J. Dairy Sci. 107:2499–2511
https://doi.org/10.3168/jds.2023-23894

Development and validation of a fully automated 2-dimensional imaging system generating body condition scores for dairy cows using machine learning

N. Siachos,[1] M. Lennox,[2] A. Anagnostopoulos,[1] B. E. Griffiths,[1] J. M. Neary,[1] R. F. Smith,[1] and G. Oikonomou[1]*
[1]Department of Livestock and One Health, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, CH64 7TE, United Kingdom
[2]CattleEye Ltd., The Innovation Centre, Queens Road, Belfast BT3 9DT, United Kingdom

ABSTRACT

Monitoring body condition score (BCS) is a useful management tool to estimate the energy reserves of an individual cow or a group of cows. The aim of this study was to develop and evaluate the performance of a fully automated 2-dimensional imaging system using a machine learning algorithm to generate real-time BCS for dairy cows. Two separate datasets were used for training and testing. The training dataset included 34,150 manual BCS (MAN_BCS) assigned by 5 experienced veterinarians during 35 visits at 7 dairy farms. Ordinal regression methods and deep learning architecture were used when developing the algorithm. Subsequently, the testing dataset was used to evaluate the developed BCS prediction algorithm on 4 of the participating farms. An experienced human assessor (HA1) visited these farms and performed 8 whole-milking-herd BCS sessions. Each farm was visited twice, allowing for 30 d (±2 d) to pass between visits. The MAN_BCS assigned by HA1 were considered the ground truth data. At the end of the validation study, MAN_BCS were merged with the stored automated BCS (AI_BCS), resulting in a testing dataset of 9,657 single BCS. A total of 3,817 cows in the testing dataset were scored twice 30 d (±2 d) apart, and the change in their BCS (ΔBCS) was calculated. A subset of cows at one farm were scored twice on consecutive days to evaluate the within-observer agreement of both the human assessor and the system. The manual BCS of 2 more assessors (HA2 and HA3) were used to assess the interobserver agreement between humans. Finally, we also collected ultrasound measurements of backfat thickness (BFT) from 111 randomly selected cows with available MAN_BCS and AI_BCS. Using the testing dataset, intra- and interobserver agreement for single BCS and ΔBCS were estimated by calculating the simple percentage agreement (PA) at 3 error levels and the weighted kappa (κ_w) for the exact agreement. A Bland-Altman plot was constructed to visualize the systematic and proportional bias. The association between MAN_BCS and AI_BCS and the BFT was assessed with Passing-Bablok regressions. The system had an almost perfect repeatability with a κ_w of 0.99. The agreement between MAN_BCS and AI_BCS was substantial, with an overall κ_w of 0.69. The overall PA at the exact, ± 0.25-unit, and ± 0.50-unit BCS error range between MAN_BCS and AI_BCS was 44.4%, 84.6%, and 94.8%, respectively, and greater than the PA obtained between HA1 and HA3. The Bland-Altman plot revealed a minimal systematic bias of −0.09 with a proportional bias at the extreme scores. Furthermore, despite the low κ_w of 0.20, the overall PA at the exact and ± 0.25-unit of BCS error range between MAN_BCS and AI_BCS regarding the ΔBCS was 45.7 and 88.2%, respectively. A strong linear relationship was observed between BFT and AI_BCS ($\rho = 0.75$), although weaker than that between BFT and MAN_BCS ($\rho = 0.91$). The system was able to predict single BCS and ΔBCS with satisfactory accuracy, comparable to that obtained between trained human scorers.

Key words: artificial intelligence, cattle, convolutional neural network, body condition score

INTRODUCTION

Dairy cows, like all mammals, rely on mobilization of stored adipose tissue reserves to support milk production and nurture their offspring (Bauman and Currie, 1980). Dry matter intake increases gradually after calving, but in high yielding cows, it is insufficient to meet their energy requirements of early lactation (Grummer et al., 2004). Hence, most dairy cows are in a negative energy balance during the postpartum period (Drackley, 1999; Herdt, 2000), which has negative effects on overall health and productivity (Butler, 2005; Esposito

Received June 22, 2023.
Accepted October 30, 2023.
*Corresponding author: goikon@liv.ac.uk

et al., 2014). Dairy cows were found to mobilize significant amounts of fat and protein during the first 5 wk after calving (Komaragiri and Erdman, 1997).

Body condition scoring is considered a valuable tool to monitor the energy reserves of a dairy cow or a group of cows. Changes in body condition over time reflect the amount of dietary energy consumption relative to energy requirements under the current level of milk production (Roche et al., 2009). Body condition score can predict the total amount of fat in a dairy cow (Wright and Russel, 1984). A one-unit increase in BCS has been associated with an average increase in BW of ~50 kg (Otto et al., 1991).

Several methods have been developed for the estimation of BCS using different scales based on either a tactile or visual assessment. A comparison of these systems has been provided by Roche et al. (2004). The visual assessment method using the 1 to 5 scale with increments of 0.25 developed by Ferguson et al. (1994) by modifying the system developed by Edmonson et al. (1989) is currently the most commonly referenced system for Holstein cattle globally. Cows with a BCS of 1 are considered emaciated and cows with a BCS of 5 are extremely obese. This method requires the observation of a cow from the side and the rear, and the assessment (including tactile assessment) of fat deposition in several anatomic features in the pelvis, tailhead, and lumbar area. It can be performed quickly by a trained person. Interobserver percentage agreement (**PA**) following this method was 58% at the exact score and 91% within the 0.25-point level of error (Ferguson et al., 1994).

It has been well documented that BCS at calving, the nadir BCS after calving, and the loss during the postpartum period have significant implications for overall health, fertility, and milk production (Roche et al., 2009, 2013). A target range of 3.00 to 3.25 at calving is recommended to optimize health and production outcomes (Roche et al., 2009). However, the changes in body condition during critical periods in a cow's production cycle are even more important than single BCS at specific stages of lactation. Cows that gained body condition after calving had a significantly shorter calving to first ovulation interval and more pregnancies per artificial inseminations compared with those that maintained or lost body condition (Barletta et al., 2017). Low BCS has also been associated with an increased risk of developing lameness (Randall et al., 2015).

Although the importance of BCS monitoring in dairy herd management has been emphasized for many decades, it is rarely performed consistently in most herds unless for research purposes (Hady et al., 1994). When performed, either by nutritionists, veterinarians, or farmers, a representative small sample of cows per group is usually scored, and rarely are any records kept (Caraviello et al., 2006). Its assessment is labor-intensive and time-consuming, especially in large herds, requiring a trained person to perform it.

