J. Dairy Sci. TBC
https://doi.org/10.3168/jds.2024-25940

Evaluation of a fully automated 2-dimensional imaging system for real-time cattle lameness detection using machine learning

N. Siachos,* B. E. Griffiths, J. P. Wilson, C. Bedford, A. Anagnostopoulos, J. M. Neary, R. F. Smith, and G. Oikonomou
Department of Livestock and One Health, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, Neston, CH64 7TE, United Kingdom

ABSTRACT

Early detection and prompt treatment of lame cows are crucial for proactive lameness management. This study aimed to evaluate a fully automated 2-dimensional imaging system for real-time lameness detection using artificial intelligence. Data were collected from 11 dairy farms in the UK Four trained veterinarians performed 42 mobility scoring sessions using a 0–3 4-grade scoring system, with scores 2 and 3 representing lameness. On each session, individual weekly average scores were calculated. This resulted in 40,116 paired human mobility scores (HMS) and weekly average mobility scores generated using artificial intelligence (AIMS) matched to a cow ID. Categorical agreement for the 4-grade scale was estimated by calculating the weighted Cohen's kappa (κ_w) and Gwet's agreement coefficient (AC_2), and for the 2-grade scale (nonlame vs. lame) by calculating the percentage agreement (PA), unweighted Cohen's kappa (κ) and Gwet's coefficient (AC_1). A trained veterinarian recorded the presence and severity of any lesion of 2,515 cows, which also had an AIMS assigned. A subset of 758 cows were also assigned an HMS 1–3 d before trimming. Sensitivity (Se), specificity (Sp), and accuracy (Acc) were calculated to describe the system's and human's ability to detect cows with foot lesions. Additionally, automated mobility scores were retrieved for cows with foot lesion records up to 30 d before trimming. Linear mixed effects models (LMM) were built to assess the association of the lesion status at trimming with the daily scores. The average (mAVG), maximum (mMAX), minimum (mMIN) and the percentage of scores that a cow was identified as lame (mPLS) during the 30 d before foot trimming were calculated and their Se, Sp and Acc in detecting foot lesions were determined. Lastly, longitudinal data were obtained from 143 cows tracking daily scores from 5 to 64 DIM. The association of lesion status at the early lactation routine trim (ELRT) with the daily scores was assessed by fitting LMM. Regarding the 4-grade scale agreement between HMS and AIMS, κ_w (0.24–0.34) represented fair agreement, whereas AC_2 (0.81–0.93) almost perfect agreement. For the 2-grade scale agreement, PA was consistently above 80%, κ (0.23–0.38) represented fair agreement, and AC_1 (0.76–0.83) showed substantial to almost perfect agreement. The AIMS detected cows bearing severe lesions with Se = 0.53 and Sp = 0.74, whereas the HMS achieved Se = 0.60 and Sp = 0.78. Using optimal thresholds for mAVG, mMAX, mMIN, and mPLS, the system achieved higher Se than HMS. Moreover, cows with severe lesions had increased scores from 23 d before trimming compared with cows with mild and moderate lesions. Longitudinal data showed that cows with severe lesions at ELRT had higher mobility scores during the first 60 DIM compared with those with mild or moderate lesions. Overall, the system's performance was comparable to that of experienced human assessors in detecting lame cows and cows with foot lesions. Finally, its capability to detect mobility changes before the development of severe lesions highlights its potential for early intervention, which could enhance lameness management in dairy herds.

Key words: artificial intelligence, convolutional neural network, foot pathologies, locomotion, mobility

INTRODUCTION

Lameness in dairy cattle is described as a clinical symptom, representing underlying pathologies, with foot lesions being the most common cause (Murray et al., 1996). Digital dermatitis (**DD**) is the most important infectious cause of lameness, whereas claw horn disruption lesions, the collective term used for lesions such as sole ulcers (**SU**), sole hemorrhage (**SH**), and white line disease (**WL**), are the main noninfectious lameness causing lesions (Murray et al., 1996; Cramer et al., 2008). Lameness is prevalent worldwide (Thomsen et al., 2023)

Received October 29, 2024.
Accepted January 8, 2025.
*Corresponding author: nektarios.siachos@liverpool.ac.uk

The list of standard abbreviations for JDS is available at adsa.org/jds-abbreviations-25. Nonstandard abbreviations are available in the Notes.

and associated with significant and wide-ranging adverse effects to cow welfare (Whay et al., 1997; Whay and Shearer, 2017) and production efficiency (Charfeddine and Pérez-Cabal, 2017; Omontese et al., 2020). Furthermore, it has the potential to seriously damage public perception of the industry, as it is an easily recognized indicator of poor animal welfare (Jackson et al., 2022).

Chronically lame cows have a much-reduced response rate to treatment compared with animals treated promptly (Thomas et al., 2015, 2016). This is thought to be due to pathological changes to the pedal bone and digital cushion structures, which compromise their functionality, creating an environment conducive to a reduced treatment response and an increased risk of developing future lesions (Newsome et al., 2016; Randall et al., 2018; Wilson et al., 2021). Early detection and prompt and effective treatment is a key component in reducing lameness prevalence on dairy farms (Pedersen and Wilson, 2021) and is hypothesized to reduce the risk of pathological change, thereby improving treatment outcomes (Wilson et al., 2022).

Lameness detection has traditionally relied upon humans performing visual assessment of mobility using mobility and locomotion scoring systems. Depending on the system, posture, gait, or both are assessed to detect discomfort (Sprecher et al., 1997; Whay et al., 2003; Flower and Weary, 2006). The Agricultural and Horticultural Development Board (**AHDB**) 4-grade mobility scoring system (scores range 0–3) is predominantly used in the United Kingdom (Whay et al., 2003). Mobility scoring is inexpensive and unobtrusive, and can facilitate early treatment when employed frequently, resulting in improved cure rates (Alawneh et al., 2012; Leach et al., 2012; Groenevelt et al., 2014). However, the frequency of mobility scoring undertaken on UK dairy farms varies considerably, which in the authors' experience is often determined by requirements from milk processors to improve animal welfare, with some farms scoring weekly, whereas others score quarterly. Furthermore, some farms do not routinely mobility score, and instead rely on an ad hoc approach to detect lameness by observing cows when walking into the milking parlor, or through the detection of lesions at foot trimming (Griffiths et al., 2018). Farmer estimated lameness, without the use of mobility scoring systems, have been shown to be a poor lameness detection method (Espejo et al., 2006; Fabian et al., 2014; Beggs et al., 2019).

Human mobility scoring does, however, have some drawbacks. It is time consuming, particularly for large herds, and labor intensive, both of which are listed by farmers as considerable barriers to implementation (Leach et al., 2012). Human mobility scoring is subjective by nature. The Register of Mobility Scorers (**RoMS**) in the United Kingdom aims to ensure that accredited scorers follow consistent professional standards (RoMS, 2024). The background and training of the observer, as well as location, environment, and cow flow can all create variability contributing to low intra- and interobserver reliability (Van Nuffel et al., 2015; Nejati et al., 2023; Siachos et al., 2024b). The presence of an observer can alter cow behavior, which further complicates the accuracy of mobility scoring, with mild to moderate lameness often hidden in an effort to mask vulnerability (Van Nuffel et al., 2015). Yet alterations in cow behavior are not uniform and have been shown to be farm specific, reflecting the interaction between individual cow factors such as age and cattle handling (Waiblinger et al., 2003).

Technology is increasingly being adopted in modern dairy farming to address welfare challenges. There has been an increasing number of systems developed to identify lame cows using various kinetic, kinematic, and indirect methods at different levels of automation and applicability (O'Leary et al., 2020; Nejati et al., 2023; Siachos et al., 2024b). However, of the current welfare-based sensors available commercially, only a few have been independently validated (Stygar et al., 2021). One such system has been recently developed and commercialized by CattleEye Ltd. (Belfast, UK). Initial validation across 3 farms has identified that this system performs comparably to 2 well-trained observers (Anagnostopoulos et al., 2023). Furthermore, when examining the system's ability to detect lesions during foot trimming in a limited number of cows, low sensitivities and high specificities were described for visual mobility scoring, with automated lameness detection displaying greater sensitivity than visual mobility scoring, but reduced specificity (Anagnostopoulos et al., 2023). As causes of lameness, lameness prevalence, herd demographics, and environmental conditions could vary substantially across farms, and so there is a need to further validate this system across more farms and using larger datasets.

The timing of the initial corium insult leading to noninfectious lesions and the temporal relationship between lameness and lesion development remain unclear (Hoblet and Weiss, 2001). Moreover, infectious lesions, cases of DD in particular, show a dynamic transition from active and painful lesions to healed or chronic cases that may serve as reservoir of the causative *Treponema* spp. in the environment (Döpfer et al., 2012; Nielsen et al., 2012). A longitudinal analysis of daily mobility scores, alongside the development of lesions could provide insights into these relationships. By identifying cows at risk of lesion development, the intervention would prevent the development of more severe lesions, thereby improving cure rates (Leach et al., 2012; Groenevelt et al., 2014). Swartz et al. (2024) recently collected foot-trimming records from 3 North American dairy farms using the CattleEye system. They demonstrated that cows with lesions had

increased median weekly scores across 4 wk before the trimming date compared with those without any reported lesions. Cows with a SU had the highest median weekly scores preceding trimming. Cows with a WL had the largest score increase, whereas cows with a case of DD had the lowest median scores and relative score increase among cows with lesions.

Our objective was to further evaluate the lameness detection performance of the CattleEye system in dairy cows using a large dataset of mobility scores and foot lesion records. To achieve this objective, we investigated (1) the agreement among a large number of mobility scores across many farms provided by CattleEye and those provided by multiple assessors using a visual mobility scoring system, (2) the accuracy of the mobility scores provided by CattleEye in detecting the presence of foot lesions recorded consistently by a trained human assessor (**HA**) across a large number of cows, and (3) the temporal association between mobility scores provided by the CattleEye system and the development of foot lesions during the lactation period using longitudinal data.

MATERIALS AND METHODS

Ethics Statement

The study was approved by the University of Liverpool Veterinary Research Ethics Committee (Reference VREC1079).

Farms and Animals

We collected data from July 2022 to March 2024. Eleven commercial large-size dairy farms designated as A through K, located in Wales and West and South England, participated in this study. Farms were milking ~1,000, 2,300, 800, 2,100, 760, 800, 600, 2,100, 1,500, 630, and 2,800 Holstein cows 3 times per day.

Automated Mobility Scoring System

The automated mobility scoring system evaluated here has been developed and commercialized by CattleEye Ltd. All participating farms were equipped with a 2-dimensional surveillance camera placed over a passageway at the exit of the milking parlor at a height of 4 m above the ground. Details about the camera setup and the functional characteristics of the system have been provided in a previous publication (Anagnostopoulos et al., 2023). Briefly, the camera captures overhead footage of cows walking through a passageway. Footage during one milking is sent to the company's servers, stored in the cloud, and processed. At first, an object-tracking algorithm identifies the outline of each cow, coat pattern, and head shape and assigns the identification number (ID) of the individual animal to the recording. The system can also pull information about the cow ID from the sorting gates or the radio-frequency identification system available in the farm. Specific anatomical key points are marked, and their coordinates are followed across frames. These are then processed by a convolutional neural network architecture, which produces a mobility score prediction. The system produces a mobility score on a continuous scale from 0 to 100 (from perfect mobility to severe lameness), with each 25-point increment corresponding to one grade (0–3) on the 4-grade UK AHDB scoring system, with scores 2 and 3 considered as lame (Whay et al., 2003).

For the purpose of this study, individual daily mobility records were available for each cow, and weekly average scores were also calculated. The system's 4-grade converted weekly average mobility score will be hereinafter referred as **AIMS**, and the binary converted score (nonlame: scores 0 and 1; lame: scores 2 and 3) as **AIMS_BIN**.

Human Mobility Scoring Records

Four HA (numbered 1–4; namely HA1 [NS], HA2 [BG], HA3 [AA], and HA4 [GO]) performed a total of 42 whole-milking-herd mobility scoring sessions. All 4 assessors were qualified veterinarians and experienced mobility scorers, with HA1, HA2, and HA3 being RoMS accredited (Wimborne, UK), and HA4 having 20 yr of experience in cattle lameness research.

During each session, a single HA mobility scored the entire milking herd using the 4-grade AHDB scoring system as cows were exiting the milking parlor during the midday milking, as the cows walked on level, good-quality concrete. Several sessions performed on farm D (8 sessions) and one on farm H included only specific milking groups and not the whole herd. Recording was performed mainly using a voice recorder or by manually writing down the cow ID (freeze brand number located at the rear thigh area on either side of the tail, or ear-tag number when the freeze brand was not clear) and the mobility score on spreadsheets attached in a clipboard. All records were then transcribed into Excel (Microsoft Corp.) spreadsheets. The 4-grade (0, 1, 2, 3) and the binary converted human mobility scores (0 or 1; 2 or 3) will be hereinafter referred to as **HMS** and **HMS_BIN**, respectively.

