serving as the ground truth, we integrated pairwise responses across all 435 video pairs from experienced assessors. It is important to note that one specific video pair was consistently utilized as a positive attention check in all 54 Human Intelligence Tasks (HITs). In order to keep the dataset complete yet free from duplicates, we chose to include only the responses from the initial HIT for this particular video pair.

In our exploratory data analysis, we tested several algorithms to develop a lameness hierarchy. The algorithms we explored included EloSteepness (Neumann and Fischer, 2022), Elo-rating (Neumann et al., 2011), Borda counting (De Borda, 1781), and merge sort (Cole, 1988). We will focus on reporting results from EloSteepness in this paper, as it outperformed the others. Although the hierarchy generated by Elo-rating and Borda counting were similar to that of EloSteepness (SI Appendix, Fig. S2 – S3), the latter was preferred for its ability to provide a probability distribution for each cow's ranking, in contrast to the point-estimates from Elo-rating and Borda counting. Borda counting assigns ranks based on the total number of wins for each cow. In future studies where the comparison graph may not be fully connected (i.e., not all cows are compared against each other), or if the number of assessors for each video pair varies, a cow could attain a higher lameness rank due to more frequent assessments or comparisons, not necessarily due to being more lame. Given the potential variability in the number of assessors per video pair and the difficulty of ensuring an equal number of pairwise comparisons per cow in future studies, we opted for EloSteepness. Merge sort, while computationally efficient and requiring fewer comparisons (about 100 for a group of 30 cows), produced different hierarchies when initial cow ID sequences were altered (SI Appendix, Fig. S4 – S5). This instability requires multiple random initializations to generate a probability distribution of each cows' rank, thus requiring a fully connected graph (435 comparisons for 30 cows) and rendered merge sort impractical for future use.

EloSteepness, initially created for forming animal social dominance hierarchies through pairwise agonistic interactions (Neumann and Fischer, 2022; SI Appendix, Fig. S1), was adapted for the construction of lameness hierarchy. EloSteepness is rooted in Bayesian analysis and provides insights into uncertainty by outlining the probability distribution for each cow's cumulative Elo winning probability. Since initial Elo ratings are not predetermined, EloSteepness estimates the starting Elo rating for each individual. The Elo rating is then updated after each interaction.

For each interaction, the winner's Elo rating increases by $(1-p) \times k$, while the loser's decreases by the same amount. Here, 'p' signifies the winning probability of the winner over the loser based on their Elo rating differences prior to the interaction, and 'k' denotes the maximum potential point shift in a single interaction. Expected wins yield fewer

points for the actor than unexpected victories. EloSteepness computes the final Elo rating for each individual at the end of the interaction sequence. To ascertain an individual's winning probability against any potential competitor, the method uses the formula:

$$p_{AB} = \frac{1}{1 + \exp(Elo_B - Elo_A)}$$

Where p_{AB} is cow A's winning probability against cow B, and Elo_A and Elo_B are their respective Elo ratings. The summed winning probabilities for each individual are the aggregate of their winning probabilities against all potential adversaries.

We used the EloSteepness package (Neumann and Fischer, 2022) to calculate lameness ranks based on the accumulated Elo winning probabilities for each cow, thereby establishing a hierarchy from random sequences of pairwise lameness assessments. In our case, a cow was identified as the 'winner' if deemed more lame than her counterpart in a pairwise comparison, making the other cow the 'loser'. When both cows were assessed as equally lame, we recorded this as two distinct interactions: one where cow A was the winner, and another with cow B as the winner. In empirical data analysis, we considered the extent of lameness disparity, as assessors had 7 different options to indicate the level of lameness difference (Fig. 3). For example, if cow A was assessed as 2 degrees more lame than cow B, we noted 2 interactions with cow A winning. However, we noted that the hierarchies constructed, whether accounting for the degree of lameness differences or not, were strikingly similar. Therefore, in all subsequent hierarchy construction, we did not factor in the degree of lameness difference.

Initially, we constructed individual lameness hierarchies based on each assessor's responses separately. Then, a hierarchy utilizing responses from all experienced assessors across every video pair was constructed. We computed the ICC among hierarchies from different assessors to represent the overall inter-observer reliability. To validate the lameness hierarchy further and connect it with traditional lameness assessment techniques, we used the Spearman correlation coefficients to compare each cow's average gait score with its cumulative Elo winning probability, considering the average gait score as a benchmark.

2.4.2. Lameness hierarchy from crowd workers

We also established a lameness hierarchy based on responses from crowd workers. ICC was then computed to assess the agreement between lameness hierarchies derived from experienced assessors and crowd workers.

2.5. Subsampling workers and comparisons

Lameness hierarchies were first developed using every possible pairwise comparison for 30 cows. However, assessing every pair may not be necessary to establish a reliable lameness hierarchy, and it is not clear how the number of assessors per video pair impacts the reliability of the hierarchy. We thus created an algorithm aimed at reducing the number of assessors and the number of comparisons required.

The first approach involved using milestone cows for comparison. We selected 5 such milestone cows, representing a range from the healthiest to the most lame, based on their average gait scores from 15 rounds of scoring (i.e., cow 1 with an average score of 1.9; cow 2 with a score of 2.4; cow 3 with a score of 2.9; cow 4 with a score of 3.1; cow 5 of 3.9 (all score using a 1–5 scale)). An algorithm was then developed to compare each of the other 25 cows against these milestones. The comparison process for each cow began with the healthiest milestone and continued through successive milestones until the cow was judged to be less lame than a milestone (assessed on average as at least 1 degree less lame) or until all milestones were evaluated. The outcome of all comparisons was then used to construct a lameness hierarchy using EloSteepness.