Over the last 20 years, many systems using computer vision technology have been developed attempting to automate BCS in order to overcome the obstacles of manual estimations. Initial attempts included the placement of a 2-dimensional (**2D**) camera to capture images (Coffey, 2003; Ferguson et al., 2006; Bewley et al., 2008). The main drawback of these systems was that they were not fully automated. Thermal sensing cameras have also been evaluated with promising results (Halachmi et al., 2008, 2013). Advancements in technology and the applicability of computerized 3-dimensional (**3D**) imaging systems has led to an increased development of systems with improved automatization in image processing (Weber et al., 2014; Fischer et al., 2015; Kuzuhara et al., 2015). The use of deep learning techniques is a breakthrough in automated livestock monitoring systems. These applications ensure a real-time and fully automated generation of BCS (Rodríguez Alvarez et al., 2018; Alvarez et al., 2019; Yukun et al., 2019). Although machine learning algorithms to date rely on 3D computer vision, it was recently reported that well-calibrated 2D algorithms could achieve comparable accuracies (O'Mahony et al., 2022).

CattleEye Ltd. (Belfast, United Kingdom) has developed a fully automated imaging system that captures 2D footage with an overhead view of cows as they walk through a race and generates real-time locomotion scores using convolutional neural networks (**CNN**); this system is now commercially available (Anagnostopoulos et al., 2023). Our aim here was to develop a machine learning algorithm capable of automatically generating BCS using the same inexpensive equipment. Additionally, we tested the performance of this system for single BCS estimation and changes in BCS over time, using the manual scores of an experienced human assessor as the ground truth data.

MATERIALS AND METHODS

Ethics Statement

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

System Setup

All farms that participated in this study were equipped with a 2D surveillance camera placed above the exit passageway of the milking parlor at a height of

4 m above the ground. Top-down footage of each cow was captured, stored in the cloud, and processed by CattleEye Ltd. (Belfast, UK).

Datasets

A first set of data was collected from July 2022 to January 2023 by 5 human assessors (**HA**) in 7 farms, and these data were used to train the algorithm. A second dataset was collected for testing our model from February 2023 to March 2023 by the same HA in 4 of the participating farms.

Training Dataset

Five HA, namely HA1 (author NS), HA2 (GO), HA3 (JN), HA4 (AA), and HA5 (BG), performed a total of 34 whole-milking-herd BCS sessions in 7 commercial dairy farms (designated as A–G) selected for convenience, located in England and Wales. Farms were milking approximately 1,000, 2,300, 800, 2,100, 760, 800, and 600 Holstein cows, respectively. Cows were milked 3 times per day in rotary parlors, apart from cows in farm E, which were milked in a rapid-exit herringbone parlor. The total number of cows scored per farm and by assessor at each visit are presented in Table 1. One additional visit was scheduled in farm A only to score and record cows with extreme BCS ($\leq$2.50 or $\geq$4.00).

All HA were qualified veterinarians with experience in dairy cattle research. A total of 34,150 manual scores (**MAN_BCS**) from human assessors with correct cow identification were recorded using a scale of 1 to 5 with 0.25-point increments developed by Ferguson et al. (1994). On all farms, MAN_BCS were collected in the milking parlor during the mid-day milking, except for farm E, where HA scored the cows standing at the feed bunk immediately after milking by walking behind the cows inside the barns.

Table 1. Dataset used for training of the algorithm, showing the number of cows per farm assigned a manual BCS by 5 human assessors (HA) during 35 visits in 8 dairy farms with a milking herd size ranging from approximately 600 to 2,300 Holstein cows

Farm	Scorer	Visit order	No. of cows
A	HA4	1	984
	HA4	2	1,035
	HA4	3	1,005
	HA1	4	885
	HA1	5	974
	HA1	6	945
	HA1	7	228
B	HA4	1	2,182
	HA4	2	1,913
	HA1	3	1,999
	HA1	4	1,965
C	HA4	1	749
	HA4	2	814
	HA5	3	704
	HA2	4	752
	HA1	5	761
	HA1	6	793
	HA1	7	739
	HA1	8	734
D	HA4	1	1,935
	HA1	2	1,942
	HA1	3	1,954
E	HA3	1	364
	HA4	2	652
	HA5	3	355
	HA1	4	542
	HA1	5	480
F	HA3	1	746
	HA4	2	800
	HA5	3	778
	HA1	4	764
	HA1	5	711
	HA1	6	629
	HA1	7	744
G	HA2	1	593

Algorithm Development

Our algorithm was developed using the training dataset comprising various farms, which allowed us to build robust image datasets for predicting a cow's BCS. To address the issue of imbalanced classes, we employed a stratified sampling technique to ensure a sufficient representation of all score categories in each dataset and reduce bias toward any dominant scores. The training dataset was split using a farm- and score-based stratified sample, with 80% of the unique IDs on each farm forming a training dataset and 20% forming the algorithm validation dataset. The training set, which constituted the most significant portion of the dataset, was used for training the deep learning model. The validation set was used for early stopping to improve training efficiency and prevent model degradation.

For each farm, we used a top-down camera positioned in a strategic area on-site to capture footage of the herd exiting the milking parlor providing a complete view of the animal. As part of the study, we gathered footage during one milking from the various farms when the HA were on-site to provide scores for each animal. These videos were then processed by a pre-trained cow detection and pose estimation pipeline to track and identify the key points on each animal as it passed under the camera. Once the pipeline had determined the initial detections, we removed any examples where the average key point confidence was below 90%, ensuring the consistency and quality of our image datasets. All final detections were then cropped and resized to a fixed resolution (256 $\times$ 256 pixels).

Due to the nature of this study, we opted to use an ordinal regression method when developing the algorithm. Because the MAN_BCS possess a natural order

(i.e., ranking), we can retain the ordinal nature of the scores by applying this method instead of a traditional regression or classification approach. We used a state-of-the-art deep learning architecture, EfficientNetV2 (Tan and Le, 2021), to form the algorithm's backbone. Once an image is passed through this algorithm, global average pooling is used to summarize the features produced by the EfficientNetV2 model. These summarized features are passed into one final dense layer to produce a prediction. We used the RMSprop algorithm to optimize the algorithm at a learning rate of 1e-4. During training, we performed rigorous data augmentation techniques such as RandAugment (Cubuk et al., 2020) and GridMask (unpublished data; P. Chen, S. Liu, H. Zhao [The Chinese University of Hong Kong, Ma Liu Shui, Hong Kong], X. Wang [City University of Hong Kong, Kowloon, Hong Kong], and J. Jia [The Chinese University of Hong Kong, Ma Liu Shui, Hong Kong]) to mitigate potential algorithm biases further. RandAugment applies a combination of image transformations to provide random variations of the original training images. GridMask simulates structured occlusions of the cow during training in the form of a grid-like mask. By applying both methods during training, we were able to increase the variety of the training images, which in turn reduces overfitting and improves the algorithm's ability to generalize. Ordinal cross-entropy loss was used for training the algorithm, which frames the ordinal regression problem as a set of binary classification subproblems (Cao et al., 2020). During training, the cross-entropy loss for each binary subproblem is determined and aggregated to form the overall loss.