The AIMS at the same visit day were also available and stored. Human assessors had no access to the automated mobility records; CattleEye Ltd. did not have access to HMS. At the end of the study, HMS and AIMS were matched using the date and the cow ID.

To assess the interobserver agreement between trained human scorers, HA1 and HA2 visited farm D on the same

day and scored ~780 cows during the morning and the afternoon milking, respectively.

Foot Lesions Data

We collected data during 61 foot-trimming sessions in 5 of the participating farms (A, D, H, I, and K). Cows in these farms were housed all year round in typical 2-row and 3-row freestall barns with grooved concrete floors and were foot-bathed daily. All sessions were performed by professional foot trimmers, and they included both routine and therapeutic trims, with HA1 being blind to which cows were presented. The presence of any lesion in all 4 feet of 2,698 cows was consistently recorded according to the International Committee for Animal Recording claw health atlas (Egger-Danner et al., 2014), and the severity of each lesion was graded. Definition and grading methodology used is described in Supplemental Table S1 (see Notes). More than 90% of the assessments were performed by HA1, and the rest were performed by HA3, a qualified veterinarian who followed the same definition and grading methodology.

Cows were classified into 3 categories according to their foot lesion status as follows:

- Status 1 or "mild" included cows with no lesions or bearing mild lesions: double sole, heel horn erosion, SH of grade 1, WL of grade 1, axial wall fissure (**AWF**) of grade 1, and DD of grade 1.
- Status 2 or "moderate" included cows bearing at least one lesion of moderate severity: SU of grade 1, SH of grade 3, WL of grade 2, AWF of grade 2, interdigital hyperplasia (**IH**) of grade 1 and 2, interdigital phlegmon (**IP**) of grade 1, and DD of grade 2.
- Status 3 or "severe" included cows bearing at least one severe lesion: SU of grade >1, WL of grade 3, AWF of grade 3, toe ulcer (**TU**) of grade >0, IH of grade 3, IP of grade 2, and DD of grade 3.

We also recorded any cow presented to the trimmer as lame with an obvious upper limb case of lameness. These cases included injuries, large abscesses, swollen joints, or joint luxation. Moreover, cows presented to the trimmer for re-examination having a hoof block already applied were also recorded. Both types of cases were excluded from the analysis regardless of the presence of foot lesions. Finally, information about the parity and the latest calving date of each cow were collected from each farm's herd management software.

Longitudinal Study Data

One hundred forty-three cows on farm B that calved between July 28 and September 28, 2023, were prospectively enrolled in a longitudinal study to compare daily automated mobility scores over time between cows that did or did not develop foot lesions during early lactation. The studied population consisted of 62 primiparous and 81 multiparous cows. The hind feet of all cows were examined by HA1 within 4 to 10 DIM by removing a thin layer of horn and modeling to examine for the presence of any lesion. Front feet at this stage were not examined to minimize handling stress. The same researcher was present during the early lactation routine trimming (**ELRT**) sessions, which were performed on this farm at a median of 94 DIM (ranging from 64 to 146 DIM) by a professional foot trimmer, who recorded the presence and graded the severity of any lesion in all 4 feet. The same definition of lesions and grading methodology was followed for the "fresh cow" trim (**FCT**) and the early lactation routine trim, as previously described. At the end of this study, the individual daily automated mobility scores on a continuous scale from 5 to 64 DIM were made available to us.

Daily Automated Mobility Scoring Data up to 30 d Before Trimming

For cows with foot lesion records we retrieved the individual daily automated mobility scores on a scale from 0 to 100 and we created 2 new datasets. The first dataset (**PriorDATA1**) included all cows, for whom we retrieved daily scores from 30 d to 1 d before trimming date. Only cows having at least 10 daily scores recorded were included, and for cows with multiple trimming sessions, we chose the earliest one if the interval between sessions was less than 30 d. A total of 1,986 cows met these criteria.

The second dataset (**PriorDATA2**) consisted only of cows that were trimmed between 60 and 120 DIM, and we retrieved daily scores from calving day to 60 DIM. Similar to the first dataset, only cows having at least 10 scores were included, and for cows with multiple trimming sessions, we chose only the earliest one. A total of 615 cows met these criteria.

Statistical Analysis

Data were handled and analyzed with IBM SPSS v.28 (IBM Corp., Armonk, NY) and R Studio (v4.3.1; R Core Team, 2023).

Interobserver Agreement

The categorical interobserver agreement between AIMS and the HMS of each human scorer was assessed by calculating the quadratically weighted Cohen's kappa coefficient (κ_w) and the quadratically weighted Gwet's coefficient (AC_2). Similarly, the agreement between AIMS_BIN and HMS_BIN of each human scorer was assessed by calculating the percentage agreement (**PA**), the unweighted Cohen's kappa coefficient (κ), and the unweighted Gwet's coefficient (AC_1). The same metrics were used to assess the interobserver agreement between HA1 and HA2. The Gwet's coefficients were computed using R Studio (v4.3.1; R Core Team, 2023) and the irrCAC package (Gwet, 2001). We chose to include in our analysis both Cohen's kappa, for consistency and comparability with previous studies, and Gwet's agreement coefficients, which are considered more robust measures of chance-corrected agreement, particularly in situations with low or high prevalence of the tested trait or marginal imbalance (Gwet, 2008).

To interpret PA, we used the conventionally accepted benchmark of accepted reliability of 80% (McHugh, 2012). To interpret κ and κ_w, and AC_1 and AC_2 estimates, we used the recommendations by Landis and Koch (1977), as follows: slight agreement (0.00–0.20), fair agreement (0.21–0.40), moderate agreement (0.41–0.60), substantial agreement (0.61–0.80), and almost perfect agreement (0.81–1.00). Values above 0.60 have been considered as representing an acceptable level of interobserver categorical agreement for various health and welfare indices (Gibbons et al., 2012; Schlageter-Tello et al., 2014).

Accuracy in Predicting the Presence of Foot Lesions

Using the dataset with foot lesion records, we created confusion matrixes for the overall study population and within each parity class, to calculate sensitivity (**Se**), specificity (**Sp**), and classification accuracy (**Acc**) for AIMS_BIN and HMS_BIN in accurately predicting the presence of "severe" (status 3) and of "moderate and severe" (merged status 2 and 3) foot lesions, using the lesion identification and grading methodology previously described as the ground truth.

The formulas to calculate Se, Sp, and Acc were:

$$Se = \frac{True\,Positives}{True\,Positives + False\,Negatives},$$

$$Sp = \frac{True\,Negatives}{True\,Negatives + False\,Positives}, \text{ and}$$

$$Acc = Se \times P + Sp \times \left(1 - P\right),$$

where P = prevalence.

The 95% binomial proportion CI for Se, Sp, and Acc were calculated with the "exact" Clopper-Pearson method (Clopper and Pearson, 1934).

Additionally, we calculated Se, Sp, and Acc for AIMS_BIN and HMS_BIN in accurately predicting the presence of SH of grade 3, SU of grade ≥1, WL of grade 3, TU of grade ≥1, or DD of grade 3, separately, for the overall study population. Cases of AWF of grade 3 were merged with those of WL of grade 3, as considered of similar origin and create the same level of pain and discomfort to the cow. Only cases without severe co-existing lesions were considered as negative controls.

Data from Longitudinal Study

Cows were classified according to findings at FCT, as having at least one case of moderate and severe lesion (status 1 vs. merged status 2 and 3) or not, and into 3 classes (status 1, 2, or 3) according to ELRT findings. Furthermore, we created a second dataset by excluding cows that were diagnosed at ELRT with moderate or severe lesions other than SH grade ≥2 or SU of any grade. This resulted in 130 cases with SH and SU status.

To assess the association of ELRT lesion status with the daily automated mobility scores from 5 to 64 DIM, we used linear mixed models with repeated measurements. Two separate models were fitted, with (1) ELRT lesion status (3 levels: status 1, 2, and 3), and (2) binary SH and SU status (1: cases at ELRT with SH grade ≥2 or SU of any grade; 0: the rest) as the main fixed effects of interest. Parity (2 levels: primiparous vs. multiparous), DIM, and FCT status (binary) as main effects and all 2-way interactions were the other explanatory and adjusted variables fitted as fixed effects in both models. Days in milk were used to specify the repeated measurements statement, accounting for the random effect of the cow.

Retrospective Assessment of Automated Mobility Scores up to 30 d Before Trimming. Using PriorDATA1, we used linear mixed effects models (**LMM**) with repeated measurements to assess retrospectively the association of the foot lesion status at trimming with the automated mobility scores from 30 d to 1 d before the trimming session. Three separate models were fitted with the inclusion of (1) overall lesion status (**OLS**) at 3 levels (mild, moderate, severe), (2) binary OLS (**OLS_BIN_SEV**; mild and moderate vs. severe), and (3) binary OLS (**OLS_BIN_MODSEV**; mild vs. moderate and severe) as the main fixed effect of interest. Farm (5 levels), parity (4 levels: first, second, third, and fourth or greater), and days before trimming (**DBT**) as main

effects; all 2-way interactions were the other explanatory and adjusted variables fitted as fixed effects in all models. Days before trimming were used to specify the repeated measurements statement, accounting for the random effect of the cow.

Similarly, using PriorDATA2, we fitted 3 separate models, following the same parametrization for OLS, to assess the retrospective association of the foot lesion status at 60 to 120 DIM, with the daily automated mobility scores from the day of calving to 60 DIM. In addition to OLS, farm (5 levels), parity (4 levels: first, second, third, and fourth or greater), and DIM were main effects; all 2-way interactions were the other explanatory and adjusted variables fitted as fixed effects in all models.

Data analysis using LMM described in the previous 2 subsections was undertaken using R Studio (v4.3.1; R Core Team, 2023), with the Tidyverse (Wickham et al., 2019), nlme (Pinheiro and Bates, 2023), and emmeans (Lenth, 2024) packages. Model building strategy and model fit procedures were the same across all LMM. Univariable analyses were performed as an initial exploratory analysis on the independent variables to be included in the LMM. For each model, the appropriate covariance structure producing the best fit was selected based on the lowest Akaike's information criterion value. Where the main variables of interest produced a statistically significant association, final models were built by backward eliminating explanatory variables with a nonsignificant association at $P > 0.10$. Normality and homoscedasticity of data were assessed by visual inspection of the fitted values versus residuals plot and the normal Q-Q plot, respectively. On each model, pairwise comparisons between the estimated marginal means (**EMM**) of the different classes of the lesion status variables were performed using Bonferroni's CI adjustment. The EMM were then plotted against DIM or DBT to visualize the evolution of the automated mobility scores per lesion status class across time.

Optimal Thresholds and Accuracy for Parameters Derived from Monthly Automated Mobility Scores. Using the PriorDATA1, we calculated for each cow the monthly average (**mAVG**), maximum (**mMAX**) and minimum (**mMIN**) score, and the percentage of daily scores that a cow was recorded as lame (**mPLS**). Then, receiver operating characteristic (**ROC**) curves were created for each parameter to identify optimal thresholds in accurately predicting the presence of severe (status 3) and moderate and severe (status 2 and 3) foot lesions, and Se, Sp, and Acc were calculated for each threshold, for the overall study population and within each parity class.

Using the same parameters as test variables on ROC curves, we identified optimal thresholds and calculated the Se, Sp, and Acc in accurately predicting the presence of SH of grade 3, SU of grade ≥1, WL of grade 3, TU of grade ≥1, or DD of grade 3, separately, for the overall study population.

RESULTS

In total, 47,538 HMS were recorded, out of which 44,981 were matched to a cow ID. After merging the HMS and AIMS using the date and the cow ID, 40,116 paired scores were available for statistical analysis. The number of scored cows and the lameness prevalence recorded by the HA and by the system per session are detailed in Supplemental Table S2 (see Notes). Herd-level lameness prevalence ranged from 7% to 30% based on HMS and from 2% to 30% based on AIMS.

Interobserver Agreement Between Human Scorers and the System

The interobserver categorical agreement between the weekly average mobility scores generated by artificial intelligence and the human mobility scores of each HA is summarized in Table 1 and shown in detail in Supplemental Table S3 (see Notes). Regarding the agreement on the 4-grade scale between AIMS and HMS, the Cohen's κ_w ranged from 0.24 to 0.34 representing only fair agreement, whereas Gwet's AC_2 ranged from 0.81 to 0.93 representing almost perfect agreement. Regarding the agreement on the binary converted 2-grade scale between AIMS_BIN and HMS_BIN, the PA ranged from 81.5% to 86.3% being consistently above the benchmark of accepted reliability. Moreover, Cohen's κ ranged from 0.23 to 0.38 representing only fair agreement, and Gwet's AC_1 ranged from 0.76 to 0.83 representing substantial and almost perfect agreement.