To assess the effect of different numbers of crowd workers, we chose a sample size ranging from 1 to 14 workers for each video pair. This decision was based on the variation in the number of workers completing each pair, with 14 being the lowest number of workers who evaluated each video pair. Due to the vast number of different potential combinations of assessors at each level and the extended processing time required by EloSteepness, it was impractical to evaluate every possible combination of assessors. Hence, we performed 10 iterations of random sampling at each worker count level, from 1 worker (level 1) to 14 workers (level 14). We calculated the Spearman correlation coefficients to assess the agreement between hierarchies derived from the subsampled crowd worker responses and that based on all crowd worker responses. We opted for Spearman correlation coefficients over ICC due to the inherent assumption of observer independence in ICC. This assumption does not hold here, as there is a clear lack of independence between the subsampled crowd workers and the full set of crowd workers. We also calculated ICC between hierarchies derived from the subsampled crowd worker responses and all experienced assessor responses. The mean and standard error of both correlation metrics, and the total number of comparisons used were recorded across the 10 iterations at each level of worker count.

2.6. Data and code sharing

Data and code used in this study are available at: https://github.com/skysheng7/lameness_hierarchy.git, and published on: https://doi.org/10.5683/SP3/QF1VTK.

3. Results

From the 15 rounds of gait scoring (comprising 3 rounds by 5 assessors) conducted by experienced assessors, it was determined that just 3 of the 30 cows assessed were clinically lame using our current conception of this ailment (i.e., gait score of 3 or higher on the 5-level scale; videos are shown in SI Appendix, Movie S1). Average gait score ranged from 2.0 to 3.9 across the 30 cows.

Intra-observer reliability varied among assessors, with ICC values ranging from 0.49 to 0.73, averaging (± SD) 0.62 (± 0.09). We also computed the inter-observer reliability for each of the 3 rounds of gait scoring from the 5 assessors separately. The average inter-observer reliability varied across rounds from 0.42 to 0.47, with a mean (± SD) of 0.44 (± 0.02).

As the number of assessors and scoring rounds increased, the agreement between average gait score from a subset of responses and the full dataset increased (Spearman correlation coefficient r_s ± SE from 0.59±0.03 to 0.97±0.004; Fig. 4). Securing responses from 2 experts, each contributing 3 rounds of scoring, proved to be sufficient. This approach attained a Spearman correlation coefficient $r_s = 0.82 \pm 0.02$ (mean ± SE) when compared with the responses gathered from 5 experts over 3 rounds of scoring.

The agreement of pairwise lameness assessments across all video pairs, between the mean responses of all crowd workers and that of all experienced assessors, yielded an ICC of 0.79. Out of the 435 video pairs, 175 had pronounced lameness differences (where one cow was deemed ≥ 1 degree more lame than the other based on the 7 bubble scale shown in Fig. 3, as per expert consensus in pairwise lameness assessment) between the 2 cows, whereas 260 pairs presented more subtle differences (where the perceived lameness difference was < 1 degree according to expert consensus). For the pairs with pronounced lameness differences, the agreement between crowd workers and experienced assessors was 0.91. For the pairs with subtle differences agreement was just 0.32.

Initially, we created 4 separate lameness hierarchies based on individual assessors' pairwise lameness evaluations. The inter-observer reliability among these hierarchies was high (ICC = 0.81), so we created a combined lameness hierarchy by pooling responses from all expert assessors. To benchmark this new pairwise assessment method against conventional gait scoring, we correlated the combined hierarchy

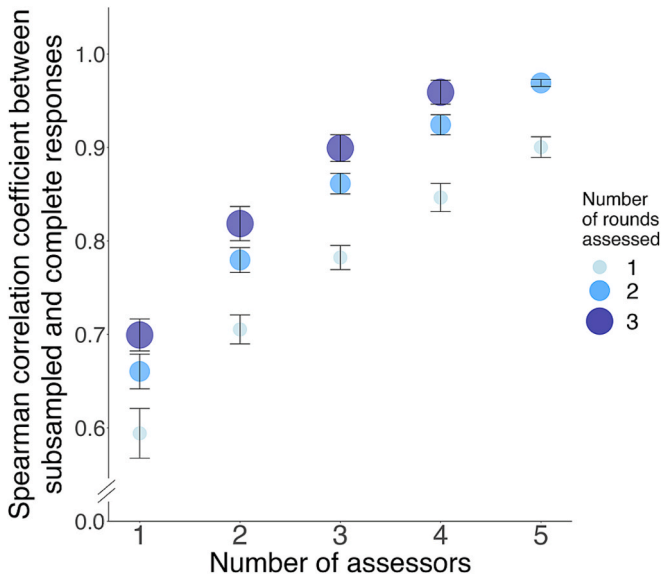


Fig. 4. This plot illustrates how the number of assessors and the number of assessment rounds influence the Spearman correlation coefficient between subsampled and complete responses. Points are differentiated by color and size to represent 1 to 3 assessment rounds, with larger, darker points indicating more rounds of assessment. Error bars reflect the standard error across all possible combinations of assessors and rounds at each level.

with the average scores from 15 traditional scoring rounds. A strong positive correlation was found between the two methods (Spearman correlation coefficient $r_s = 0.77$, $P < 0.01$; Fig. 5; SI Appendix, Movie S1). Lastly, we generated a lameness hierarchy by combining responses from all crowd workers. The inter-observer reliability between this crowd workers' hierarchy and that from experienced assessors was high, with an ICC of 0.85 (Fig. 6; SI Appendix, Movie S2).

We employed a 5-milestone subsampling algorithm to ascertain the minimum number of workers and comparisons required to establish a reliable lameness hierarchy. With increased number of crowd workers, the agreement with the full set increased (Spearman correlation coefficient $r_s \pm$ SE: 0.70±0.03 to 0.97±0.003), but stabilized once the sampled crowd worker count reached 8 ($r_s \pm$ SE: 0.96±0.004; Fig. 7A). Similarly, the average agreement between hierarchies from a subset of crowd workers and those from all experienced assessors also rose with the number of sampled crowd workers (mean ICC ± SE: 0.69±0.02 to 0.83±0.005), but reached a plateau after sampling 8 crowd workers (mean ICC ± SE: 0.83±0.007; Fig. 7B). The increase of the number of comparisons required corresponded with an increase in the count of crowd workers (average ± SE: 89.4±1.45 to 113.0±0.54 comparisons), yet this number stabilized after the crowd worker count reached 6 (average ± SE: 111.7±0.76 comparisons; Fig. 7C).