Dataset Testing

For herd demographics and body condition of individual cows to vary, a significant length of time was allowed between the scoring sessions used for training and the scoring sessions used for testing. Specifically, the time interval allowed between the last training scoring session and the first session used for testing was 4 mo for farm G and 2 mo for the other farms, except for the training session only involving cows with extreme BCS at farm A, which took place 35 d before the first session used for testing.

We tested AI_BCS using the MAN_BCS of HA1 as the ground truth data. HA1 is a qualified veterinarian with 4 years of experience in body condition scoring on several research projects. The HA1 performed 2 whole-herd BCS sessions on each farm 30 d apart, which we considered an adequate period of time to allow changes in condition. We anticipated that cows in their early and late lactation stages would likely lose and gain a detectable amount of body condition, respectively.

The HA1 visually evaluated the BCS of each cow in the milking parlor during milking, standing about 2 m behind the rear of the cow, at the same height with the cows in farms A, B, and D, and at 2 m higher in farm G, using the 5-point scale method (Ferguson et al., 1994). Recording was performed using a portable tablet with touch screen (Toughbook FZ-G2, Panasonic) by manually entering the freeze brand number of each cow (located at the rear thigh area on either side of the tail) and the BCS into an Excel spreadsheet. Records from cows scored at both sessions within each farm were used to assess the agreement between MAN_BCS and AI_BCS in detecting monthly changes in BCS (**ΔBCS**). The HA1 did not have access to the CattleEye data and vice versa. At the end of the study, we merged the records of MAN_BCS and AI_BCS according to cow identification number so that statistical analyses could be performed.

To calculate the intra-observer reliability, HA1 visited farm D 24 h after a whole-herd BCS scoring session to score a subset of the same cows. Automated BCS recordings from the same days were also stored to assess the precision of the automated system at the same farm.

The interobserver agreement of HA1 with 2 human assessors, namely HA2 and HA3, was assessed by performing 2 more sessions in farm B. The HA2 is a veterinarian with many years of experience in collecting BCS data for research. The HA3 is also a veterinarian with 15 years of research experience working with dairy cows, but not with collecting BCS data in particular. Both HA2 and HA3 teach body condition scoring to undergraduate veterinary students.

The backfat thickness (**BFT**) of 111 randomly selected cows in farm B was measured using a portable real-time B-mode ultrasonographic equipment (Draminski 4Vet mini, Draminski S.A.) with a 5.0 MHz linear transducer at a field view depth of 50 mm. Cows were minimally restrained with headlocks at the feed bunk immediately after milking. After brushing the examination site, a 50% aqueous solution of isopropyl alcohol and then ultrasound gel were applied to obtain acoustic contact between the probe surface and the skin without clipping any haircoat. The probe was placed at the sacral area vertically to an imaginary line connecting the tuber ischia (pin bone) and tuber sacrale (hook bone) at the site corresponding to the front of the first coccygeal vertebra, as described by Schröder and Staufenbiel (2006). After applying the slightest possible pressure, the obtained images were stored for subsequent analysis using ImageJ, a freeware digital image processing program provided by the National Institutes of Health (Schneider et al., 2012). The BFT measurements always included the thickness of the skin. The

BCS of each cow was manually recorded by HA1 before measuring BFT to avoid any biased estimations.

Statistical Analysis

Data were analyzed with IBM SPSS version 28. Records with missing IDs were excluded and histograms were created to assess any typing errors in BCS. Records with a BCS not corresponding to the 1 to 5 scale in 0.25-unit increments were excluded.

The within-observer agreement for both the MAN_BCS and the AI_BCS, as well as the interobserver agreement between HA and between MAN_BCS and AI_BCS, per farm and overall, were measured by calculating the PA at 3 levels of error (0.00, within ± 0.25-unit, and within ± 0.50-unit difference of BCS), and the weighted Cohen's kappa (κ_w) coefficient, an estimate of categorical agreement exceeding agreement by chance, using quadratic weights for the exact agreement. The same metrics were also used to assess the agreement between MAN_BCS and AI_BCS regarding the ΔBCS for cows that were scored in both sessions.

The commonly used benchmark of accepted reliability for PA in most studies of 80% (McHugh, 2012) was also used in our study at the ± 0.25 error range. Interpretation of κ_w was performed according to the recommendations of Landis and Koch (1977), as follows: slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–1.00) agreement. A κ_w of ≥0.60 is suggested as a threshold of accepted interobserver agreement (Gibbons et al., 2012; Schlageter-Tello et al., 2014).

Moreover, a Bland-Altman plot (Bland and Altman, 1986) of the differences between MAN_BCS and AI_BCS against the mean scores of both methods was constructed using all single scores to examine for any systematic or proportional bias in the automatically generated BCS.

Finally, the relationship of the MAN_BCS and AI_BCS with the BFT measurements of 111 cows was characterized using Passing-Bablok regressions, a nonparametric analysis allowing for comparison of different methods with measurement errors which is not sensitive to data distribution (Passing and Bablok, 1983).

RESULTS

A total of 9,657 paired MAN_BCS and AI_BCS records were available for testing and 3,817 cows were scored twice 30 d (±2 d) apart. The overall and within farm distribution of BCS as well as their monthly changes, as recorded by HA1 and the system, are presented in Tables 2 and 3. The overall mean, standard deviation, and minimum–maximum were 3.23, 0.42, and 1.75 to 4.50 for single MAN_BCS and 3.32, 0.29, and 2.25 to 4.25 for AI_BCS. Additionally, the overall mean, standard deviation, and minimum–maximum ΔBCS obtained were −0.04, 0.24, and −1.25 to 1.25 for HA1, and 0.02, 0.15, and −1.00 to 0.75 for the system, respectively.