Interobserver Agreement Between Human Scorers

Results on the interobserver agreement between HA1 and HA2 is shown in Table 2. The 4-grade scale agreement produced a Cohen's κ_w and a Gwet's AC_2 of 0.27 (95% CI: 0.21–0.33) and 0.75 (0.95% CI: 0.72–0.78) representing fair and substantial agreement, respectively. The PA for the binary converted scale was 76.7%, whereas Cohen's κ and Gwet's AC_1 were 0.27 (95% CI: 0.19–0.35) and 0.67 (95% CI: 0.61–0.72) representing fair and substantial agreement, respectively.

Foot Lesions Data

From the initial dataset of 2,698 cows with foot lesion records, 33 cows had a case of upper limb lameness. One hundred twenty cows were presented to the trimmer for re-examination and had already a hoof block applied; these cows were excluded from the analysis. Finally,

Table 1. Overall interobserver categorical agreement between the weekly average scores provided by an automated system (CE) and the visual mobility scores assigned by 4 experienced human assessors (HA) in 11 commercial dairy farms[1]

	CE vs. HA1			CE vs. HA2			CE vs. HA3			CE vs. HA4		
	n = 28,225			n = 7,225			n = 3,466			1,200		
Score	PA, %	κ/κ_w	AC_1/AC_2	PA, %	κ/κ_w	AC_1/AC_2	PA, %	κ/κ_w	AC_1/AC_2	PA, %	κ/κ_w	AC_1/AC_2
0 or 1; 2 or 3	83.7	0.38 (0.37–0.40)	0.78 (0.77–0.79)	81.5	0.23 (0.20–0.26)	0.76 (0.75–0.77)	82.1	0.32 (028–0.36)	0.76 (0.74–078)	86.3	0.34 (0.26–0.42)	0.83 (0.80–0.85)
0, 1, 2, or 3		0.34 (0.33–0.35)	0.86 (0.86–0.87)		0.33 (0.31–0.35)	0.85 (0.85–0.86)		0.24 (0.21–0.27)	0.81 (0.80–0.82)		0.26 (0.20–0.33)	0.93 (0.92–0.94)

[1]Agreement on the 4-grade scale (0–3) was estimated by calculating the quadratically weighted Cohen's kappa (κ_w) and the quadratically weighted Gwet's agreement coefficient (AC_2), with 95% CI shown in parentheses. Agreement on the binary converted 2-grade scale (0 or 1; 2 or 3) was estimated by calculating the percentage agreement (PA), unweighted Cohen's kappa (κ), and the unweighted Gwet's agreement coefficient (AC_1), with 95% CI shown in parentheses.

2,515 cows were assigned an AIMS, and 758 were scored by the same HA 1 to 3 DBT and were also assigned an HMS. The prevalence and the severity of lesions recorded on a cow level are shown in Table 3. On a descending order, SH grade 3, DD grade ≥2, WL grade 3, and SU grade ≥1 were the most prevalent lesions recorded in our population.

The overall and per parity measures of accuracy for AIMS_BIN and HMS_BIN in correctly detecting cows bearing foot lesions using the foot lesions data recorded by HA1 as ground truth, are presented in Table 4. The AIMS_BIN achieved an overall combination of Se, Sp, and Acc of 0.37 (95% CI: 0.34–0.39), 0.76 (95% CI: 0.73–0.78), and 0.58 (95% CI: 0.56–0.60), respectively, in detecting the presence of moderate and severe lesions. The HMS_BIN achieved an overall combination of Se, Sp, and Acc of 0.38 (95% CI: 0.33–0.44), 0.84 (95% CI: 0.80–0.87), and 0.62 (95% CI: 0.58–0.65), respectively, in detecting the presence of moderate and severe lesions. Measures of accuracy varied across parities. Both the automated and the human scores achieved the lowest Se in parity 1 cows (0.12 and 0.21, respectively) and the highest Se in parity 4+ cows (0.46 and 0.58, respectively).

Regarding the accuracy in detecting cows bearing at least one case of "severe" lesions, AIMS_BIN and HMS_BIN produced an overall combination of Se, Sp, and Acc of 0.53 (95% CI: 0.47–0.58), 0.74 (95% CI: 0.72–0.76), and 0.71 (95% CI: 0.69–0.73), and 0.60 (95% CI: 0.50–0.70), 0.78 (95% CI: 0.75–0.81), and 0.76 (95% CI: 0.73–0.79), respectively. Both the automated and the human scores achieved the lowest Se in parity 1 cows (0.26 and 0.33, respectively), whereas the highest Se for the automated scores we achieved in parity 3 (0.60) and for the human scores were achieved in 4+ cows (0.75).

The measures of accuracy for the automated and the human mobility scores in detecting the presence of each lesion separately are detailed in Table 5. The automated system was able to detect the presence of SH grade 3, SU, WL grade 3, TU, and DD grade 3 with Se/Sp combinations of 0.40/0.75, 0.52/0.75, 0.55/0.75, 0.64/0.75, and 0.50/0.75, respectively, whereas the human mobility scores could detect the presence of these lesions with Se/Sp combinations of 0.49/0.81, 0.63/0.81, 0.67/0.81, 1.00/0.81, and 0.38/0.81, respectively.

Table 2. Interobserver agreement between 2 trained human assessors (HA1 and HA2) mobility scoring cows on the same day in one of the participating farms (farm D) using the 4-grade AHDB mobility scoring system and binary converted 2-grade scale (0 or 1; 2 or 3)[1]

	HA1 vs. HA2			
Score	n	PA, %	κ/κ_w	AC_1/AC_2
0 or 1; 2 or 3	705	76.7	0.27 (0.19–0.35)	0.67 (0.61–0.72)
0, 1, 2, or 3			0.27 (0.21–0.33)	0.75 (0.72–0.78)

[1]Agreement on the 4-grade scale (0, 1, 2, or 3) was estimated by calculating the quadratically weighted Cohen's kappa (κ_w) and the quadratically weighted Gwet's agreement coefficient (AC_2) with 95% CI shown in parentheses. Agreement on the binary converted 2-grade scale (0 or 1; 2 or 3) was estimated by calculating the percentage agreement (PA), unweighted Cohen's kappa (κ) and the unweighted Gwet's agreement coefficient (AC_1) with 95% CI shown in parentheses.

Longitudinal Data

Days in milk, ELRT × DIM interaction, and parity were identified as significant predictors of the daily automated mobility scores variation. The lesion status at FCT did not produce any significant associations. The plotted EMM (±95% CI) for the ELRT × DIM interaction for each lesion status level are displayed in Figure 1. Cows with severe lesions at the early lactation trim had significantly greater ($P \leq 0.035$) automated mobility scores than both cows with moderate and with mild or no lesions from 36 DIM to 50 DIM, and greater scores ($P \leq 0.046$) than cows with mild or no lesions from 54 DIM to 64 DIM. Primiparous cows had overall lower daily automated mobility scores compared with multiparous by 4.7 points ($P = 0.001$). The EMM for cows with mild or no lesions were consistently below 40 (range: 32–39) without any abrupt changes.

When we excluded from the analysis cows that had lesions other than SH grade ≥2 or SU of any grade at ELRT, we found that DIM and SH or SU status × parity interaction were the only significant predictors of the daily automated mobility scores variation. The overall effect of SH or SU status was not significant ($P = 0.096$). However, the EMM for automated mobility scores of cows having at least one case of SH grade ≥2 or SU of any grade at ELRT were greater ($P \leq 0.045$) compared with cows without these lesions at specific time points (DIM 6, and DIM 61–64). Plotted EMM for the binary SH or SU status by DIM are provided in Supplemental Figure S1 (see Notes).

Table 3. Total number and percentage of foot lesions and severity grading recorded by a trained veterinarian during 61 foot-trimming sessions, including routine and therapeutic trims, performed by professional foot trimmers in 5 of the participating farms

Lesion and grade of severity	n	%
Sole hemorrhage		
1	743	29.5
2	494	19.6
3	280	11.1
Sole ulcer		
1	83	3.3
2	18	0.7
3	3	0.1
White line		
1	448	17.8
2	354	14.1
3	189	7.5
Axial wall fissure		
1	9	0.4
2	9	0.4
3	7	0.3
Toe ulcer		
1	1	0.0
2	7	0.3
3	3	0.1
Interdigital hyperplasia		
1	41	1.6
2	34	1.4
3	4	0.2
Interdigital phlegmon		
1	7	0.3
2	4	0.2
Digital dermatitis		
1	132	5.2
2	85	3.4
3	113	4.5

Assessment of Daily Automated Mobility Scores up to 30 DBT

From the LMM using PriorDATA1, parity, farm, and OLS × DBT interaction were the significant predictors of the automated mobility scores variation from 30 d to 1 d before trimming. The EMM for automated mobility scores of cows with severe lesions were significantly greater ($P \leq 0.027$) than those of cows with moderate and mild lesions from as early as 23 DBT and were consistently above 40 points from this time point onward (Figure 2). Cows at fourth or greater, third, and second parity had overall greater EMM compared with primiparous cows by 7.1, 5.2, and 3.0 points, respectively ($P < 0.001$).

Regarding the LMM with binary lesion status (mild and moderate vs. severe) as the main variable of interest, parity, OLS_BIN_SEV × DBT interaction, farm, and DBT were significant predictors of the automated mobility scores variation from 30 d to 1 d before trimming. The EMM for automated mobility scores of cows with severe lesions were consistently greater ($P < 0.001$) than those of cows with moderate and mild lesions during the entire 30 DBT and are shown in Supplemental Figure S2 (see Notes).

Regarding the LMM with binary lesion status (mild vs. moderate and severe), parity, lesion status (specifically OLS_BIN_MODSEV), farm, and DBT were significant predictors of the automated mobility scores variation from 30 d to 1 d before trimming, when merging moderate and severe lesions. The EMM for automated mobility scores of cows with moderate and severe lesions were greater ($P \leq 0.047$) than those of cows with mild lesions during most of the time before trimming, and are shown in Supplemental Figure S3 (see Notes).

From the LMM using PrioDATA2 for cows that were trimmed between 60 and 120 DIM, parity, OLS × parity interaction, farm, OLS, and OLS × DIM interaction were significant predictors of the automated mobility scores variation during the first 60 DIM. The EMM for automated mobility scores of cows with severe lesions were greater ($P \leq 0.047$) than those in cows with moderate and mild lesions on several time points from 24 to 32 DIM, at 47 DIM, and from 55 to 60 DIM (Figure 3). The EMM

Table 4. Overall and per parity measures of accuracy (sensitivity, Se; specificity, Sp; accuracy, Acc) for the binary converted human mobility scores (scores 2 and 3 on the 4-grade scale) and for the binary converted weekly average automated mobility scores (scores ≥50) in correctly detecting cows bearing at least one case of moderate and severe foot lesions using the recordings from a trained veterinarian as ground truth (exact Clopper-Pearson binomial 95% CI are shown in parentheses)

		Moderate and severe			Severe		
Parity	n	Se (95% CI)	Sp (95% CI)	Acc (95% CI)	Se (95% CI)	Sp (95% CI)	Acc (95% CI)
Human mobility scores (≥2)							
Overall	758	0.38 (0.33–0.44)	0.84 (0.80–0.87)	0.62 (0.58–0.65)	0.60 (0.50–0.70)	0.78 (0.75–0.81)	0.76 (0.73–0.79)
Parity 1	247	0.21 (0.14–0.30)	0.94 (0.89–0.97)	0.62 (0.56–0.68)	0.33 (0.15–0.57)	0.89 (0.85–0.93)	0.85 (0.80–0.89)
Parity 2	154	0.32 (0.20–0.45)	0.86 (0.77–0.92)	0.66 (0.58–0.73)	0.62 (0.32–0.86)	0.83 (0.76–0.89)	0.81 (0.74–0.87)
Parity 3	160	0.38 (0.27–0.50)	0.76 (0.66–0.85)	0.58 (0.50–0.66)	0.60 (0.41–0.77)	0.76 (0.68–0.83)	0.73 (0.66–0.80)
Parity 4+	178	0.58 (0.49–0.67)	0.67 (0.54–0.79)	0.61 (0.54–0.68)	0.75 (0.58–0.88)	0.57 (0.59–0.65)	0.61 (0.53–0.68)
Weekly average automated mobility scores (≥50)							
Overall	2,515	0.37 (0.34–0.39)	0.76 (0.73–0.78)	0.58 (0.56–0.60)	0.53 (0.47–0.58)	0.74 (0.72–0.76)	0.71 (0.69–0.73)
Parity 1	440	0.12 (0.07–0.17)	0.94 (0.90–0.97)	0.61 (0.57–0.66)	0.26 (0.12–0.45)	0.93 (0.90–0 0.95)	0.88 (0.85–0.91)
Parity 2	437	0.24 (0.18–0.32)	0.85 (0.80–0.89)	0.63 (0.59–0.68)	0.41 (0.25–0.58)	0.84 (0.80–0.87)	0.80 (0.76–0.84)
Parity 3	820	0.42 (0.37–0.48)	0.71 (0.67–0.75)	0.59 (0.56–0.63)	0.60 (0.50–0.68)	0.70 (0.66–0.73)	0.68 (0.65–0.72)
Parity 4+	797	0.46 (0.41–0.51)	0.60 (0.55–0.65)	0.52 (0.49–0.56)	0.56 (0.48–0.65)	0.59 (0.56–0.63)	0.59 (0.55–0.62)

[1]n = number of cows.

of multiparous cows were overall greater compared with primiparous cows by 13.2 points ($P < 0.001$).