Our subsampling algorithm reduced the number of necessary comparisons. For a set of 25 cows (with the other 5 cows of the 30-cow dataset used as reference points), the maximum number of comparisons needed was 116, achieving a reduction of at least 61 % compared to using all possible pairwise comparisons. To optimally balance the number of workers and comparisons while ensuring agreement with hierarchies formed from both all crowd worker responses and all experienced assessor responses, we recommend recruiting a minimum of 8 crowd workers per video pair. This approach would necessitate an average of 111.5±0.75 (average ± SE) comparisons, representing just 37 % of the total possible comparisons.

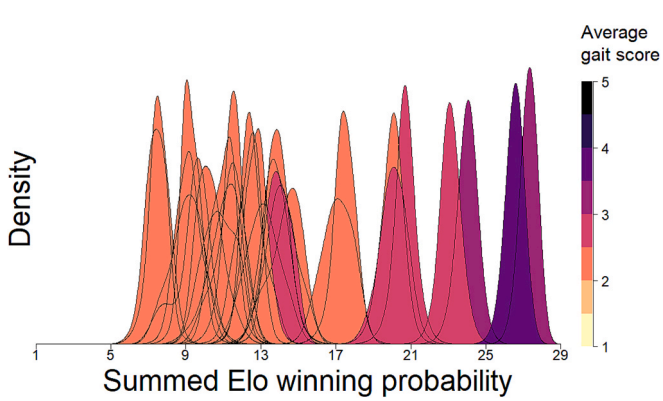


Fig. 5. Cumulative Elo winning probabilities for a group of 30 cows, based on their ranking in the lameness hierarchy. This hierarchy illustrated was established from pairwise lameness assessment by experienced assessors across 435 video pairs. The Elo winning probability represents the likelihood of a particular cow being more lame than any other cow in the group, with a higher total Elo winning probability indicating a more severe level of lameness. The summed Elo winning probability for each cow is shown as a hill distribution, where the color of each hill corresponds to the cow's average gait score (using a 1–5 scale), determined from 15 rounds of scoring by experienced assessors.

4. Discussion

In this study, we introduced a novel method of constructing a lameness hierarchy for lameness assessment, and demonstrated that it was superior in terms of inter-observer reliability compared to traditional gait scoring. Our approach achieved high inter-observer reliability among experienced assessors (ICC = 0.81), and proved to be feasible also for untrained crowd workers, likely due to the simple and intuitive nature of the task. This effectiveness is demonstrated by the high level of agreement between crowd workers from MTurk and experienced assessors in both pairwise lameness assessments and the constructed lameness hierarchies. Furthermore, using a 5-milestone subsampling algorithm, we found that recruiting only 8 crowd workers per video pair and utilizing on average 37 % of all potential pairwise comparisons was sufficient to construct a reliable lameness hierarchy. In contrast, we found low intra- (ICC = 0.49–0.73) and inter-observer reliability (ICC = 0.42–0.47) in traditional gait scoring method, highlighting the subjectivity and inconsistency of this method, echoing findings from previous studies (Channon et al., 2009).

4.1. Lameness data labeled using our algorithm through crowd-sourcing

Transfer learning techniques, such as fine-tuning pre-trained models (e.g., YOLO), have been used in the deployment of lameness assessment models in recent years (Kang et al., 2022). Transfer learning techniques involve adjusting models that have been pre-trained on extensive general datasets to new, specific tasks (Best et al., 2020). Transfer learning excels in applications with limited training data by leveraging knowledge from previously trained models, enabling these applications to achieve higher accuracy without the need for extensive datasets (Best et al., 2020). However, it is important to note that these models are usually trained on general tasks and datasets, which may significantly differ from the specialized application and intricate task of lameness assessment in dairy cows. The generalizability and performance of a pre-trained model will decline as the mismatch between the source domain (i.e., domain used in pre-training) and target domain (i.e., domain of the new task) enlarges (Ding and Fu, 2019).

Although commonly used, pre-training datasets like ImageNet contain images of livestock that may present biases related to livestock farming. Publicly available images of farm animals typically depict them on pasture or in free-range environments, but most farm animals are housed in intensive farming systems (Hagendorff et al., 2023). Models trained on images from animals on pasture may show poor performance when applied to images of animals indoors (Hagendorff et al., 2023). As a result, adapting these models to lameness detection will likely still require a considerable amount of labeled data to achieve an accuracy