Table 2. Dataset used for testing, showing the distribution of single BCS estimated manually by a human assessor (MAN_BCS) and with an automated system (AI_BCS) during 8 whole-herd sessions in 4 dairy farms

						Percentile			
Method	Farm	Session	n[1]	Mean	SD	25th	50th	75th	Min–max[2]
MAN_BCS	A	1	768	3.25	0.42	3.00	3.25	3.50	2.00–4.50
AI_BCS				3.31	0.29	3.00	3.25	3.50	2.25–4.25
MAN_BCS		2	824	3.27	0.40	3.00	3.25	3.50	2.00–4.50
AI_BCS				3.33	0.29	3.25	3.25	3.50	2.25–4.25
MAN_BCS	B	1	1,779	3.20	0.42	3.00	3.25	3.50	2.00–4.25
AI_BCS				3.32	0.31	3.25	3.25	3.50	2.25–4.00
MAN_BCS		2	1,699	3.16	0.46	3.00	3.25	3.50	2.00–4.25
AI_BCS				3.33	0.29	3.25	3.25	3.50	2.50–4.00
MAN_BCS	D	1	1,774	3.29	0.40	3.00	3.25	3.50	1.75–4.50
AI_BCS				3.35	0.30	3.00	3.25	3.50	2.50–4.25
MAN_BCS		2	1,637	3.17	0.46	3.00	3.25	3.50	2.00–4.25
AI_BCS				3.34	0.31	3.00	3.25	3.50	2.50–4.25
MAN_BCS	G	1	569	3.35	0.35	3.00	3.25	3.50	2.25–4.50
AI_BCS				3.22	0.21	3.00	3.25	3.25	2.75–4.25
MAN_BCS		2	607	3.31	0.35	3.00	3.25	3.50	2.00–4.50
AI_BCS				3.26	0.23	3.00	3.25	3.50	2.50–4.25
MAN_BCS	Total		9,657	3.23	0.42	3.00	3.25	3.50	1.75–4.50
AI_BCS				3.32	0.29	3.00	3.25	3.50	2.25–4.25

[1]n = number of single BCS.

[2]Min–max = minimum–maximum.

Table 3. Distribution of monthly changes in body condition score (ΔBCS) estimated manually by a human assessor (MAN_BCS) and by an automated system (AI_BCS) during 8 whole-herd sessions in 4 dairy farms

					Percentile			
Method	Farm	n[1]	Mean	SD	25th	50th	75th	Min–max[2]
MAN_BCS	A	592	0.07	0.21	0.00	0.00	0.25	−1.25–1.25
AI_BCS			0.06	0.14	0.00	0.00	0.25	−0.25–0.75
MAN_BCS	B	1,368	−0.03	0.24	0.00	0.00	0.00	−1.00–0.75
AI_BCS			0.02	0.16	0.00	0.00	0.00	−0.50–0.50
MAN_BCS	D	1,349	−0.12	0.24	−0.25	0.00	0.00	−1.25–0.75
AI_BCS			−0.01	0.15	0.00	0.00	0.00	−1.00–0.50
MAN_BCS	G	508	−0.25	0.20	−0.25	0.00	0.00	−0.75–1.00
AI_BCS			0.04	0.14	0.00	0.00	0.00	−0.25–0.50
MAN_BCS	Total	3,817	−0.04	0.24	−0.25	0.00	0.00	−1.25–1.25
AI_BCS			0.02	0.15	0.00	0.00	0.00	−1.00–0.75

[1]n = number of cows scored twice 30 d apart.
[2]Min–max = minimum–maximum.

The within-observer agreement of MAN_BCS produced a PA at the 0.00, ± 0.25, and ± 0.50 error range of 68.1, 96.1, and 99.8%, respectively, and a κ_w of 0.94 (95% CI: 0.92–0.95). The within-observer agreement of AI_BCS produced a PA at the 0.00, ± 0.25, and ± 0.50 error range of 96.2%, 100%, and 100%, respectively and a κ_w of 0.99 (95% CI: 0.99–0.99) (Table 4).

The correct classification rates of single AI_BCS within the MAN_BCS classes (≤2.50, 2.75–3.25, 3.50–3.75, and ≥4.00) were 9.5% (with 24.3% and 51.6% being scored as 2.75 and 3.00, respectively), 82.6% (with 15.2% being scored as 3.50), 77.2% (with 19.6% being scored as 3.25) and 22.6% (with 58.3% being scored as 3.75), respectively.

The overall PA at the 0.00, within 0.25-unit, and within 0.50-unit of BCS error range between MAN_BCS and AI_BCS was 44.4%, 84.6%, and 94.8%, respectively. The PA at the ± 0.25 error range were between 80.1% and 89.4% across farms, always above the benchmark of accepted reliability. The overall κ_w was 0.69 (95% CI: 0.68–0.70), ranging from 0.67 (95% CI: 0.65 – 0.69) to 0.75 (95% CI: 0.73–0.77) across farms, representing substantial agreement (Table 5).

Regarding ΔBCS (Table 6), the overall PA at the 0.00, ± 0.25, and ± 0.50 error range between the 2 methods was 45.7%, 88.2%, and 97.2%, respectively. The PA at the ± 0.25 level ranged from 88.0% to 91.6% across farms and always above the benchmark of accepted reliability. The overall κ_w was 0.20 (95% CI: 0.17–0.23), ranging from 0.09 (95% CI: 0.01–0.18) to 0.20 (95% CI: 0.15–0.25) across farms, representing only slight agreement.

The Bland-Altman plot (Figure 1) showed a minimal systemic bias of the system, assigning on average a higher BCS than HA1 by 0.09. The lower and upper 95% limits of agreement between the 2 methods were −0.63 and 0.45, respectively. A simple linear regression of the differences between the 2 methods against their mean values produced an R^2 of 0.253 ($P < 0.001$), revealing a proportional bias, as well. Differences between HA1 and the system increased as BCS increased.

Table 4. Intraobserver agreement of manual BCS estimations (MAN_BCS) and automatically generated BCS (AI_BCS) assessed with percentage agreement (PA) at 3 levels of error (0.00, within 0.25 unit, and within 0.50 unit) and quadratically weighted kappa coefficients (κ_w) for the exact agreement, performed 24 h apart in the same farm

Method	n[1]	Error	PA (%)	κ_w	SE[2]	*P*-value	95% CI[3]
MAN_BCS	486	0.00	68.3	0.94	0.070	<0.001	0.92–0.95
		0.25	96.1				
		0.50	99.8				
AI_BCS	2,062	0.00	96.2	0.99	0.001	<0.001	0.99–0.99
		0.25	100				
		0.50	100				

[1]n = number of cows scored twice 24 h apart.
[2]SE = standard error of κ_w.
[3]CI = confidence interval of κ_w.