Regarding the LMM with the binary lesion status, where mild and moderate lesions were merged into one group and assessed against the severe lesion group (specifically OLS_BIN_SEV), as the main independent variable of interest, OLS_BIN_SEV × DIM interaction, parity, parity × DIM interaction, DIM, farm, farm × parity interaction and OLS_BIN_SEV × parity interaction were significant predictors of the automated mobility scores variation during the first 60 DIM. The EMM for automated mobility scores of cows with severe lesions were numerically greater compared with those of cows with mild or moderate lesions during the first 60 DIM, with these being greater ($P \leq 0.042$) from 55 to 60 DIM (Supplemental Figure S4, see Notes).

Regarding the LMM with the binary lesion status, where the mild lesion group was assessed against the merged group of moderate and severe lesions (specifically OLS_BIN_MODSEV), as the main independent variable of interest, neither OLS_BIN_MODSEV nor

Table 5. Measures of accuracy (sensitivity, Se; specificity, Sp; accuracy, Acc) for the binary converted human mobility scores (scores 2 and 3 on the 4-grade scale) and for the binary converted weekly average automated mobility scores (scores ≥50) in correctly predicting the presence of specific foot lesions, using the recordings from a trained veterinarian as ground truth (exact Clopper-Pearson binomial 95% CI are shown in parentheses)

	Human mobility scores (≥2)				Weekly average automated mobility scores (≥50)			
	n[1]	Se (95% CI)	Sp (95% CI)	Acc (95% CI)	n[1]	Se (95% CI)	Sp (95% CI)	Acc (95% CI)
Sole hemorrhage (Grade 3)	87/661	0.49 (0.39–0.60)	0.81 (0.77–0.84)	0.77 (0.73–0.80)	280/2,208	0.40 (0.35–0.46)	0.75 (0.73–0.77)	0.70 (0.68–0.72)
Sole ulcer (Grade ≥1)	38/612	0.63 (0.46–0.78)	0.81 (0.77–0.84)	0.80 (0.76–0.83)	104/2,032	0.52 (0.42–0.62)	0.75 (0.73–0.77)	0.74 (0.72–0.75)
White line (Grade 3)	55/629	0.67 (0.53–0.79)	0.81 (0.77–0.84)	0.80 (0.76–0.83)	196/2,124	0.55 (0.47–0.62)	0.75 (0.73–0.77)	0.73 (0.71–0.75)
Toe ulcer (Grade ≥1)	5/579	1.00 (0.48–1.00)	0.81 (0.77–0.84)	0.81 (0.78–0.84)	11/1,939	0.64 (0.31–0.89)	0.75 (0.73–0.77)	0.75 (0.73–0.77)
Digital dermatitis (Grade 3)	34/608	0.38 (0.22–0.56)	0.81 (0.77–0.84)	0.79 (0.75–0.82)	113/2,041	0.50 (0.40–0.59)	0.75 (0.73–0.77)	0.74 (0.71–0.75)

[1]n = number of actual positive cows/number of total cows eligible, after excluding cases with concomitant severe lesions.

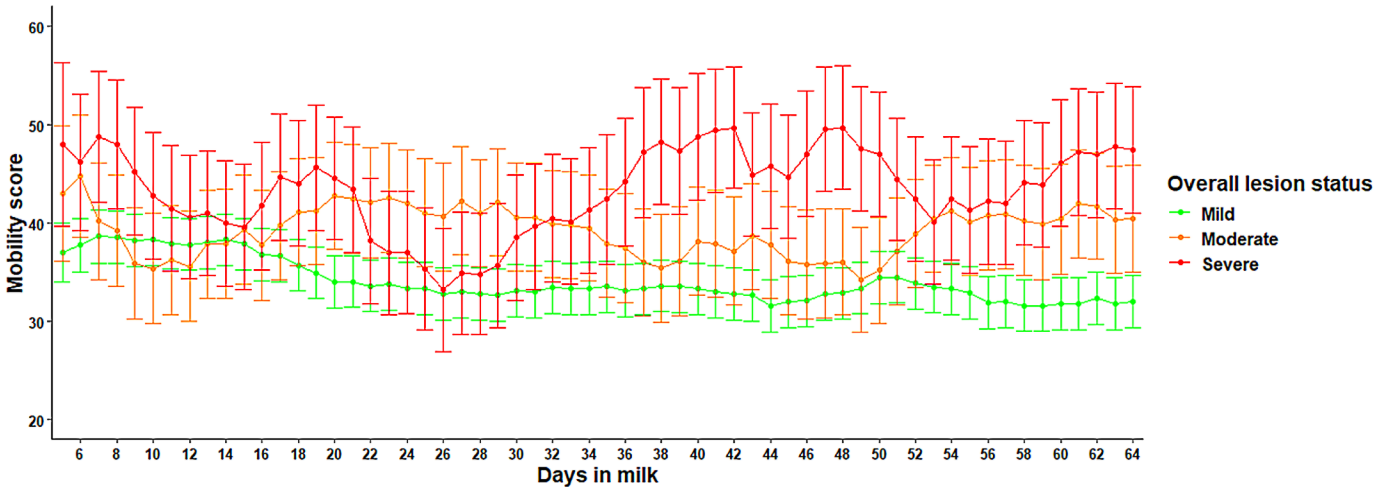

Figure 1. Estimated marginal means (±95% CI) derived from linear mixed models accounting for the effect of parity and lesion status detected immediately after calving, showing the evolution of daily automated mobility scores tracked from 5 to 64 DIM in 143 cows classified in 3 levels according to the presence and severity of foot lesions identified during the early lactation foot trim, which was performed at a median of 94 DIM. A statistically significant association of the overall lesion status × DIM interaction was observed ($P < 0.001$) with the daily mobility scores.

OLS_BIN_MODSEV × DIM interaction yielded any statistical significance.

Optimal Thresholds and Accuracy for Parameters Derived from Mobility Patterns 30 DBT

Using the PriorDATA1, the optimal thresholds for mAVG, mMAX, mMIN and mPLS in detecting cows with severe and with moderate and severe lesions, with the calculated Se, Sp, and Acc, overall and per parity, are presented in detail in Table 6.

Considering the accurate detection of cows with moderate and severe lesions, the overall threshold of 21.2% for mPLS produced the highest Se (0.48, 95% CI: 0.45–0.52), and the threshold of 58.5 for mMAX produced the best discriminative performance with the highest area under the curve (**AUC**; 0.60, 95% CI: 0.57–0.62) and Acc (0.62, 95% CI: 0.59–0.64). Sensitivities up to 0.67 (for mMAX) were achieved in cows at third parity. None

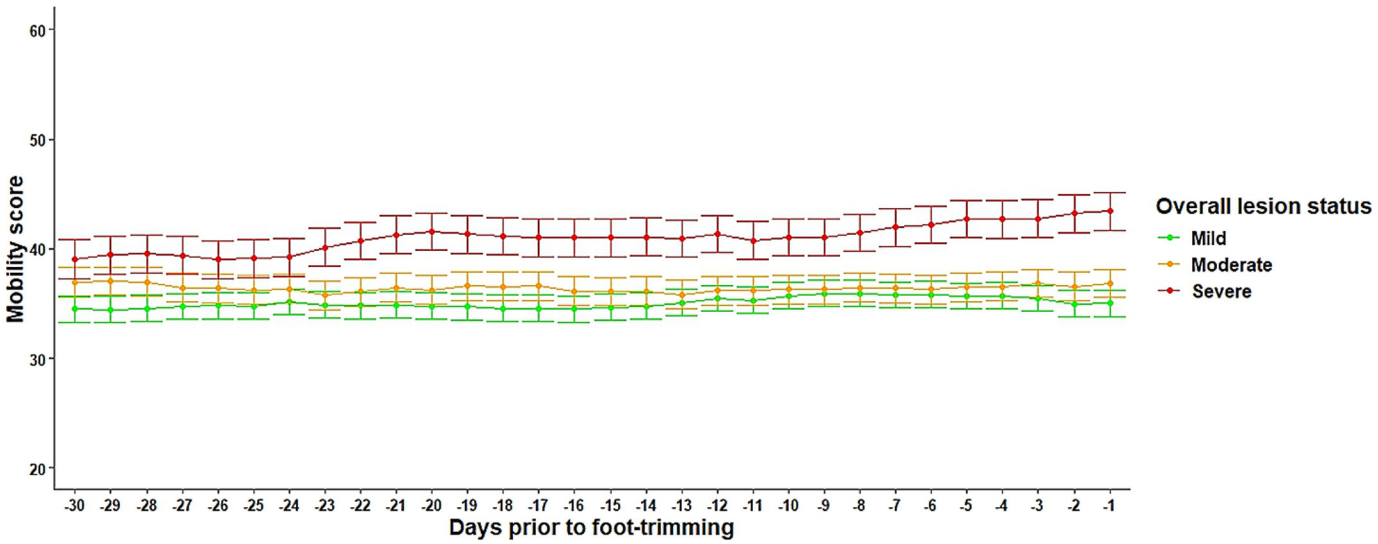

Figure 2. Estimated marginal means (±95% CI) derived from linear mixed models accounting for farm and parity effects, showing the evolution of daily automated mobility scores tracked from 30 to 1 d before trimming (DBT) in 1,986 cows of 5 farms classified in 3 levels according to the presence and severity of foot lesions identified during the trimming session. A statistically significant association of the overall lesion status ($P < 0.001$) and of the overall lesion status × DBT interaction was observed ($P = 0.025$) with the historical mobility scores.

of the parameters (mAVG, mMAX, mMIN, and mPLS) yielded a significant AUC to define a classification threshold in cows at first parity. The highest AUC with the highest upper CI bound (0.55, 95% CI: 0.49–0.61, $P = 0.095$) was produced for mAVG.

Considering the accurate detection of cows with severe lesions, the overall threshold of 11.9% for mPLS produced the highest Se (0.76, 95% CI: 0.70–0.82), whereas the threshold of 57.5 for mMAX produced the highest AUC (0.73, 95% CI: 0.0.69–0.76), and the threshold of 45.9 for mAVG produced the highest Acc (0.71, 95% CI: 0.69–0.73). Across parities, the highest Se (0.80, 95% CI: 0.70–0.87) was achieved for mPLS in cows at third parity. The threshold of 46.5 for mMAX produced a notable Se (0.76, 95% CI: 0.43–0.85) in cows at first parity, but with relatively poor discriminative performance (AUC = 0.66, 95% CI: 0.54–0.77).

The measures of accuracy for each parameter derived from the automated mobility scores 30 d to 1 DBT in detecting the presence of each lesion separately are detailed in Table 7. All parameters produced thresholds with similar AUC in detecting the presence of SH grade 3, but the threshold of 30.5 for mMIN achieved the highest Se (0.62, 95% CI: 0.55–0.68), and the threshold of 64.5 for mMAX achieved the highest Sp (0.89, 95% CI: 0.87–0.90). Similar results were observed for detecting any SU, with the threshold of 29.5 for mMIN achieving the highest Se (0.77, 95% CI: 0.65–0.86), and the threshold of 65.5 for mMAX achieving the highest Sp (0.90, 95% CI: 0.89–0.92). Regarding detection of WL grade 3, the threshold of 57.5 for mMAX produced the best AUC (0.79, 95% CI: 0.74–0.83) with moderate Se (0.74, 95% CI: 0.65–0.81) and Sp (0.71, 95%: 0.68–0.73), whereas mPLS at 11.9% produced a notably high Se (0.81, 95% CI: 0.74–0.88), but was followed by a low Sp (0.53, 95% CI: 0.51–0.56). None of the examined parameters (mAVG, mMAX, mMIN, and mPLS) yielded a significant AUC to define a classification threshold in detecting the few TU cases recorded. The highest AUC with the highest upper CI bound (0.61, 95% CI: 0.34–0.88, $P = 0.371$) was produced for mAVG. Finally, the thresholds of 54.5 for mMAX and 20.3% for mPLS produced the best Se (0.66 and 0.65, respectively) in detecting DD grade 3, whereas the threshold of 46 for mAVG produced the highest Sp (0.74, 95% CI: 0.72–0.76) and Acc (0.73, 95% CI: 0.71–0.75).