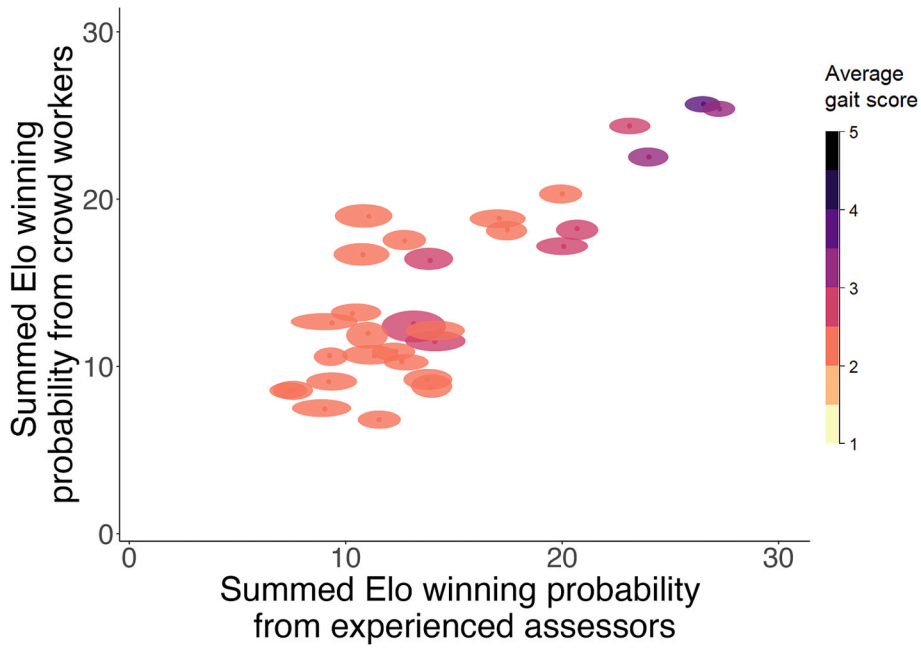


Fig. 6. The relationship between the lameness hierarchy for 30 cows derived from pairwise assessments by experienced assessors versus assessments from crowd workers. The Elo winning probability represents the likelihood of each cow being more lame compared to others in the group, with a higher Elo winning probability indicating greater lameness. Each cow is represented by an ellipse in the plot, where the color corresponds to the cow's average gait score (using a 1–5 scale), determined from 15 rounds of scoring by experienced assessors. The center of each eclipse indicates the average cumulative Elo winning probability for that cow. The horizontal and vertical axes of the ellipse represent the 90% creditable interval of the summed Elo winning probabilities as determined by experienced assessors and crowd workers, respectively.

sufficient for successful application in actual farm settings.

Additionally, lameness assessment is a complex cognitive process involving recognizing subtle changes in body movement patterns (Li et al., 2023). It is essential to label a substantial number of videos to capture the diverse manifestations of lameness in different individuals, as cows can display varying body movement patterns even within the same lameness category (Viazzi et al., 2013). It is thus crucial to train deep learning models with a dataset covering a wide range of lameness features (Li et al., 2023), as opposed to programming the model to focus on just 1 or 2 specific features like back curvature and head bob (Jiang et al., 2022). In addition, the complexity of lameness detection is compounded by the diverse environments in which dairy cows are housed. Variations in flooring—from soft rubber to hard concrete—can influence a cow's gait (Flower et al., 2007). This necessitates a model trained on data that encompasses a wide array of environmental conditions to further enhance the model's performance and adaptability to new environments upon deployment (Kang et al., 2021). Therefore, despite the advantages of transfer learning, there remains a need for labeling large datasets that capture a wide range of scenarios. The lameness hierarchy method we propose enables rapid and accurate labeling of a large volume of lameness data using crowd workers.

Traditional gait scoring methods vary widely from simple 2-level systems (Brenninkmeyer et al., 2007), to more complex 3-level (Van Nuffel et al., 2009), 4-level (Brenninkmeyer et al., 2007), and 5-level scales (Sprecher et al., 1997; Flower and Weary, 2006); these differences pose a challenge in creating an unified lameness database. Establishing a standardized approach to lameness assessment is essential. In this regard, our pairwise lameness assessment approach, which simply determines which cow is more lame, offers a universal solution. We can generate new pairwise lameness comparisons between cows from different datasets. The Elo-rating algorithm can use these new comparison results and previous comparison results between cows in the same datasets to reconstruct a combined lameness hierarchy bridging different datasets, paving the way for the creation of a comprehensive, global database.

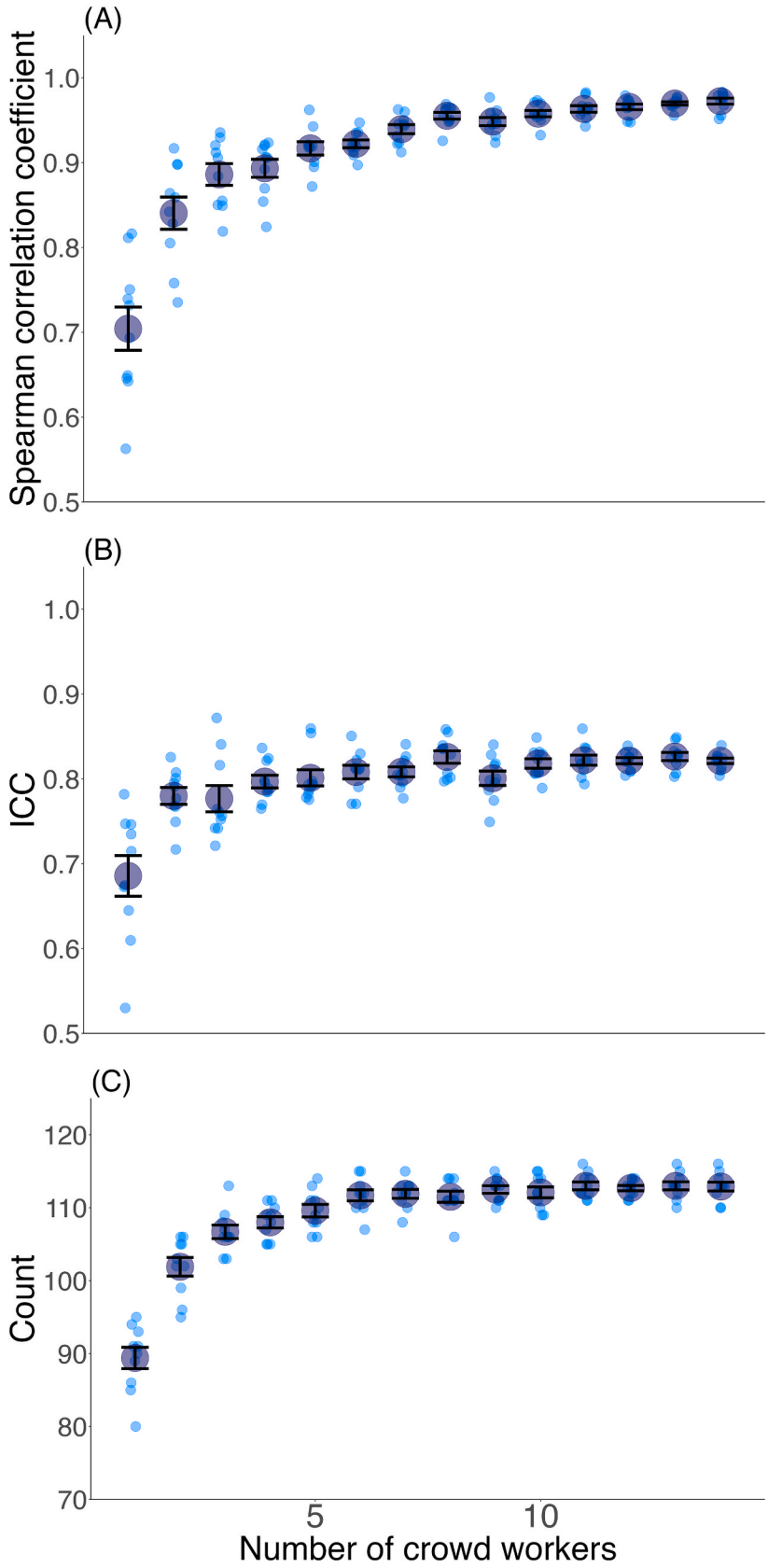
4.2. Lameness data labeled through traditional gait scoring