Table 5. Categorical agreement assessed with percentage agreement (PA) at 3 levels of error (0.00, within 0.25 unit, and within 0.50 unit) and quadratically weighted kappa coefficients (κ_w) for the exact agreement, between manual estimations of BCS by a human assessor and automatically generated BCS from 8 whole-herd sessions performed in 4 dairy farms

Farm	n[1]	Error	PA (%)	κ_w	SE[2]	*P*-value	95% CI[3]
A	1,592	0.00	53.1	0.75	0.010	<0.001	0.73–0.77
		0.25	89.4				
		0.50	96.2				
B	3,478	0.00	41.8	0.67	0.008	<0.001	0.65–0.69
		0.25	80.1				
		0.50	93.1				
D	3,411	0.00	44.1	0.71	0.007	<0.001	0.70–0.73
		0.25	85.3				
		0.50	94.6				
G	1,176	0.00	41.2	0.64	0.014	<0.001	0.61–0.67
		0.25	86.9				
		0.50	98.4				
Total	9,657	0.00	44.4	0.69	0.005	<0.001	0.68–0.70
		0.25	84.6				
		0.50	94.8				

[1]n = number of single BCS.
[2]SE = standard error of κ_w.
[3]CI = confidence interval of κ_w.

The interobserver agreement between human assessors is presented in Table 7. The PA between HA1 vs. HA2 and HA1 vs. HA3 at the 0.00, ± 0.25-unit, and ± 0.50-unit of BCS error range was 53.4%, 95.8%, and 99.4%, and 29.7%, 75.3%, and 91.7%, respectively. The κ_w was 0.82 (95% CI: 0.80–00.84) and 0.77 (95% CI: 0.73–0.80), representing almost perfect and substantial agreement, respectively.

The Passing-Bablok regression (Figure 2) revealed a strong linear relationship of both MAN_BCS and AI_BCS with the BFT measurements, producing Spearman's rank correlation coefficients of $\rho = 0.91$ and 0.75 ($P < 0.001$), respectively.

DISCUSSION

The aim of this study was to develop and evaluate the performance of a low-cost, fully automated 2D surveillance system for BCS monitoring of Holstein cattle on commercial dairy farms in the UK. First, we assessed the precision of the system, which was almost perfect in terms of assigning exactly the same BCS to individual

Table 6. Categorical agreement assessed with percentage agreement (PA) at 3 levels of error (0.00, within 0.25 unit, and within 0.50 unit) and quadratically weighted kappa coefficients (κ_w) for the exact agreement, between manual estimations of BCS by a human assessor and automatically generated BCS in detecting monthly changes in BCS of 3,817 cows in 4 dairy farms

Farm	n[1]	Error	PA (%)	κ_w	SE[2]	*P*-value	95% CI[3]
A	592	0.00	50.2	0.17	0.038	<0.001	0.10–0.25
		0.25	91.6				
		0.50	97.8				
B	1,368	0.00	43.0	0.16	0.025	<0.001	0.11–0.21
		0.25	88.0				
		0.50	97.2				
D	1,349	0.00	46.5	0.20	0.024	<0.001	0.15–0.25
		0.25	86.2				
		0.50	96.2				
G	508	0.00	45.5	0.09	0.043	0.017	0.01–0.18
		0.25	90.1				
		0.50	98.8				
Total	3,817	0.00	45.7	0.20	0.015	<0.001	0.17–0.23
		0.25	88.2				
		0.50	97.2				

[1]n = number of cows scored twice 30 d apart.
[2]SE = standard error of κ_w.
[3]CI = confidence interval of κ_w.

Table 7. Categorical interobserver agreement assessed with percentage agreement (PA) at 3 levels of error (0.00, within 0.25 unit, and within 0.50 unit) and quadratically weighted kappa coefficients (κ_w) for the exact agreement, between single BCS estimations by human assessors (HA) performed in different sessions in the same farm

Observer	n[1]	Error	PA (%)	κ_w	SE[2]	*P*-value	95% CI[3]
HA1 vs. HA2	1,577	0.00	53.4	0.82	0.010	<0.001	0.80–0.84
		0.25	95.8				
		0.50	99.4				
HA1 vs. HA3	573	0.00	29.7	0.77	0.018	<0.001	0.73–0.80
		0.25	75.3				
		0.50	91.7				

[1]n = number of cows.
[2]SE = standard error of κ_w.
[3]CI = confidence interval of κ_w.

cows within successive days. Although the categorical agreement with the human assessor for the exact score was substantial, with almost 85% of the scores being within the 0.25-unit error range of single manual BCS, the system was less accurate in identifying cows with very low (≤2.50) or very high (>4.00) BCS. A poor κ_w was produced regarding the monthly changes in BCS, although 88% of them were within the 0.25 error range of the manual ΔBCS. We also observed a strong linear association of the ultrasound measurements of backfat thickness with the automatically generated BCS, weaker though than that with the manual BCS.

In this study, we calculated the PA at 3 different levels of error and the quadratic κ_w at the exact score as appropriate measures of within- and interobserver

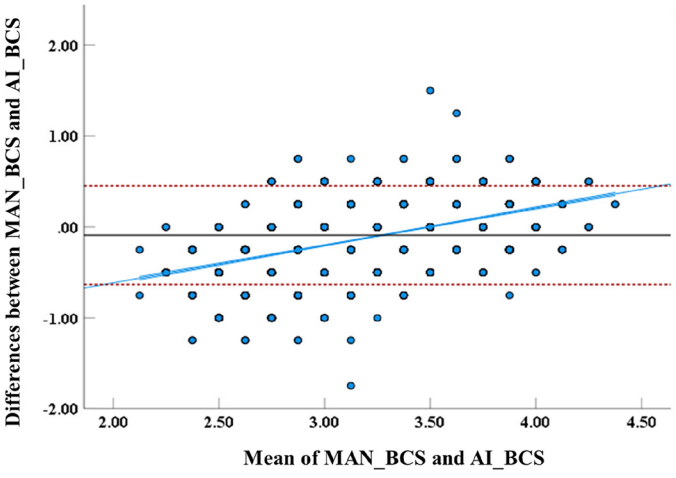

Figure 1. Bland-Altman plot of the differences between the manual BCS (MAN_BCS) and the automatically generated BCS (AI_BCS) against the mean values of both methods. Solid black horizontal line represents the mean of differences, showing a systematic error of −0.09. Dashed horizontal lines represent the 95% limits of agreement (±1.96 × SD) at −0.63 and 0.45. The regression line represents a proportional bias ($R^2 = 0.253$, $P < 0.001$).