DISCUSSION

This study evaluated the performance of a fully automated 2-dimensional imaging system using machine learning for real-time lameness detection across a large dataset of mobility scores and foot lesions records, collected from 11 commercial UK dairy farms. We demonstrated that the system achieved substantial to almost perfect agreement with trained human observers for detecting lameness (when evaluated using Gwet's agreement coefficients) and identified cows with foot lesions with comparable accuracy. Notably, the system showed improved sensitivity in detecting severe lesions compared with human mobility scores, when using optimal threshold of parameters describing the mobility pattern

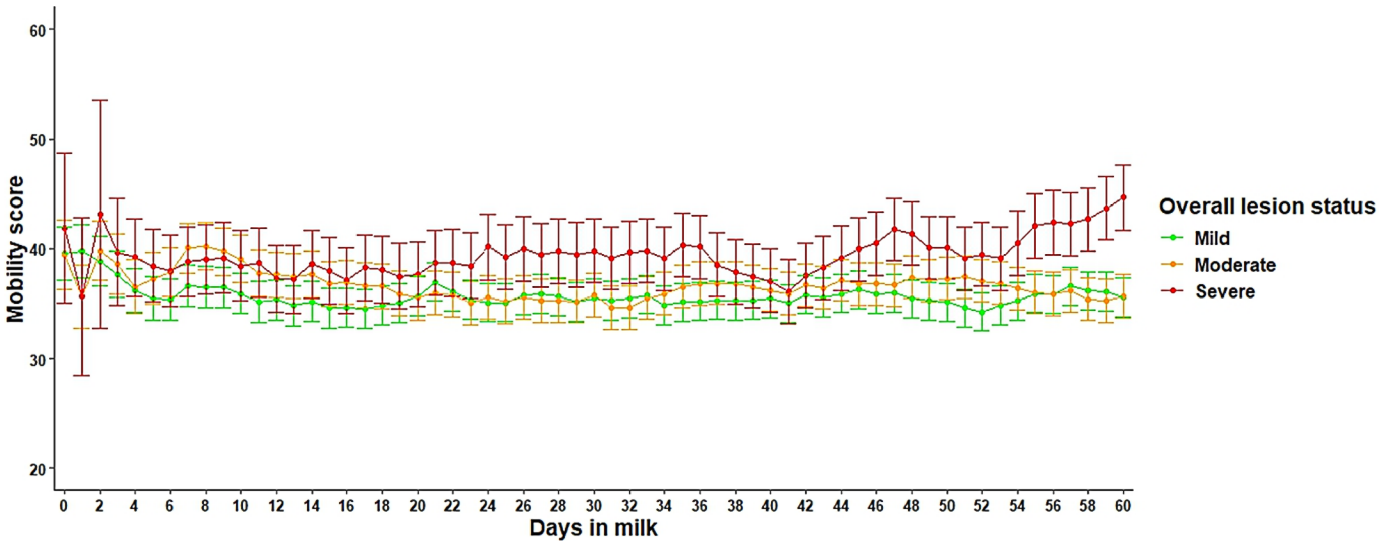

Figure 3. Estimated marginal means (±95% CI) derived from linear mixed models accounting for farm and parity effects, showing the evolution of daily automated mobility scores tracked during the first 60 DIM in 615 cows of 5 farms that were trimmed between 60 and 120 DIM classified in 3 levels according to the presence and severity of foot lesions identified during the trimming session. A statistically significant association of the overall lesion status ($P = 0.042$) and of the overall lesion status × DIM interaction was observed ($P < 0.001$) with the historical mobility scores.

Table 6. Overall and per parity measures of accuracy Se, Sp, Acc) for optimal thresholds, derived from receiver operating characteristic curves, for the average (mAVG), maximum (mMAX), minimum (mMIN), and percentage of scores that a cow was scored as lame (mPLS) during the past 30 d before foot trimming in 1,986 cows of 5 dairy farms, in correctly detecting cows bearing at least one case of moderate and severe foot lesions using the recordings from a trained veterinarian as ground truth (exact Clopper-Pearson binomial 95% CI for Se, Sp and Acc are shown in parentheses)

	Moderate and severe						Severe					
Item	Cut-off	AUC[1] (95% CI)	*P*-value	Se	Sp	Acc	Cut-off	AUC (95% CI)	*P*-value	Se	Sp	Acc
Overall, n[2] = 837/1,986							Overall, n = 228/1,986					
mAVG	44.5	0.58 (0.56–0.61)	<0.001	0.44 (0.40–0.47)	0.70 (0.67–0.73)	0.59 (0.57–0.62)	45.9	0.69 (0.65–0.73)	<0.001	0.55 (0.48–0.61)	0.73 (0.71–0.75)	0.71 (0.69–0.73)
mMAX	58.5	0.60 (0.57–0.62)	<0.001	0.41 (0.37–0.44)	0.77 (0.74–0.79)	0.62 (0.59–0.64)	57.5	0.73 (0.69–0.76)	<0.001	0.65 (0.59–0.72)	0.70 (0.68–0.72)	0.69 (0.67–0.71)
mMIN	32.5	0.57 (0.54–0.60)	<0.001	0.45 (0.42–0.49)	0.66 (0.63–0.69)	0.57 (0.55–0.59)	32.5	0.65 (0.61–0.69)	<0.001	0.57 (0.50–0.64)	0.64 (0.61–0.66)	0.63 (0.61–0.65)
mPLS	21.2	0.58 (0.55–0.60)	<0.001	0.48 (0.45–0.52)	0.64 (0.61–0.66)	0.57 (0.55–0.59)	11.9	0.69 (0.65–0.72)	<0.001	0.76 (0.70–0.82)	0.52 (0.50–0.55)	0.55 (0.53–0.57)
Parity 1, n = 142/348							Parity 1, n = 21/348					
mAVG	NA[3]	0.55 (0.49–0.61)	0.095	NA	NA	NA	37.0	0.64 (0.52–0.76)	0.029	0.71 (0.48–0.89)	0.58 (0.52–0.63)	0.59 (0.53–0.64)
mMAX	NA	0.54 (0.48–0.60)	0.192	NA	NA	NA	46.5	0.66 (0.54–0.77)	0.015	0.76 (0.43–0.85)	0.53 (0.50–0.61)	0.54 (0.51–0.61)
mMIN	NA	0.54 (0.48–0.60)	0.196	NA	NA	NA	NA	0.60 (0.47–0.73)	0.122	NA	NA	NA
mPLS	NA	0.51 (0.44–0.57)	0.860	NA	NA	NA	NA	0.61 (0.48–0.73)	0.102	NA	NA	NA
Parity 2, n = 125/367							Parity 2, n = 29/367					
mAVG	43.6	0.57 (0.51–0.63)	0.030	0.38 (0.29–0.47)	0.76 (0.70–0.81)	0.63 (0.58–0.68)	41.6	0.67 (0.57–0.77)	0.002	0.66 (0.46–0.82)	0.65 (0.60–0.71)	0.65 (0.60–0.70)
mMAX	56.5	0.61 (0.54–0.67)	0.001	0.41 (0.32–0.50)	0.78 (0.72–0.83)	0.65 (0.60–0.70)	54.5	0.72 (0.62–0.82)	<0.001	0.72 (0.53–0.87)	0.69 (0.63–0.74)	0.69 (0.64–0.74)
mMIN	NA	0.53 (0.47–0.59)	0.333	NA	NA	NA	25.5	0.63 (0.52–0.73)	0.025	0.79 (0.60–0.92)	0.42 (0.36–0.47)	0.45 (0.40–0.50)
mPLS	17.0	0.59 (0.53–0.65)	0.004	0.44 (0.35–0.53)	0.73 (0.67–0.78)	0.63 (0.58–0.68)	21.8	0.69 (0.59–0.79)	0.001	0.66 (0.46–0.82)	0.73 (0.67–0.77)	0.72 (0.67–0.76)
Parity 3, n = 254/669							Parity 3, n = 88/669					
mAVG	40.8	0.56 (0.52–0.61)	0.009	0.66 (0.60–0.72)	0.44 (0.39–0.49)	0.52 (0.49–0.56)	46.0	0.68 (0.62–0.74)	<0.001	0.58 (0.47–0.68)	0.70 (0.66–0.73)	0.68 (0.64–0.72)
mMAX	59.5	0.58 (0.54–0.63)	<0.001	0.39 (0.33–0.45)	0.78 (0.74–0.82)	0.63 (0.59–0.67)	62.5	0.74 (0.68–0.80)	<0.001	0.55 (0.44–0.65)	0.85 (0.81–0.87)	0.81 (0.77–0.84)
mMIN	29.5	0.56 (0.51–0.60)	0.013	0.67 (0.61–0.73)	0.43 (0.38–0.48)	0.52 (0.48–0.56)	29.5	0.62 (0.56–0.68)	<0.001	0.77 (0.67–0.86)	0.41 (0.37–0.46)	0.46 (0.42–0.50)
mPLS	17.0	0.56 (0.51–0.60)	0.017	0.54 (0.48–0.60)	0.53 (0.48–0.58)	0.53 (0.50–0.57)	12.3	0.68 (0.62–0.74)	<0.001	0.80 (0.70–0.87)	0.48 (0.44–0.52)	0.52 (0.48–0.56)
Parity 4+, n = 311/590							Parity 4+, n = 90/590					
mAVG	44.6	0.59 (0.55–0.64)	<0.001	0.61 (0.55–0.66)	0.57 (0.51–0.63)	0.59 (0.55–0.63)	46.1	0.66 (0.59–0.72)	<0.001	0.69 (0.58–0.78)	0.56 (0.51–0.60)	0.58 (0.54–0.62)
mMAX	58.5	0.62 (0.58–0.67)	<0.001	0.53 (0.47–0.59)	0.70 (0.64–0.75)	0.61 (0.57–0.65)	57.5	0.70 (0.64–0.76)	<0.001	0.73 (0.63–0.82)	0.59 (0.55–0.63)	0.61 (0.57–0.65)
mMIN	35.5	0.58 (0.54–0.63)	<0.001	0.49 (0.43–0.54)	0.67 (0.62–0.73)	0.58 (0.53–0.61)	32.5	0.62 (0.56–0.69)	<0.001	0.69 (0.58–0.78)	0.48 (0.44–0.53)	0.51 (0.47–0.56)
mPLS	38.2	0.58 (0.54–0.63)	0.001	0.49 (0.43–0.55)	0.66 (0.57–0.71)	0.57 (0.53–0.61)	41.6	0.65 (0.58–0.71)	<0.001	0.58 (0.47–0.68)	0.63 (0.59–0.67)	0.63 (0.58–0.66)

[1]AUC = area under the curve.

[2]n = number of actual positive cows.

[3]NA = not applicable.

during the last 30 DBT. Additionally, the system's ability to track mobility changes over time highlighted its potential for earlier detection of cows at risk of developing foot lesions, supporting its use in proactive lameness management.

The interobserver agreement between the automated system and the human scorers presented variability depending on the metrics used. Regarding the agreement on the 4-level absolute scores between AIMS and HMS, obtained Cohen's κ_w indicated only fair agreement, whereas the quadratically weighted Gwet's AC_2 were within the almost perfect agreement range. Accordingly, HA1 and HA2 attained a Cohen's κ_w that fell within the range of estimates obtained between the system and the human scorers. Gwet's AC_2 was in the substantial agreement range, but it was lower than the overall estimates obtained between the system and any human scorer.

When we evaluated the 2-level scale agreement between AIMS_BIN and HMS_BIN, we found that PA consistently exceeded the benchmark of accepted reliability, Gwet's AC_1 indicated substantial and almost perfect agreement, and Cohen's κ fell within the fair agreement range. The PA between HA1 and HA2 was lower than that for overall measurements between the system and any assessor. Cohen's κ fell within the range of estimates obtained between the system and the human scorers, and Gwet's AC_1 indicated substantial agreement, but was again lower than the overall estimates obtained between the system and any human scorer.

This discrepancy between the different metrics could be attributed to a statistical phenomenon called the "kappa paradox," which is defined by low kappa values in the presence of high percent agreement and is affected by marginal distributions and the low or high prevalence of the trait being studied (Byrt et al., 1993; Vanhoudt et al., 2019). Use of kappa has been questioned in several medical studies due to paradoxically poor reliability in disharmony with the percentage level of agreement (Wongpakaran et al., 2013; Cibulka and Strube, 2021). Gwet's coefficient is considered an improved alternative to kappa and a more stable estimate of chance-corrected agreement under low prevalence of the examined trait scenarios (Gwet, 2008).

A few studies have evaluated the interobserver agreement among different assessors when scoring cows on farm. Thomsen et al. (2008) found weighted kappa values between 0.24 and 0.68 among 5 different observers using a 5-level scale. Linardopoulou et al., (2022) found a wide range of kappa values (0.00–0.57) among human scorers on a 2-level agreement. Anagnostopoulos et al. (2023) reported a Cohen's κ_w of 0.41 and Gwet's AC_2 of 0.85 for the 4-grade scoring, and PA of 88.2%, Cohen's κ of 0.42 and Gwets's AC_1 of 0.81 for the binary converted classification into lame or nonlame. Improved interobserver agreement metrics have been reported in studies evaluating mobility from video recordings. Gardenier et al. (2021) reported PA of 56% and κ_w of 0.59 for the 4-level scale, and PA of 79% and κ of 0.57 for the 2-level scale, respectively, using the Dairy Australia Healthy Hooves 4-point locomotion system on videos from 50 cows. Similarly, Schlageter-Tello et al. (2014) reported PA of 57% and κ_w of 0.65 on the 5-level scale, and PA of 85.2% and κ of 0.70 for the 2-level scale, among 10 observers using a 5-point locomotion scoring system on 58 videos of cows equally representing all locomotion scores.