Previous computer vision models have relied on limited training videos labeled using traditional gait scoring methods, with 1 or 2 assessors scoring each cow just once (Abdul Jabbar et al., 2017; Shrestha et al., 2018). Our findings indicate that such labeling methods are likely to be unreliable; our data showed a Spearman correlation coefficient r_s of only 0.59±0.03 (mean ± SE) for a single assessor and 0.71±0.02 (mean ± SE) for 2 assessors. Furthermore, one-time live lameness assessments are likely to be untrustworthy; as in our trial, an experienced lameness assessor conducted live evaluations on site on the same day when the videos were taken; this assessor had estimated a lameness prevalence of 37 % among the 30 cows selected. Our expert assessments, based on 15 rounds of gait scoring for each cow, indicated a prevalence of just 10 %. This discrepancy suggests that one-time assessments by a single assessor are not suitable as a ground truth in training automatic lameness detection systems (Sahar et al., 2022).

One advantage of using traditional gait scoring to label lameness videos is that each assessor needs to conduct fewer assessments compared to the lameness hierarchy method. For instance, in a herd of 25 cows, each assessor would need to perform 25 absolute scoring assessments for one round, whereas the lameness hierarchy method would require about 116 pairwise comparisons among the 25 cows. However, due to the low intra- and inter-observer reliability of traditional gait scoring, it necessitates multiple assessors and rounds of scoring. Additionally, this method can only be performed by experienced lameness assessors who require extensive training, making it impractical for large-scale labeling tasks.

4.3. Future application

The agreement between expert and crowd worker responses in pairwise lameness assessment documented in the current study (ICC = 0.79) is lower than our previous result using a different sample of cows (ICC = 0.89; Sheng et al., 2023; $P < 0.01$). This difference can be



(caption on next page)

Fig. 7. The graph illustrates the effects of increasing the number of crowd workers on: (A) Spearman correlation coefficients between hierarchies created using subsampled dataset and the full dataset from crowd workers, (B) Intraclass correlation coefficient (ICC) between hierarchies created from subsampled crowd workers and the full set of experienced assessors, (C) the number of pairwise comparisons used. We sequentially sampled between 1 and 14 crowd workers for every video pair, carrying out 10 iterations of random sampling at each worker count level. Small light blue dots illustrate the results from a single random sampling iteration, while the mean across all random iterations is represented by large dark blue dots. Jitter was added to avoid overplotting. The error bars indicate the standard error across all sampling iterations.

attributed to the lower prevalence of clinical lameness in our current dataset (10 %) compared to the earlier dataset (28 %; Sheng et al., 2023). The current dataset consisted of predominantly healthy cows, resulting in more video pairs with subtle lameness differences, reducing the overall ICC; for video pairs with clear lameness differences, the ICC was 0.91 versus just 0.32 for pairs with subtle differences. ICC assesses the proportion of variance in the true score (T) accounted for by the observed score's variance, inclusive of measurement error (E): $\frac{Var(T)}{Var(T)+Var(E)}$ (Hallgren, 2012). For video pairs with subtle lameness differences, $Var(T)$ decreases due to a narrower distribution of lameness differences (i.e., the range of lameness differences for video pairs with subtle differences is $(-1, 1)$ compared to that for video pairs with distinguishable differences ranges $[-3, -1] \cup [1, 3]$), and $Var(E)$ increases due to ambiguity in discerning which cow is more lame. The low prevalence of lameness in our dataset can be viewed as a strength rather than a limitation for testing and evaluating our lameness hierarchy algorithm. In real-world conditions, where lameness prevalence is typically higher (average 22.8 % globally; Thomsen et al., 2023), applying our algorithm should increase the number of video pairs showing easily distinguishable lameness differences, improving agreement between assessors.

Identifying cows at relatively early stages of lameness is a critical challenge, as early detection can yield substantial benefits for cows and farmers, acting as an incentive for investing in automated lameness detection systems (Alsaaod et al., 2019). Experienced assessors often find it challenging to reach a consensus on cows exhibiting subtle signs of lameness and to differentiate between cows with gait scores of 2 and 3 (Channon et al., 2009). However, when applied to the current dataset, our algorithm demonstrated high performance (interobserver reliability ICC = 0.81) in this critical range, where all cows received an average gait score of 2 or higher but less than 4. The lameness hierarchy described in our study positions cows on a continuous scale. This method has the potential to help farmers to identify which cows in their herd may warrant further attention and potential treatment. One approach for identifying these cows is by comparing them to a benchmark cow with a known lameness status and focusing on those ranked as more lame than the benchmark. Alternatively, farmers could opt to assess the 'k' most lame cows within the hierarchy, where 'k' is a flexible number determined by the farmer, depending on their available time and resources.