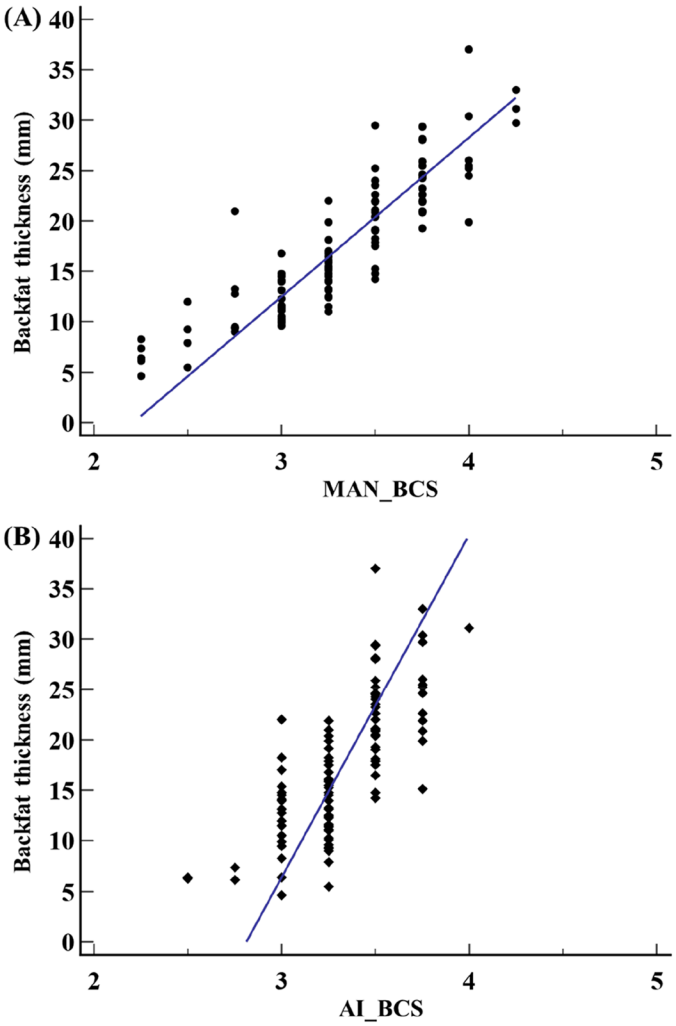

Figure 2. Passing-Bablok regressions of ultrasound backfat thickness measurements in 111 Holstein cows against (A) manual BCS (MAN_BCS), with a Spearman's rank correlation coefficient ρ = 0.905 (95% CI: 0.86–0.93, $P < 0.001$); and (B) automatically generated BCS (AI_BCS), with a Spearman's rank correlation coefficient ρ = 0.751 (95% CI: 0.66–0.82, $P < 0.001$).

categorical agreement. Weights for kappa coefficient are preferred over unweighted Cohen's kappa when disagreement is more "serious," as codes are farther apart, which is the case for ordinal variables with multiple codes. The quadratic weights are commonly used, and under specific conditions, they are equivalent to the intraclass correlation coefficient (Schuster, 2004).

The Pearson correlation coefficient (r), which is often reported in relevant studies, measures a linear relationship between continuous variables and is not a suitable metric for the agreement of measures in ordinal scale (Bland and Altman, 1986; Lin, 1989). Nevertheless, and only for comparison purposes with previous studies, we calculated the Pearson correlation coefficient for the overall single BCS (data not shown) and found it to be r = 0.76 ($P < 0.001$).

The day-to-day repeatability of the system in assigning the same BCS to individual cows was almost perfect and exceeded that of HA1, which was already high relative to previous studies with human assessors. Reportedly, κ_w coefficients for the within-observer agreement at the exact score ranged from 0.22 to 0.75 among bovine veterinarians, from 0.86 to 0.98 between instructors (Kristensen et al., 2006), from 0.62 to 0.80 between experienced dairy scientists (Vasseur et al., 2013), and from 0.52 to 0.72 between trained assessors (Song et al., 2019).

The system showed a substantial overall agreement with MAN_BCS for the exact score, with 85% of all scores being within the 0.25-unit error range, and with a negligible systematic bias. The performance varied across farms, but κ_w and PA were always within the substantial agreement range and above the benchmark of accepted reliability at the 0.25 error level, respectively. The system was significantly less accurate in correctly classifying cows assigned a MAN_BCS $\leq$2.50 and >4.00. This was also evident from the proportional bias observed in the Bland-Altman plot. These extreme scores were, as expected, the least represented in our dataset. The decreased accuracy at very low or very high BCS is a common challenge for automated systems. However, one of the main advantages of deep learning applications compared with conventional approaches is that they typically give the system the capacity to learn from experience and improve without necessarily being programmed to do so (Sarker, 2021). Therefore, we consider that increasing the algorithm's training with more data focusing on extreme BCS would progressively improve the accuracy of the system.

The interobserver agreement between HA1 and HA2 was within the almost perfect range and better than that between AI_BCS and MAN_BCS. Both assessors were very experienced in collecting BCS data for research. This level of agreement is greater than what is typically observed in practice. The agreement between HA1 and HA3 was within the substantial range regarding the κ_w and close to that between HA1 and HA2. However, the PA at the exact score and the $\pm$ 0.25 error level was lower than that observed between AI_BCS and MAN_BCS. The interobserver agreement at the exact score in the literature varied, and was reportedly κ_w = 0.17 to 0.78 among bovine veterinarians (Kristensen et al., 2006), κ_w = 0.67 between trainers and mean κ_w = 0.42 between trainer and trainees (Vasseur et al., 2013), and κ_w = 0.76 to 0.89 between a veterinarian and 2 animal health technicians following review and training on BCS methodology before scoring (Morin et al., 2017). Hence, the agreement obtained between AI_BCS and MAN_BCS here was similar to that achieved between trained and experienced human observers in previous studies.

The interobserver agreement between human assessors in detecting changes in BCS has been investigated only by Morin et al. (2017), with 3 observers scoring 57 cows in the first 3 wk after calving and again in 6 to 8 wk after calving. The range of ΔBCS recorded was -0.75 to 1.00 points. The produced agreement among 3 observers was moderate, with a mean quadratic κ_w of 0.49, which was higher than the agreement observed in our study. The low coefficient in our study could be partly explained by the limited capacity of the system to correctly classify cows at BCS $\leq$2.50. However, the PA at the 0.00 and $\pm$ 0.25 error levels in our study were both higher than the 33.3% and 83.6%, respectively, reported again by Morin et al. (2017). Percentage agreement is a direct measure, whereas κ_w is an estimate of agreement exceeding chance that is influenced by the observer's accuracy, the number of codes, the prevalence of each code, and the observer's bias (McHugh, 2012). This could be the case accounting for the observed discrepancy that although the PA in our study between manual and automated ΔBCS was higher than those in Morin et al. (2017), the κ_w was significantly lower. The codes for ΔBCS that HA1 assigned were higher than those of the system and of those in Morin et al. (2017). The prevalence within each code was also different from Morin et al. (2017), as a larger number of cows were scored at various stages of lactation. We are not aware of other studies evaluating the accuracy of an automated system in detecting ΔBCS. Based on the obtained PA in the current study, we can reasonably support that the system was adequately accurate in detecting ΔBCS and within the agreement levels expected between human observers.