Our results suggest that the system's weekly average mobility scores align well with human observations, comparable to the level of agreement reported in the literature between trained HA scoring on site and to that between HA1 and HA2. Unlike humans, the system is capable of consistently scoring large numbers of cows daily without fatigue or disruptions in cow flow, minimizing the variability linked to different backgrounds and levels of experience (Garcia et al., 2015).

The automated system showed reasonable accuracy in predicting the presence of moderate and severe lesions. Using AIMS_BIN, the system achieved the same overall Se as with HMS_BIN (0.37 vs. 0.38), and even surpassed the human scorer in parity 3 cows, although the human was always more specific. Both the system and the human were less sensitive and more specific in heifers, this would lead to more heifers bearing foot lesions to remain undetected relative to older cows. The AIMS was more sensitive in older cows, meaning that more cows with foot lesions were correctly identified. This variability in the obtained Se across parities, is in accordance with previous findings with human mobility scoring reported by Logan et al. (2024) and implies that signs of lameness in heifers are more subtle than in older cows. The higher prevalence of lesions in older cows could likely be another explanation. Although Se and Sp are theoretically unaffected by the prevalence of the tested trait, evidence suggests that higher prevalence is associated with improved Se and lower Sp estimations (Murad et al., 2023). Considering this, a parity-specific calibration of the system's algorithm or lowering the predetermined cut-off of 50 to define lameness in heifers may be worth considering. Logan et al. (2024) reported similar Se and Sp for mobility scoring using the AHDB scoring system in detecting cows with moderate lesions (case definition 2 in their study); a classification that excluded minor lesions and is comparable to the merged status 2 and 3 used in our study. However, the way the mobility scoring was performed, and the single scorer's background, training, and level of experience are not clearly described in their study. Further validation of mobility scoring as a means

to identify mild lesions is required, with clear descriptors of mobility scoring training and implementation.

Both AIMS_BIN and HMS_BIN demonstrated improved Se and Acc in detecting cows with at least one severe lesion. The system's performance was comparable to that of the human, but the human was generally more sensitive (0.60 vs. 0.53). The HMS_BIN achieved notably high Se in parity 4+ cows and AIMS_BIN in parity 3 cows, but at the expense of Sp. Although cows with obvious upper limb lameness were excluded, a thorough

Table 7. Overall and per parity measures of accuracy (Se, Sp, Acc) for optimal thresholds, derived from receiver operating characteristic curves, for the average (mAVG), maximum (mMAX), minimum (mMIN) and percentage of scores that a cow was scored as lame (mPLS) during the past 30 d before foot trimming in 1,986 cows of 5 dairy farms, in correctly predicting the presence of specific foot lesions, using the recordings from a trained veterinarian as ground truth[1]

Item	Cut-off	AUC[2] (95% CI)	*P*-value	Se (95% CI)	Sp (95% CI)	Acc (95% CI)
Sole hemorrhage n[3] = 201/1,770						
mAVG	43.6	0.61 (0.57–0.66)	<0.001	0.53 (0.46–0.60)	0.64 (0.62–0.67)	0.63 (0.61–0.65)
mMAX	64.5	0.60 (0.56–0.65)	<0.001	0.28 (0.22–0.35)	0.89 (0.87–0.90)	0.82 (0.80–0.84)
mMIN	30.5	0.61 (0.57–0.65)	<0.001	0.62 (0.55–0.68)	0.56 (0.54–0.59)	0.57 (0.54–0.59)
mPLS	35.5	0.59 (0.55–0.63)	<0.001	0.41 (0.34–0.48)	0.75 (0.73–0.77)	0.71 (0.69–0.73)
Sole ulcer n = 69/1,638						
mAVG	47.3	0.69 (0.63–0.75)	<0.001	0.51 (0.38–0.63)	0.78 (0.76–0.80)	0.77 (0.75–0.79)
mMAX	65.5	0.70 (0.63–0.76)	<0.001	0.42 (0.30–0.55)	0.90 (0.89–0.92)	0.88 (0.87–0.90)
mMIN	29.5	0.68 (0.62–0.74)	<0.001	0.77 (0.65–0.86)	0.51 (0.48–0.54)	0.52 (0.50–0.55)
mPLS	23.4	0.66 (0.59–0.73)	<0.001	0.65 (0.53–0.76)	0.65 (0.62–0.67)	0.65 (0.62–0.67)
White line n = 129/1,698						
mAVG	43.5	0.72 (0.68–0.77)	<0.001	0.70 (0.61–0.78)	0.64 (0.61–0.66)	0.64 (0.62–0.67)
mMAX	57.5	0.79 (0.74–0.83)	<0.001	0.74 (0.65–0.81)	0.71 (0.68–0.73)	0.71 (0.69–0.73)
mMIN	30.5	0.67 (0.62–0.71)	<0.001	0.69 (0.60–0.77)	0.56 (0.54–0.59)	0.57 (0.55–0.60)
mPLS	11.9	0.71 (0.67–0.76)	<0.001	0.81 (0.74–0.88)	0.53 (0.51–0.56)	0.55 (0.53–0.58)
Toe ulcer n = 6/1,575						
mAVG	NA	0.61 (0.34–0.88)	0.371	NA	NA	NA
mMAX	NA	0.61 (0.37–0.86)	0.371	NA	NA	NA
mMIN	NA	0.55 (0.27 – 0.84)	0.658	NA	NA	NA
mPLS	NA	0.56 (0.29 – 0.83)	0.608	NA	NA	NA
Digital dermatitis n = 83/1,652						
mAVG	46.0	0.66 (0.60–0.72)	<0.001	0.52 (0.41–0.63)	0.74 (0.72–0.76)	0.73 (0.71–0.75)
mMAX	54.5	0.66 (0.60–0.72)	<0.001	0.66 (0.55–0.76)	0.60 (0.58–0.63)	0.60 (0.58–0.63)
mMIN	32.5	0.63 (0.57–0.70)	<0.001	0.57 (0.45–0.67)	0.65 (0.63–0.67)	0.65 (0.62–0.67)
mPLS	20.3	0.66 (0.60–0.72)	<0.001	0.65 (0.54–0.75)	0.62 (0.60–0.64)	0.62 (0.60–0.65)

[1]The exact Clopper-Pearson binomial 95% CI for Se, Sp and Acc are shown in parentheses[1]

[2]AUC = area under the curve.

[3]n = number of actual positive cows/number of total cows eligible, after excluding cases with concomitant severe lesions.

clinical examination was not conducted systematically, which may have allowed musculoskeletal issues unrelated to foot lesions in older cows to go undetected, reducing specificity. It is interesting to note that in our study both the system and the human achieved higher Se compared with the findings of Logan et al. (2024) in detecting cows with severe lesions (case definition 3 in their study). However, human mobility scoring in Logan et al. (2024) was highly specific (overall Sp = 0.94). In their study, Logan et al. (2024) reported much lower Se in heifers for detecting the presence of moderate and severe lesions (0.07 and 0.09, respectively) compared with our study, although the CI were wide. Variability across farms between the 2 studies, especially in the way mobility scoring was performed and case definition (i.e., the threshold for a case definition of lameness may have been lower in the current study), are likely causes for the observed differences in detecting foot lesions.

Using individual automated mobility scores tracking back 30 DBT, we determined optimal thresholds for mAVG, mMAX, mMIN, and mPLS. The use of any of these parameters resulted in improved Se in detecting moderate and severe lesions over the AIMS_BIN and the HMS_BIN, without significant decreases in Sp and Acc, although the human scorer remained more specific throughout. However, this was not the case for first parity cows where none produced a significant improvement in classification. Generally, the best approach to maximize Se (i.e., produce more true positives and fewer false negatives), would be to target cows that were scored as lame by the system for approximately more than a fifth of the times they were scored, whereas to maximize Sp (i.e., produce more true negatives and fewer false positives), it is best to target cows whose maximum score in the past month exceeded 58.5.

The use of any of the parameters derived from mobility patterns 30 DBT (mAVG, mMAX, mMIN, and mPLS) led to improved Se in detecting the presence of severe lesions over the AIMS_BIN, but at the expense of Sp. The thresholds produced for mMAX and mPLS achieved an overall Se higher than that of HMS_BIN, but the human scorer was more specific. Remarkably high sensitivities were obtained for mAVG and mMAX in detecting severe lesions in heifers (0.71 and 0.76, respectively), specificity though was poor. To maximize Se, targeting cows identified by the automated system as lame over 12% of the time in the past month is advisable. Whereas to maximize Sp, it is advisable to target cows with an average score above 46 in the past month.

To the best of our knowledge, there are no studies evaluating the accuracy of any mobility or locomotion scoring system detecting different foot lesions separately. The HMS_BIN was able to correctly detect all cows with TU, although cases were few, and cows with severe WL or with SU of any grade with sufficient Se (over 0.60). Detection of severe SH had a moderate Se, but Se for detection of DD grade 3 (the active M.2 lesions) was poor. The AIMS_BIN could detect TU, severe WL, SU of any grade, DD of grade 3 or severe SH with moderate Se (between 0.40 and 0.60), which were lower than those obtained by HMS_BIN for all foot lesion types except DD. Because we excluded cases with concomitant severe lesions to calculate the actual negatives for each lesion, this led to the production of the same Sp across all lesions, which can be interpreted as an overall Sp of human or automated mobility scoring to accurately detect any of these lesions. We should note that the system's Se in detecting DD grade 3 was higher than that of the human scorer. This implies that humans fail to detect the potentially abnormal gait of cows with painful active DD lesions by a single mobility assessment. The improved performance of AIMS_BIN could be attributed to the weekly average score's ability to detect the dynamic alterations in cow's gait over a week's course without the presence of a human interfering with the normal walk of the affected cows. This is especially important for younger cows in which DD is more prevalent (Smits et al., 1992; Somers et al., 2005) and which have a shorter flight distance and are more likely to exhibit fleeing behavior even in pain when a human is present (Phillips, 2002).

When using mAVG, mMAX, mMIN, and mPLS as parameters to describe the individual mobility pattern up to 30 DBT, we found that they provided improved measures of accuracy compared with AIMS_BIN for all lesions except TU. For TU, there were only 6 actual positive cases in the dataset, and none of the parameters could correctly discriminate them. In detecting severe SH and SU of any grade, mMIN had the best Se, and mMAX had the best Sp. Most parameters showed adequate Se in detecting severe WL, with mMAX producing the best Se-Sp-Acc combination. Lastly, in detecting DD grade 3, all tested parameters showed similar discrimination, with mMAX and mPLS having the best Se and mAVG having the best Sp. When looking at these parameters combined, the system could outperform the human scorer much the same as with the detection of moderate and of severe lesions previously described. With these results, it becomes clear that analyzing the mobility patterns in cows with foot lesions and even using this data to train the algorithm for early detection of lesion development is necessary.

The longitudinal study was performed on a single farm but provided useful insights into the temporal dynamics of automated mobility scores in cows that developed lesions during the early lactation stage. The examination of the hind feet of the enrolled cows within the first 4 to 10 DIM ensured a known history for each cow around calving and gave us the possibility to account for the presence

of pre-existing lesions that would affect mobility as we progress into lactation. By tracking the daily automated mobility scores from 5 to 64 DIM we were able to detect changes in what could be considered as a new phenotype for cattle lameness research and herd health management, the automated daily mobility score pattern of a cow. Our results showed that cows that were diagnosed with severe and even with moderate lesions at the early lactation foot trim, had higher scores throughout the first 2 mo of lactation and a notably greater day-to-day variation. Cows with severe lesions had higher scores that were clearly separated from as early as 36 DIM from cows with moderate and with mild lesions, indicating the potential to identify earlier cows at higher risk to develop severe foot pathologies during early lactation. Even when we performed the same analysis but focused only on cows that developed sole lesions, collectively referring to SH and SU, we observed a tendency for higher scores across the first 2 mo after calving. The incidence of sole lesions, peaks at 3 to 5 mo of lactation (Leach et al., 1997; Barker et al., 2009). These findings indicate the importance of carrying out a first routine trim early into lactation, as the associated changes in mobility occurred before 40 DIM in the study herds enrolled. Detecting these changes could help identify higher risk cows for an early lactation routine trim, while leaving those with good mobility for later. Research has indicated that a targeted early lactation intervention on heifers is cost beneficial over trimming all lame and nonlame heifers (Maxwell et al., 2015), and that foot trimming cows with good mobility induces stress leading to short-term decrease in activity, rumination, and milk production (Van Hertem et al., 2014). The efficacy of the system in detecting animals at risk of early-stage lesions, and the potential benefit derived from intervening in these animals warrants further investigation.