The approach outlined in this study extends beyond lameness assessment, offering wide-ranging applications across domains. Health-related data labeling typically demands the expertise of veterinarians and other medical professionals. However, our work has shown the potential of using pairwise comparisons to label health-related data by crowd workers. Similar methodologies have been employed for tasks like ranking images based on cell counting (Shah and Wainwright, 2018), and ranking the performance of large language models and text-to-image generative models based on pairwise comparisons of text outputs or image outputs (i.e., Chatbot Arena; Chiang et al., 2024). The concept of creating rankings through pairwise comparisons introduces possibilities for other image labeling tasks.

4.4. Limitations

Our lameness hierarchy, developed through pairwise assessments, shows a strong positive correlation with traditional gait scoring methods, but it is crucial to recognize that these two methodologies are fundamentally different. A perfect alignment between the lameness hierarchy and the rankings based on average gait scores from 15 rounds of scoring was not expected. Pairwise assessments, which align more naturally with human cognitive processes, provide finer granularity compared to absolute scoring systems (Gulliksen, 1946). Differences observed between the rankings derived from the lameness hierarchy and those based on sorted average gait scores of each cow should be interpreted cautiously, as the real ground truth – which arguably could be better determined by augmenting these data with the presence of hoof pathologies and how animals respond to treatment with analgesics (Flower and Weary, 2006) – was not established in this study. Future research could employ the lameness hierarchy approach we have outlined to gain a deeper understanding of how gait changes correlate with hoof and leg health pathologies.

Different types of hoof lesions can alter nociceptive thresholds in different ways, resulting in varying degrees of impact on the locomotion of cows (Whay et al., 1998). For instance, Flower and Weary (2006) found no differences in the gait of healthy cows and those with sole hemorrhages. When lesions are limited to skin inflammation, gait may not be altered at all (Kester et al., 2014). We acknowledge that the lameness hierarchy method may not identify every cow with a visible lesion. However, for cows that do walk abnormally, our method is superior to traditional gait scoring in detecting these abnormalities. It is possible that some cows with visible lesions might not be detected by either traditional scoring methods or our lameness hierarchy. In such instances, visual gait assessment would be inadequate, and alternative approaches, such as monitoring changes in lying, feeding, and social behaviors (Alsaaod et al., 2019), should be considered to identify lame cows that do not exhibit altered gait.

Visual detection of hoof and leg lesions is also prone to error. For example, injuries to the corium can occur 8 to 10 weeks before a lesion becomes visible on the hoof, and it is unlikely that corium injury is still painful when the hemorrhage appears (Flower and Weary, 2006). Therefore, the absence of visible lesions does not mean the cow is not in pain, and the presence of visible lesions does not necessarily indicate pain. It is also important to note that these studies were conducted comparing different cows to detect gait alteration. We encourage future research to use pairwise lameness comparisons described here to compare the gait of the same individual at different time points to capture more nuanced information and potential gait alterations as lesions progress.

Although a link between lesions and lameness has been identified, a causal relationship has not been established (Kester et al., 2014). A review by Kester et al. (2014) concluded that while severe lesions can affect cows' gait, numerous studies suggest that lameness is more likely to precede hock lesions and elevate the risk of lesions, rather than the other way around. This is because lame cows tend to spend more time lying down and have difficulty standing up, which increases their exposure to lying surfaces and the likelihood of collisions with cubicles (Kester et al., 2014). Our proposed lameness hierarchy approach, which excels at identifying cows with abnormal gaits, might be effective in identifying cows at higher risk of developing lesions before they occur.

Furthermore, farmers often identify cows requiring hoof inspection using signs of abnormal gait. Visual inspection of cows' walking, whether conducted manually or automatically, is likely to remain a common method for identifying cows that require further examination. Our proposed lameness hierarchy method offers a more reliable approach to identifying cows with altered gait compared to traditional

gait scoring methods.

An additional limitation of the current study is the way in which crowd workers and experienced assessors responded to our question "how much more lame" during pairwise assessments. Experienced assessors, drawing on their training, may have been more likely to base their responses on specific lameness traits observed in the two cows being compared. In contrast, crowd workers might base their decisions on their confidence level in discerning which cow was more lame, a judgment that might not reflect differences in lameness levels.

We also noted that some cows were ranked higher by crowd workers than by lameness experts due to factors unrelated to lameness, such as slipping, walking slowly, or stopping mid-walk. Future studies should consider removing such videos from the dataset. Developing an automatic, video processing pipeline to sort videos based on quality would be beneficial in ensuring the reliability of data used for lameness assessments (Kang et al., 2021; Kang et al., 2022).

Moreover, the videos for this pilot study were collected from a single farm. Future research should include videos from multiple farms to evaluate the validity of this labeling method across different environments.

5. Conclusion

We described a method of constructing lameness hierarchy based on pairwise lameness assessments results. Our pairwise lameness assessment method provides a reliable alternative to conventional gait scoring for evaluating lameness in dairy cows and proved sufficiently intuitive for execution by crowd workers. The adoption of a 5-milestone subsampling algorithm reduced the number of workers and comparisons required to establish a robust lameness hierarchy. Our method opens up new possibilities for labeling large quantities of lameness videos through crowd-sourcing, laying the groundwork for the creation of datasets that can be used to train models for automated lameness detection. Moreover, the lameness hierarchy provides more granular data than conventional methods and may be applicable to a wide range of medical and health-related domains where expert-based data labeling has been traditionally used.

Declaration of Generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the first author used OpenAI ChatGPT to rephrase portions of the manuscript. After using this tool/service, all authors reviewed and edited the content as needed. Collectively all authors take full responsibility for the content of the publication.