The BFT measurement using ultrasonography is considered a more objective estimation of a cow's subcutaneous fat reserves than manual BCS estimations (Schröder and Staufenbiel, 2006) presenting excellent

precision and high correlation with actual carcass backfat thickness (Brethour, 1992). The correlation observed between MAN_BCS and BFT measurements was strong and within the range of coefficients (0.82–0.98) reported in previous studies (Hussein et al., 2013; Strieder-Barboza et al., 2015; Siachos et al., 2021). Weber et al. (2014) developed a 3D optical system which managed to estimate weekly BFT measurements with a notably high correlation coefficient of 0.96. In our study, the AI_BCS produced a strong linear relationship with BFT, though weaker than that of the MAN_BCS, but failed to correctly classify most thin cows with BFT <10 mm, which received a MAN_BCS of 2.75 or less.

Automated systems are capable of identifying individual cows and recording and storing large datasets frequently and effortlessly. On the other hand, a whole-herd BCS session in large dairy herds is challenging both physically and mentally for a human scorer. It is worth mentioning that HA1 was never able to score the entire milking herd in any session and recorded 4.7% to 19.9% fewer cows across sessions compared with the system. The main reasons for this were unclear IDs, fast movement of the rotaries, typing errors when scoring, and mental fatigue after many hours of repeatedly scoring. The milking time in the farms we visited ranged from 3.5 to 6.5 h. An automated system is by definition not prone to such errors.

Over the last 2 decades, a growing body of research has reported the development of systems with image processing for BCS estimation in dairy cows using 2D and 3D digital or thermal cameras, at different levels of automation (fully or semi-automated). These systems focus on a wide range of anatomical features, use different types of models to estimate BCS, and have been validated with variable numbers of cows, showing variable performance in terms of intra- and interobserver reliability. Unfortunately, little uniformity is present in the reported metrics of agreement, accuracy, and precision among studies.

One of the first systematic attempts to automatically generate BCS using 2-dimensional imaging was described by Coffey (2003), which involved projecting laser light and extracting data manually from fitted curves of the cow's contour around the tailhead using a digital camera. An $r = 0.62$ with the average manual BCS of 3 human assessors was achieved, in cows within a narrow, though, range of BCS (2.25–3.25).

To date, all developed and evaluated systems using a 2D camera had either low or medium levels of automation, meaning that the anatomical features on the acquired images had to be labeled or the input of the images had to be selected manually. In terms of performance, Ferguson et al. (2006) used a mixed regression model with an R^2 of 0.68 to 0.80, Bewley et al. (2008) reported an accuracy of 92.8% at the ± 0.25 error range. Also using a mixed regression model, Azzaro et al. (2011) reported an error rate of 0.31 in predicting the actual BCS using polynomial kernel principal component analysis, and Bercovich et al. (2013) reported an R^2 of 0.77 of a partial least square regression model and Fourier descriptors of cows' signatures with an accuracy of 58% at the ± 0.25 error interval. Although the system developed by Bewley et al. (2008) achieved a markedly high level of accuracy, the level of automation was low. Digital images were captured automatically, but the image processing was carried out manually. Finally, Nagy et al. (2023) evaluated a 2D camera using deep learning CNN and achieved an accuracy of approximately 60% at the 0.25 error level and an unweighted kappa of approximately 0.30 for the exact agreement at 12 BCS classes. When they re-evaluated the agreement at 3 BCS classes designated as being above, within, or below the recommended BCS range per stage of lactation, the unweighted kappa increased at levels between 0.60 and 0.80, indicative of substantial agreement.

Halachmi et al. (2013) used thermal imaging and reported $r = 0.94$ between the thermal sensed and the manual BCS by measuring the deviation of the cow's contour against a fitted parabola. Both accuracy and level of automation were significantly ameliorated compared with their previous work, where the correlation between the predicted BCS by the thermal camera and the manual BCS produced a Spearman's rank correlation coefficient of 0.315 (Halachmi et al., 2008).

The majority of recently published studies assessed the feasibility of 3D imaging for automatically generated BCS. Level of automation and performance were highly variable. Fischer et al. (2015) used principal components analysis and achieved a high correlation ($r = 0.89$–0.96) and a mean absolute error of 0.28. However, it must be noted that the datasets selected for calibration and validation were small and selected to cover a wide range of BCS, which may not comply with the usual BCS distribution seen in dairy herds. Similarly, Kuzuhara et al. (2015) achieved good correlations between predicted and manual BCS with an R^2 of 0.74 and a root mean square error of 0.18, using principal components analysis with the inclusion of geodesic lines connecting specific anatomical features. Both image capturing and measurement of the length of the geodesic lines were, however, performed manually. Spoliansky et al. (2016) used automated image processing and reported $r = 0.72$ between predicted and manual BCS and 74% accuracy at the ± 0.25 error range, using polynomial regression models with an R^2 of 0.75 and a mean absolute error of 0.26. Song et al. (2019) obtained an overall sensitivity of 0.72 in correctly scor-

ing cows at the ± 0.50 error range and a mean absolute error of 0.15-unit of BCS using a polynomial regression model. However, the whole system's setup had to be moved manually behind the cow and image capturing and image processing were performed manually, as well. Martins et al. (2020) developed a regression model with an R^2 of 0.61 to 0.63 and root mean square error of 0.16 to 0.17, by manually capturing and processing a lateral and a dorsal 3D image from each cow. Finally, Liu et al. (2020) achieved an accuracy of 76% at the ± 0.25 error range by developing a fully automated system using an ensemble learning model on 3D images.

The following studies dealt with the validation of systems, relying again on 3D images and using machine learning algorithms to generate BCS. All these systems have the advantage of being fully automated, allowing for real-time acquisition of BCS estimations. Rodríguez Alvarez et al. (2018) reported an accuracy of 78% at the ± 0.25 error range using CNN, which was improved up to 81.5% after using transfer learning and techniques of ensemble modeling (Alvarez et al., 2019). Yukun et al. (2019) assessed a DenseNet CNN model that acquired data from manual BCS and BFT measurements and achieved accuracies of 45% and 77% at the 0.00 and ± 0.25 error range, respectively. O'Mahony et al. (2022) reported poor agreement with manual scores, with accuracies at the ± 0.25 error interval ranging from 22% to 39% and Krippendorf's alpha coefficients ranging from −0.08 to 0.03 for the exact agreement with manual scores, again with the use of deep learning CNN. A 2D neural network yielded even lower accuracies. Zhao et al. (2023) reported an accuracy of 45% and 91.2% at the exact score and at the ± 0.25 error range in depth images using a model based on an EfficientNet network. However, extreme scores of ≤2.25 and ≥4.0 were classified by the authors as being 2.25 and 3.75, respectively. Furthermore, Shi et al. (2023) developed a point cloud 3D feature extraction network with attention guiding, which was accurate by 49 and 80% at the exact score and the ± 0.25 error range, a performance very similar to ours.