Moreover, a remark should be made about the increased scores and variability of cows with severe and with moderate lesions compared with those with no or mild lesions observed during the first 5 to 10 DIM: Pathological, systemic inflammation around calving could be a valid explanation. Sole lesions are understood to occur because of inflammation and deranged horn production caused by a biomechanical insult applied to keratinocytes within the corium (Bergsten, 1994). However, systemic inflammation could also act as the initial trigger for the development of claw horn disruption lesions by disrupting blood circulation inside the corium and driving change to the functional anatomy of the foot (Watson et al., 2022; Wilson et al., 2022). Identifying cows undergoing pathological systemic inflammation immediately after calving just from the mobility pattern would be an interesting field for further research.

Guided by the results we obtained from this longitudinal study, we looked into PriorDATA2, which included cows that were trimmed between 60 and 120 DIM. We then conducted the same analysis to assess whether these findings can be replicated in a larger sample size of cows from multiple farms. We still observed increased mobility scores during the first 60 DIM in cows with severe lesions, with differences being most apparent toward the end of this period. We were unable to confirm the differences observed shortly after calving in the longitudinal study. The fact that we did not have a known history for these cows is a limitation, although lesion status immediately after calving was not significantly associated with the evolution of mobility scores in the longitudinal study.

Using PriorDATA1, we observed that cows with severe lesions regardless of the stage of lactation had higher automated mobility scores from as early as 23 d before the trimming session. This indicates that the system's daily scores can provide early warnings for potential severe cases of lameness. Even when combining moderate and severe cases and comparing them to mild ones, this separation was still noticeable. The ability for early detection is essential for any lameness management protocol and if accompanied by timely and proper intervention, it could be a valuable tool in our efforts to control lameness and improve overall herd health and increase farm profitability. It has been shown that farmers are commonly underestimating the lameness prevalence in their herds and recognize milder, and even severe cases with a significant delay (Alawneh et al., 2012; Leach et al., 2012). Early detection and intervention are crucial in preventing the development of severe pathologies, promoting recovery and reducing recurrent cases (Leach et al., 2012; Groenevelt et al., 2014). The system's capacity to provide regular and frequent mobility scores provides a substantial advantage over the farmer's observations or even the human-conducted mobility score assessment, as it minimizes the chances of missing early signs of lameness that might be overlooked in less frequent, or inconsistent assessments of lameness.

We also highlighted the farm and parity effects on the daily scores, meaning that farm- and parity-specific adjustments may improve the algorithm's accuracy and reliability. Primiparous cows had lower mobility scores than older cows. Whether this is because primiparous cows have a lower lameness prevalence or because they manifest pain and lameness differently, remains unclear. Both seem plausible, the latter assumption though is corroborated by the lower thresholds identified from ROC analysis to predict foot lesions in primiparous cows. More data from longitudinal observations is required to address this issue. Optimizing the system's performance across different environments and herd demographics is a goal for future improvements. Artificial intelligence

applications can handle complex data and learn from experience without being programmed to and without compromising overall accuracy (Sarker, 2021).

Limitations

The present study has some limitations that need to be acknowledged. Although the study involved 4 experienced and trained researchers, it did not assess intra-observer variability. Including more observers with different backgrounds and levels of experience, would allow for a more rigorous assessment of the interobserver agreement between humans and make the study more representative of how human mobility scoring is performed in practice. The study was conducted on 11 commercial dairy farms housing in total more than 15,000 milking cows, which renders generalizability of our findings to intensively managed high-yielding Holstein cows. However, the fact that some of the same farms were also used for training the algorithm could be considered as a limitation. Nonetheless, because mobility, herd demographics, and environmental conditions are dynamic and we obtained similar results across farms, we consider this issue to have a minimum influence on our findings. Furthermore, the fact that a trained researcher collected the foot lesions data from a large number of cows is a strength of our study. Although these data were collected during routine and therapeutic trims, the random selection of several cows on each of the participating farms for foot inspection could provide more reliable results on the system's accuracy in detecting cows with foot lesions. Lastly, the longitudinal study was conducted on a single farm and the fact that foot trimming history for cows with daily mobility scores was unknown, offers indicative findings, but does not allow us to draw definite conclusions.

CONCLUSIONS

The automated 2-dimensional imaging system tested in this study demonstrated substantial agreement with human mobility scoring according to Gwet's agreement estimates, providing reliable detection of lame cows and cows with foot lesions across various dairy farms. The system showed sensitivity and specificity in detecting foot lesions comparable to those of a well-trained HA. Its capability to score cows frequently, consistently, and unobtrusively, providing daily mobility scores offered an advantage over the HA in terms of sensitivity. Moreover, we highlighted the system's potential for early intervention. Adoption of this system under a lameness management protocol could be useful in selecting cows for foot trimming, reducing lameness prevalence, preventing the development of severe lesions and improving overall health and welfare of dairy cows.

NOTES

The present study was funded by Innovate UK (Swindon, United Kingdom; Farming Innovation Programme Small R&D Partnership Projects, project no. 10027372). The authors are grateful to the farmers who participated in the study and to the professional foot trimmers who allowed us to collect data. Supplemental material for this article is available at http://doi.org/10.17632/533d5ttydp.1 (Siachos, 2024a). The study was approved by the University of Liverpool Veterinary Research Ethics Committee (Reference VREC1079). The authors have not stated any conflicts of interest.

Nonstandard abbreviations used: AC1 = unweighted Gwet's agreement coefficient; AC2 = quadratically weighted Gwet's agreement coefficient; ACC = classification accuracy; AHDB = Agricultural and Horticultural Development Board; AIMS = 4-grade converted weekly average mobility score; AIMS_BIN = binary converted weekly average mobility score; AUC = area under the curve; AWF = axial wall fissure; CE = XXXX; DBT = days before trimming; DD = digital dermatitis; ELRT = early lactation routine trim; EMM = estimated marginal means; FCT = fresh cow trim; HA = human assessors; HMS = human mobility scores; HMS_BIN = binary converted human mobility scores; IH = interdigital hyperplasia; IP = interdigital phlegmon; κ = unwighted Cohen's kappa; κ_w = quadratically weighted Cohen's kappa; LMM = linear mixed effects models; mAVG = monthly average score; mMAX = monthly maximum score; mMIN = monthly minimum score; mPLS = percentage of daily scores that a cow was recorded as lame; OLS = overall lesion status; OLS_BIN_SEV = binary OLS, mild and moderate vs. severe; OLS_BIN_MODSEV = binary OLS, mild vs. moderate and severe; PA = percentage agreement; PriorDATA1 = first dataset; PriorDATA2 = second dataset; ROC = receiver operating characteristic; RoMS = Register of Mobility Scorers; Se = sensitivity; SH = sole hemorrhage; Sp = specificity; SU = sole ulcer; TU = toe ulcer; WL = white line disease.

REFERENCES

Alawneh, J. I., R. A. Laven, and M. A. Stevenson. 2012. Interval between detection of lameness by locomotion scoring and treatment for lameness: A survival analysis. Vet. J. 193:622–625. https://doi.org/10.1016/J.TVJL.2012.06.042.

Anagnostopoulos, A., B. E. Griffiths, N. Siachos, J. Neary, R. F. Smith, and G. Oikonomou. 2023. Initial validation of an intelligent video surveillance system for automatic detection of dairy cattle lameness. Front. Vet. Sci. 10. https://doi.org/10.3389/fvets.2023.1111057.

Barker, Z. E., J. R. Amory, J. L. Wright, S. A. Mason, R. W. Blowey, and L. E. Green. 2009. Risk factors for increased rates of sole ulcers, white line disease, and digital dermatitis in dairy cattle from twenty-seven farms in England and Wales. J. Dairy Sci. 92:1971–1978. https://doi.org/10.3168/JDS.2008-1590.

Beggs, D. S., E. C. Jongman, P. E. Hemsworth, and A. D. Fisher. 2019. Lame cows on Australian dairy farms: A comparison of farmer-identified lameness and formal lameness scoring, and the position of lame cows within the milking order. J. Dairy Sci. 102:1522–1529. https://doi.org/10.3168/JDS.2018-14847.

Bergsten, C. 1994. Haemorrhages of the sole horn of dairy cows as a retrospective indicator of laminitis: An epidemiological study. Acta Vet. Scand. 35:55–66. https://doi.org/10.1186/BF03548355/METRICS.

Byrt, T., J. Bishop, and J. B. Carlin. 1993. Bias, prevalence and kappa. J. Clin. Epidemiol. 46:423–429.

Charfeddine, N., and M. A. Pérez-Cabal. 2017. Effect of claw disorders on milk production, fertility, and longevity, and their economic impact in Spanish Holstein cows. J. Dairy Sci. 100:653–665. https://doi.org/10.3168/JDS.2016-11434.

Cibulka, M. T., and M. J. Strube. 2021. The conundrum of kappa and why some musculoskeletal tests appear unreliable despite high agreement: A comparison of Cohen kappa and Gwet AC to assess observer agreement when using nominal and ordinal data. Phys. Ther. 101:1–5. https://doi.org/10.1093/PTJ/PZAB150.

Clopper, C. J., and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 25:404–413.

Cramer, G., K. D. Lissemore, C. L. Guard, K. E. Leslie, and D. F. Kelton. 2008. Herd- and cow-level prevalence of foot lesions in Ontario dairy cattle. J. Dairy Sci. 91:3888–3895. https://doi.org/10.3168/JDS.2008-1135.

Döpfer, D., M. Holzhauer, and M. van Boven. 2012. The dynamics of digital dermatitis in populations of dairy cattle: Model-based estimates of transition rates and implications for control. Vet. J. 193:648–653. https://doi.org/10.1016/J.TVJL.2012.06.047.

Egger-Danner, C., P. Nielsen, A. Fiedler, K. Müller, T. Fjeldaas, D. Döpfer, V. Daniel, C. Bergsten, G. Cramer, A. M. Christen, K. F. Stock, G. Thomas, M. Holzhauer, A. Steiner, J. Clarke, N. Capion, N. Charfeddine, E. Pryce, E. Oakes, J. Burgstaller, B. Heringstad, C. Ødegård, and J. Kofler. 2014. ICAR Claw Health Atlas. 2nd ed. ICAR Technical Series. ICAR, Rome, Italy.

Espejo, L. A., M. I. Endres, and J. A. Salfer. 2006. Prevalence of lameness in high-producing Holstein cows housed in freestall barns in Minnesota. J. Dairy Sci. 89:3052–3058. https://doi.org/10.3168/JDS.S0022-0302(06)72579-6.

Fabian, J., R. A. Laven, and H. R. Whay. 2014. The prevalence of lameness on New Zealand dairy farms: A comparison of farmer estimate and locomotion scoring. Vet. J. 201:31–38. https://doi.org/10.1016/J.TVJL.2014.05.011.

Flower, F. C., and D. M. Weary. 2006. Effect of hoof pathologies on subjective assessments of dairy cow gait. J. Dairy Sci. 89:139–146. https://doi.org/10.3168/JDS.S0022-0302(06)72077-X.

Garcia, E., K. König, B. H. Allesen-Holm, I. C. Klaas, J. M. Amigo, R. Bro, and C. Enevoldsen. 2015. Experienced and inexperienced observers achieved relatively high within-observer agreement on video mobility scoring of dairy cows. J. Dairy Sci. 98:4560–4571. https://doi.org/10.3168/jds.2014-9266.

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

Gibbons, J., E. Vasseur, J. Rushen, and A. M. De Passillé. 2012. A training programme to ensure high repeatability of injury scoring of dairy cows. Anim. Welf. 21:379–388. https://doi.org/10.7120/09627286.21.3.379.

Griffiths, B. E., D. G. White, and G. Oikonomou. 2018. A cross-sectional study into the prevalence of dairy cattle lameness and associated herd-level risk factors in England and Wales. Front. Vet. Sci. 5:65. https://doi.org/10.3389/FVETS.2018.00065/BIBTEX.

Groenevelt, M., D. C. J. Main, D. Tisdall, T. G. Knowles, and N. J. Bell. 2014. Measuring the response to therapeutic foot trimming in dairy cows with fortnightly lameness scoring. Vet. J. 201:283–288. https://doi.org/10.1016/J.TVJL.2014.05.017.

Gwet, J. 2001. Handbook of Inter-Rater Reliability. Gaithersburg, MD: STATAXIS Publishing Company.

Gwet, K. L. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. 61:29–48. https://doi.org/10.1348/000711006X126600.

Hoblet, K. H., and W. Weiss. 2001. Metabolic hoof horn disease claw horn disruption. Vet. Clin. North Am. Food Anim. Pract. 17:111–127. https://doi.org/10.1016/S0749-0720(15)30057-8.