CRediT authorship contribution statement

Kehan Sheng: Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Borbala Foris:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Marina A.G. von Keyserlingk:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Tiffany-Anne Timbers:** Writing – review & editing, Visualization, Supervision, Methodology, Funding acquisition. **Varinia Cabrera:** Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation. **Daniel M. Weary:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to Ruan Daros, Sarah Bolton, Kiyomi Ito, Hanna Eriksson, Anne-Marieke Smid, and Tiago Facury Moreira for volunteering their time in gait scoring and pairwise lameness assessment. We thank Mohammad W. Sahar, Negar Sadrzadeh, and Joseph Krahn from the Animal Welfare Program (AWP), Vancouver BC Canada for collecting videos during this trial. We are also thankful for the undergraduate work-learn students at the Animal Welfare Program, especially Sora Jeong, Kay Yang, and Kanish Khanna for their help in data cleaning and preprocessing. We thank Alexandra Protopopva, Lesley Dampier, Bianca Vandresen, Camila Cavalli, Michael Brunt, Bailey Eagan, Meghan Lok, Emeline Nogues, Katie Koralesky, Seonpil Yoo, Lara Sirovica from the AWP for testing the user interface of the pairwise lameness assessment task. We also thank the students and staff at The University of British Columbia Dairy Education and Research Centre (Agassiz, BC, Canada).

This research was supported by a Natural Science and Engineering Research Council (NSERC; Ottawa, Canada) Discovery grant (RGPIN-2023-04350). General funding for the UBC Animal Welfare Program is provided by the NSERC Industrial Research Chair in Dairy Cattle Welfare together with our industrial partners: the Dairy Farmers of Canada (Ottawa, ON, Canada), Saputo Inc. (Montreal, QC, Canada), British Columbia Dairy Association (Burnaby, BC Canada), Alberta Milk (Edmonton, AB, Canada), Intervet Canada Corporation (Kirkland, QC, Canada), Boehringer Ingelheim Animal Health (Burlington, ON, Canada), BC Cattle Industry Development Fund (Kamloops, BC, Canada), The Semex Alliance (Guelph, ON, Canada), Lactanet (Sainte-Anne-de-Bellevue, QC, Canada), Dairy Farmers of Manitoba (Winnipeg, MB, Canada), and the Saskatchewan Milk Marketing Board (Regina, SK, Canada).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2025.110206.

Data availability

All data and code are freely available via a link available in the paper

References

Abdul Jabbar, K., Hansen, M.F., Smith, M.L., Smith, L.N., 2017. Early and non-intrusive lameness detection in dairy cows using 3-dimensional video. *Biosyst. Eng.* 153, 63–69. https://doi.org/10.1016/j.biosystemseng.2016.09.017.

Alsaaod, M., Fadul, M., Steiner, A., 2019. Automatic lameness detection in cattle. *Vet. J.* 246, 35–44. https://doi.org/10.1016/j.tvjl.2019.01.005.

Amazon Web Services Inc. (2020). Amazon MTurk Application Programming Interface. [Computer software].

Best, N., Ott, J., Linstead, E.J., 2020. Exploring the efficacy of transfer learning in mining image-based software artifacts. *Journal of Big Data* 7 (1). https://doi.org/10.1186/s40537-020-00335-4.

Brenninkmeyer, C., Dippel, S., March, S., Brinkmann, J., Winckler, C., Knierim, U., 2007. Reliability of a subjective lamess scoring system for dairy cows. *Anim Welf.* 16 (2), 127–129. https://doi.org/10.1017/s0962728600031158.

Channon, A.J., Walker, A.M., Pfau, T., Sheldon, I.M., Wilson, A.M., 2009. Variability of manson and leaver locomotion scores assigned to dairy cows by different observers. *Vet. Rec.* 164 (13), 388–392. https://doi.org/10.1136/vr.164.13.388.

Chiang W. L, Zheng L, Sheng Y, Angelopoulos A. N, Li T, Li D, Zhang H, Zhu B, Jordan M, Gonzalez J. E, Stoica I. (2024). Chatbot arena: An open platform for evaluating llms by human preference. arXiv preprint arXiv:2403.04132.

Cole, R., 1988. Paralle Merge Sort. *SIAM J. Comput.* 17 (4), 770–785. https://epubs.siam.org/terms-privacy.

De Borda, J., 1781. Mémoire sur les élections au scrutin. *Histoire de L'Académie Royale Des. Sciences* 102, 657–665. https://doi.org/10.4236/jwarp.2022.147026.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition* 248–255.

Denis-Robichaud, J., Kelton, D., Fauteux, V., Villettaz-Robichaud, M., Dubuc, J., 2020. Short communication: Accuracy of estimation of lameness, injury, and cleanliness prevalence by dairy farmers and veterinarians. *J. Dairy Sci.* 103 (11), 10696–10702. https://doi.org/10.3168/jds.2020-18651.

Ding, Z., Fu, Y., 2019. Deep Transfer Low-Rank Coding for Cross-Domain Learning. *IEEE Trans. Neural Networks Learn. Syst.* 30 (6), 1768–1779. https://doi.org/10.1109/TNNLS.2018.2874567.

Flower, F.C., De Passillé, A.M., Weary, D.M., Sanderson, D.J., Rushen, J., 2007. Softer, higher-friction flooring improves gait of cows with and without sole ulcers. *J. Dairy Sci.* 90 (3), 1235–1242. https://doi.org/10.3168/jds.S0022-0302(07)71612-0.

Flower, F.C., Weary, D.M., 2006. Effect of Hoof Pathologies on Subjective Assessments of Dairy Cow Gait. *J. Dairy Sci.* 89 (1), 139–146. https://doi.org/10.3168/JDS.S0022-0302(06)72077-X.

Gardenier, J., Underwood, J., Weary, D.M., Clark, C.E.F., 2021. Pairwise comparison locomotion scoring for dairy cattle. *J. Dairy Sci.* 104 (5), 6185–6193. https://doi.org/10.3168/jds.2020-19356.

Gulliksen, H., 1946. Paired comparisons and the logic of measurement. *Psychol. Rev.* 53 (4), 199–213. https://doi.org/10.1037/h0061673.

Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology* 8 (1), 23.

Harary, F., 1962. The maximum connectivity of a graph. *Proc. Natl. Acad. Sci.* 48 (7), 1142–1146.

Jiang, B., Song, H., Wang, H., Li, C., 2022. Dairy cow lameness detection using a back curvature feature. *Comput. Electron. Agric.* 194, 106729. https://doi.org/10.1016/j.compag.2022.106729.

Kang, X., Li, S., Li, Q., Liu, G., 2022. Dimension-reduced spatiotemporal network for lameness detection in dairy cows. *Comput. Electron. Agric.* 197, 106922. https://doi.org/10.1016/j.compag.2022.106922.

Kang, X., Zhang, X.D., Liu, G., 2021. A review: Development of computer vision-based lameness detection for dairy cows and discussion of the practical applications. *Sensors* 21 (3), 753. https://doi.org/10.3390/s21030753.

Kester, E., Holzhauer, M., Frankena, K., 2014. A descriptive review of the prevalence and risk factors of hock lesions in dairy cows. *Vet. J.* 202 (2), 222–228. https://doi.org/10.1016/j.tvjl.2014.07.004.

Li, Q., Chu, M., Kang, X., Liu, G., 2023. Temporal aggregation network using micromotion features for early lameness recognition in dairy cows. *Comput. Electron. Agric.* 204, 107562. https://doi.org/10.1016/j.compag.2022.107562.

Neumann, C., Duboscq, J., Dubuc, C., Ginting, A., Irwan, A.M., Agil, M., Widdig, A., Engelhardt, A., 2011. Assessing dominance hierarchies: Validation and advantages of progressive evaluation with Elo-rating. *Anim. Behav.* 82 (4), 911–921. https://doi.org/10.1016/j.anbehav.2011.07.016.

Neumann, C., Fischer, J., 2022. Extending Bayesian Elo-rating to quantify the steepness of dominance hierarchies. *Methods Ecol. Evol.* https://doi.org/10.1111/2041-210x.14021.

Neumann, C., & Kulik, L. (2020). *EloRating: Animal Dominance Hierarchies by Elo Rating* (R package version 0.46.11). https://cran.r-project.org/package=EloRating.

R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. [Computer software] https://www.R-project.org/.

Sahar, M.W., Beaver, A., Daros, R.R., von Keyserlingk, M.A.G., Weary, D.M., 2022. Measuring lameness prevalence: Effects of case definition and assessment frequency. *J. Dairy Sci.* 105 (9), 7728–7737. https://doi.org/10.3168/jds.2021-21536.

Shah, N.B., Wainwright, M.J., 2018. Simple, Robust and Optimal Ranking from Pairwise Comparisons. *J. Mach. Learn. Res.* 18 (199), 1–38. http://jmlr.org/papers/v18/16-206.html.

Sheng, K., Foris, B., von Keyserlingk, M.A.G., Gardenier, J., Clark, C., Weary, D.M., 2023. Crowd sourcing remote comparative lameness assessments for dairy cattle. *J. Dairy Sci.* 106 (8), 5715–5722. https://doi.org/10.3168/jds.2022-22737.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* 35 (5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162.

Shrestha, A., Loukas, C., Le Kernec, J., Fioranelli, F., Busin, V., Jonsson, N., King, G., Tomlinson, M., Viora, L., Voute, L., 2018. Animal lameness detection with radar sensing. *IEEE Geosci. Remote Sens. Lett.* 15 (8), 1189–1193. https://doi.org/10.1109/LGRS.2018.2832650.

Sprecher, D.J., Hostetler, D.E., Kaneene, J.B., 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. *Theriogenology* 47 (6), 1179–1187.

Swartz, D., Shepley, E., Gaddis, K.P., Burchard, J., Cramer, G., 2024. Descriptive evaluation of a camera-based dairy cattle lameness detection technology. *J. Dairy Sci.* https://doi.org/10.3168/jds.2024-24851.

Thomsen, P.T., Shearer, J.K., Houe, H., 2023. Prevalence of lameness in dairy cows. *Vet. J.* 295, 105975. https://doi.org/10.1016/j.tvjl.2023.105975.

Van Nuffel, A., Sprenger, M., Tuyttens, F.A.M., Maertens, W., 2009. Cow gait scores and kinematic gait data: Can people see gait irregularities? *Anim Welf.* 18 (4), 433–439. https://doi.org/10.1017/s0962728600000841.

van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. CreateSpace. [Computer software].

Viazzi, S., Bahr, C., Schlageter-Tello, A., Van Hertem, T., Romanini, C.E.B., Pluk, A., Halachmi, I., Lokhorst, C., Berckmans, D., 2013. Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. *J. Dairy Sci.* 96 (1), 257–266. https://doi.org/10.3168/jds.2012-5806.

von Keyserlingk, M.A., Barrientos, A., Ito, K., Galo, E., Weary, D.M., 2012. Benchmarking cow comfort on North American freestall dairies: Lameness, leg injuries, lying time, facility design, and management for high-producing Holstein dairy cows. *J. Dairy Sci.* 95 (12), 7399–7408. https://doi.org/10.3168/jds.2012-5807.

Whay, H.R., Waterman, A.E., Webster, A.J.F., O'brien, J.K., 1998. The influence of lesion type on the duration ofhyperalgesia associated with hindlimb lameness in dairy cattle. *Vet. J.* 156 (1), 23–29. https://doi.org/10.1016/S1090-0233(98)80058-0.

Ying, X., 2019. An Overview of Overfitting and its Solutions. *J. Phys. Conf. Ser.* 1168, 022022. https://doi.org/10.1088/1742-6596/1168/2/022022.