Jackson, A., M. J. Green, and J. Kaler. 2022. Fellow cows and conflicting farmers: Public perceptions of dairy farming uncovered through frame analysis. Front. Vet. Sci. 9:995240. https://doi.org/10.3389/FVETS.2022.995240/BIBTEX.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159. https://doi.org/10.2307/2529310.

Leach, K. A., D. N. Logue, S. A. Kempson, J. E. Offer, H. E. Ternent, and J. M. Randall. 1997. Claw lesions in dairy cattle: Development of sole and white line haemorrhages during the first lactation. Vet. J. 154:215–225. https://doi.org/10.1016/S1090-0233(97)80024-X.

Leach, K. A., D. A. Tisdall, N. J. Bell, D. C. J. Main, and L. E. Green. 2012. The effects of early treatment for hindlimb lameness in dairy cows on four commercial UK farms. Vet. J. 193:626–632. https://doi.org/10.1016/J.TVJL.2012.06.043.

Lenth, R. 2024. emmeans: Estimated marginal means, aka least squares means. R package version 1.10.1. Accessed May 17, 2024. https://CRAN.R-project.org/package=emmeans.

Linardopoulou, K., L. Viora, F. Fioranelli, J. Kernec, Q. Abbasi, G. King, E. Borelli, and N. Jonsson. 2022. Time-series observations of cattle mobility: accurate label assignment from multiple assessors, and association with lesions detected in the feet. Page 297 in Proceedings of the 31st World Buiatrics Congress, Madrid, Spain.

Logan, F., C. G. McAloon, E. G. Ryan, L. O'Grady, M. Duane, B. Deane, and C. I. McAloon. 2024. Sensitivity and specificity of mobility scoring for the detection of foot lesions in pasture-based Irish dairy cows. J. Dairy Sci. 107:3197–3206. https://doi.org/10.3168/JDS.2023-23928.

Maxwell, O. J. R., C. D. Hudson, and J. N. Huxley. 2015. Effect of early lactation foot trimming in lame and non-lame dairy heifers: A randomised controlled trial. Vet. Rec. 177:100. https://doi.org/10.1136/vr.103155.

McHugh, M. L. 2012. Interrater reliability: The kappa statistic. Biochem. Med. (Zagreb) 22:276. https://doi.org/10.11613/bm.2012.031.

Murad, M. H., L. Lin, H. Chu, B. Hasan, R. A. Alsibai, A. S. Abbas, R. A. Mustafa, and Z. Wang. 2023. The association of sensitivity and specificity with disease prevalence: Analysis of 6909 studies of diagnostic test accuracy. Can. Med. Assoc. J. 195:E925–E931. https://doi.org/10.1503/cmaj.221802/tab-related-content.

Murray, R. D., D. Y. Downham, M. J. Clarkson, W. B. Faull, J. W. Hughes, F. J. Manson, J. B. Merritt, W. B. Russell, J. E. Sutherst, and W. R. Ward. 1996. Epidemiology of lameness in dairy cattle: Description and analysis of foot lesions. Vet. Rec. 138:586–591. https://doi.org/10.1136/VR.138.24.586.

Nejati, A., A. Bradtmueller, E. Shepley, and E. Vasseur. 2023. Technology applications in bovine gait analysis: A scoping review. PLoS One 18:e0266287. https://doi.org/10.1371/journal.pone.0266287.

Newsome, R., M. J. Green, N. J. Bell, M. G. G. Chagunda, C. S. Mason, C. S. Rutland, C. J. Sturrock, H. R. Whay, and J. N. Huxley. 2016. Linking bone development on the caudal aspect of the distal phalanx with lameness during life. J. Dairy Sci. 99:4512–4525. https://doi.org/10.3168/JDS.2015-10202.

Nielsen, B. H., P. T. Thomsen, L. E. Green, and J. Kaler. 2012. A study of the dynamics of digital dermatitis in 742 lactating dairy cows. Prev. Vet. Med. 104:44–52. https://doi.org/10.1016/J.PREVETMED.2011.10.002.

O'Leary, N. W., D. T. Byrne, A. H. O'Connor, and L. Shalloo. 2020. Invited review: Cattle lameness detection with accelerometers. J. Dairy Sci. 103:3895–3911. https://doi.org/10.3168/jds.2019-17123.

Omontese, B. O., R. Bellet-Elias, A. Molinero, G. D. Catandi, R. Casagrande, Z. Rodriguez, R. S. Bisinotto, and G. Cramer. 2020. Association between hoof lesions and fertility in lactating Jersey cows. J. Dairy Sci. 103:3401–3413. https://doi.org/10.3168/jds.2019-17252.

Pedersen, S., and J. Wilson. 2021. Early detection and prompt effective treatment of lameness in dairy cattle. Livestock (Lond.) 26:115–121. https://doi.org/10.12968/live.2021.26.3.115.

Phillips, C. 2002. The relationship between cattle and man. Pages 217–224 in Cattle Behaviour & Welfare, C. Phillips, ed. Blackwell Science Ltd.

Pinheiro, J., and D. Bates. 2023. nlme: Linear and nonlinear mixed effects models. R package version 3.1–162. Accessed May 17, 2024. https://cran.r-project.org/package=nlme.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Randall, L. V., M. J. Green, L. E. Green, M. G. G. Chagunda, C. Mason, S. C. Archer, and J. N. Huxley. 2018. The contribution of previous lameness events and body condition score to the occurrence of lameness in dairy herds: A study of 2 herds. J. Dairy Sci. 101:1311–1324. https://doi.org/10.3168/JDS.2017-13439.

RoMS (Register of Mobility Scorers). 2024. The Register of Mobility Scorers. Accessed Sep. 3, 2024. https://roms.org.uk/.

Sarker, I. H. 2021. Machine learning: Algorithms, real-world applications and research directions. SN Comput. Sci. 2:160. https://doi.org/10.1007/s42979-021-00592-x.

Schlageter-Tello, A., E. A. M. Bokkers, P. W. G. Groot Koerkamp, T. Van Hertem, S. Viazzi, C. E. B. Romanini, I. Halachmi, C. Bahr, D. Berckmans, and K. Lokhorst. 2014. Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement. J. Dairy Sci. 97:5533–5542. https://doi.org/10.3168/jds.2014-8129.

Siachos, N. 2024a. Supplemental material: Evaluation of a fully automated 2D imaging system for real-time cattle lameness detection using machine learning. Mendeley Data, V1. 10.17632/533d5ttydp.1.

Siachos, N., J. M. Neary, R. F. Smith, and G. Oikonomou. 2024b. Automated dairy cattle lameness detection utilizing the power of artificial intelligence; current status quo and future research opportunities. Vet. J. 304:106091. https://doi.org/10.1016/j.tvjl.2024.106091.

Smits, M. C. J., K. Frankena, J. H. M. Metz, and J. P. T. M. Noordhuizen. 1992. Prevalence of digital disorders in zero-grazing dairy cows. Livest. Prod. Sci. 32:231–244. https://doi.org/10.1016/S0301-6226(12)80004-2.

Somers, J. G. C. J., K. Frankena, E. N. Noordhuizen-Stassen, and J. H. M. Metz. 2005. Risk factors for digital dermatitis in dairy cows kept in cubicle houses in the Netherlands. Prev. Vet. Med. 71:11–21. https://doi.org/10.1016/j.prevetmed.2005.05.002.

Sprecher, D. J., D. E. Hostetler, and J. B. Kaneene. 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. Theriogenology 47:1179–1187. https://doi.org/10.1016/S0093-691X(97)00098-8.

Stygar, A. H., Y. Gómez, G. V. Berteselli, E. Dalla Costa, E. Canali, J. K. Niemi, P. Llonch, and M. Pastell. 2021. A systematic review on commercially available and validated sensor technologies for welfare assessment of dairy cattle. Front. Vet. Sci. 8:634338. https://doi.org/10.3389/fvets.2021.634338/bibtex.

Swartz, D., E. Shepley, K. P. Gaddis, J. Burchard, and G. Cramer. 2024. Descriptive evaluation of a camera-based dairy cattle lameness detection technology. J. Dairy Sci. 107. https://doi.org/10.3168/JDS.2024-24851.

Thomas, H. J., G. G. Miguel-Pacheco, N. J. Bollard, S. C. Archer, N. J. Bell, C. Mason, O. J. R. Maxwell, J. G. Remnant, P. Sleeman, H. R. Whay, and J. N. Huxley. 2015. Evaluation of treatments for claw horn lesions in dairy cows in a randomized controlled trial. J. Dairy Sci. 98:4477–4486. https://doi.org/10.3168/JDS.2014-8982.

Thomas, H. J., J. G. Remnant, N. J. Bollard, A. Burrows, H. R. Whay, N. J. Bell, C. Mason, and J. N. Huxley. 2016. Recovery of chronically lame dairy cows following treatment for claw horn lesions: A randomised controlled trial. Vet. Rec. 178:116. https://doi.org/10.1136/VR.103394.

Thomsen, P. T., L. Munksgaard, and F. A. Togersen. 2008. Evaluation of a lameness scoring system for dairy cows. J. Dairy Sci. 91:119–126. https://doi.org/10.3168/jds.2007-0496.

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

Van Hertem, T., Y. Parmet, M. Steensels, E. Maltz, A. Antler, A. A. Schlageter-Tello, C. Lokhorst, C. E. B. Romanini, S. Viazzi, C. Bahr, D. Berckmans, and I. Halachmi. 2014. The effect of routine hoof trimming on locomotion score, ruminating time, activity, and milk yield of dairy cows. J. Dairy Sci. 97:4852–4863. https://doi.org/10.3168/jds.2013-7576.

Van Nuffel, A., I. Zwertvaegher, L. Pluym, S. Van Weyenberg, V. M. Thorup, M. Pastell, B. Sonck, and W. Saeys. 2015. Lameness detection in dairy cows: Part 1. How to distinguish between non-lame and lame cows based on differences in locomotion or behavior. Animals (Basel) 5:838–860. https://doi.org/10.3390/ani5030387.

Vanhoudt, A., D. A. Yang, T. Armstrong, J. N. Huxley, R. A. Laven, A. D. Manning, R. F. Newsome, M. Nielen, T. van Werven, and N. J. Bell. 2019. Interobserver agreement of digital dermatitis M-scores for photographs of the hind feet of standing dairy cattle. J. Dairy Sci. 102:5466–5474. https://doi.org/10.3168/jds.2018-15644.

Waiblinger, S., C. Menke, and D. W. Fölsch. 2003. Influences on the avoidance and approach behaviour of dairy cows towards humans on 35 farms. Appl. Anim. Behav. Sci. 84:23–39. https://doi.org/10.1016/S0168-1591(03)00148-5.

Watson, C., M. Barden, B. E. Griffiths, A. Anagnostopoulos, H. M. Higgins, C. Bedford, S. Carter, A. Psifidi, G. Banos, and G. Oikonomou. 2022. Prospective cohort study of the association between early lactation mastitis and the presence of sole ulcers in dairy cows. Vet. Rec. 190:e1387. https://doi.org/10.1002/vetr.1387.

Whay, H. R., D. C. J. Main, L. E. Green, and A. J. F. Webster. 2003. Assessment of the welfare of dairy cattle using animal-based measurements: Direct observations and investigation of farm records. Vet. Rec. 153:197–202. https://doi.org/10.1136/vr.153.7.197.

Whay, H. R., and J. K. Shearer. 2017. The impact of lameness on welfare of the dairy cow. Vet. Clin. North Am. Food Anim. Pract. 33:153–164. https://doi.org/10.1016/j.cvfa.2017.02.008.

Whay, H. R., A. E. Waterman, and A. J. F. Webster. 1997. Associations between locomotion, claw lesions and nociceptive threshold in dairy heifers during the peri-partum period. Vet. J. 154:155–161. https://doi.org/10.1016/S1090-0233(97)80053-6.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. 2019. Welcome to the Tidyverse. J. Open Source Softw. 4:1686. https://doi.org/10.21105/joss.01686.

Wilson, J. P., M. J. Green, L. V. Randall, C. S. Rutland, N. J. Bell, H. Hemingway-Arnold, J. S. Thompson, N. J. Bollard, and J. N. Huxley. 2022. Effects of routine treatment with nonsteroidal anti-inflammatory drugs at calving and when lame on the future probability of lameness and culling in dairy cows: A randomized controlled trial. J. Dairy Sci. 105:6041–6054. https://doi.org/10.3168/JDS.2021-21329.

Wongpakaran, N., T. Wongpakaran, D. Wedding, and K. L. Gwet. 2013. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. BMC Med. Res. Methodol. 13:61. https://doi.org/10.1186/1471-2288-13-61.

ORCIDS

N. Siachos, https://orcid.org/0000-0001-7670-4950
J. P. Wilson, https://orcid.org/0000-0001-7096-8007
A. Anagnostopoulos, https://orcid.org/0000-0002-5193-858X
J. M. Neary, https://orcid.org/0000-0001-8438-2234
R. F. Smith, https://orcid.org/0000-0003-0944-310X
G. Oikonomou https://orcid.org/0000-0002-4451-4199