To the best of our knowledge, 3 published studies have evaluated the performance of commercially available fully automated BCS recording systems for dairy cows (Hansen et al., 2018; Mullins et al., 2019; O'Leary et al., 2020). All 3 systems were equipped with 3D cameras. Hansen et al. (2018) validated a fully automated system based on a 3D rolling ball algorithm allowing for simultaneous BCS, lameness, and BW monitoring, with accuracies of 66.4% at the ± 0.25 and 80% at the ± 0.34 error range and a mean error of 0.21. Validation was performed using 119 cows. Moreover, Mullins et al. (2019) reported r = 0.76 to 0.78 and displayed a Bland-Altman plot showing a systematic bias of 0.12. Visual assessment of the plot reveals a clear proportional bias. The authors stated that the system was accurate within the 3.00 to 3.75 range of BCS. A refinement on daily BCS data gathering by fitting a locally estimated scatterplot smoothing function was proposed by Albornoz et al. (2022) to ameliorate the system's precision. Finally, O'Leary et al. (2020) reported r = 0.72 and a Lin's concordance correlation coefficient of 0.67, respectively. A systematic bias of 0.11 was also produced in a Bland-Altman plot, but they did not assess the presence of a proportional error. Neither of the last 2 studies included any metric of categorical agreement to compare with our findings.

CONCLUSIONS

We demonstrated that a fully automated system using a machine learning algorithm to generate real-time BCS from 2D footage is able to predict single BCS and changes in BCS with an adequate accuracy, comparable to that obtained between trained human scorers. Lower accuracy at the low scores was observed but can be improved with more training data at the extremes.

ACKNOWLEDGMENTS

This study was funded by Innovate UK (Swindon, United Kingdom; Farming Innovation Programme Small R&D Partnership Projects, project no. 10027372). The second author (ML) works for CattleEye Ltd. (Belfast, United Kingdom), which aims to commercialize the system described here. He was responsible for algorithm development. He was involved in discussions regarding study design but was not involved in data collection and analysis of the testing dataset. The authors have not stated any other conflicts of interest.

REFERENCES

Albornoz, R. I., K. Giri, M. C. Hannah, and W. J. Wales. 2022. An improved approach to automated measurement of body condition score in dairy cows using a three-dimensional camera system. Animals (Basel) 12:72. https://doi.org/10.3390/ani12010072.

Alvarez, J. R., M. Arroqui, P. Mangudo, J. Toloza, D. Jatip, J. M. Rodriguez, A. Teyseyre, C. Sanz, A. Zunino, C. Machado, and C. Mateos. 2019. Estimating body condition score in dairy cows from depth images using convolutional neural networks, transfer learning and model ensembling techniques. Agronomy (Basel) 9:20.

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. 10:1111057. https://doi.org/10.3389/fvets.2023.1111057.

Azzaro, G., M. Caccamo, J. D. Ferguson, S. Battiato, G. M. Farinella, G. C. Guarnera, G. Puglisi, R. Petriglieri, and G. Licitra. 2011. Objective estimation of body condition score by modeling cow body shape from digital images. J. Dairy Sci. 94:2126–2137. https://doi.org/10.3168/jds.2010-3467.

Barletta, R. V., M. Maturana Filho, P. D. Carvalho, T. A. Del Valle, A. S. Netto, F. P. Rennó, R. D. Mingoti, J. R. Gandra, G. B. Mourão, P. M. Fricke, R. Sartori, E. H. Madureira, and M. C. Wiltbank. 2017. Association of changes among body condition score during the transition period with NEFA and BHBA concentrations, milk production, fertility, and health of Holstein cows. Theriogenology 104:30–36. https://doi.org/10.1016/j.theriogenology.2017.07.030.

Bauman, D. E., and W. B. Currie. 1980. Partitioning of nutrients during pregnancy and lactation: A review of mechanisms involving homeostasis and homeorhesis. J. Dairy Sci. 63:1514–1529. https://doi.org/10.3168/jds.S0022-0302(80)83111-0.

Bercovich, A., Y. Edan, V. Alchanatis, U. Moallem, Y. Parmet, H. Honig, E. Maltz, A. Antler, and I. Halachmi. 2013. Development of an automatic cow body condition scoring using body shape signature and Fourier descriptors. J. Dairy Sci. 96:8047–8059. https://doi.org/10.3168/jds.2013-6568.

Bewley, J. M., A. M. Peacock, O. Lewis, R. E. Boyce, D. J. Roberts, M. P. Coffey, S. J. Kenyon, and M. M. Schutz. 2008. Potential for estimation of body condition scores in dairy cattle from digital images. J. Dairy Sci. 91:3439–3453. https://doi.org/10.3168/jds.2007-0836.

Bland, J. M., and D. G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310. https://doi.org/10.1016/S0140-6736(86)90837-8.

Brethour, J. R. 1992. The repeatability and accuracy of ultrasound in measuring backfat of cattle. J. Anim. Sci. 70:1039–1044. https://doi.org/10.2527/1992.7041039x.

Butler, W. R. 2005. Nutrition, negative energy balance and fertility in the postpartum dairy cow. Cattle Pract. 13:13–18.

Cao, W., V. Mirjalili, and S. Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit. Lett. 140:325–331. https://doi.org/10.1016/j.patrec.2020.11.008.

Caraviello, D. Z., K. Weigel, P. Fricke, M. Wiltbank, M. Florent, N. Cook, K. Nordlund, N. Zwald, and C. Rawson. 2006. Survey of management practices on reproductive performance of dairy cattle on large US commercial farms. J. Dairy Sci. 89:4723–4735. https://doi.org/10.3168/jds.S0022-0302(06)72522-X.

Coffey, M. P. 2003. A phenotypic and genetic analysis of energy balance in dairy cows. PhD Thesis. School of Biological Sciences, University of Edinburgh, Edinburgh, UK.

Cubuk, E. D., B. Zoph, J. Shlens, and Q. V. Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. Pages 702–203 in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops. The Computer Vision Foundation.

Drackley, J. K. 1999. Biology of dairy cows during the transition period: The final frontier? J. Dairy Sci. 82:2259–2273. https://doi.org/10.3168/jds.S0022-0302(99)75474-3.

Edmonson, A. J., I. J. Lean, L. D. Weaver, T. Farver, and G. Webster. 1989. A body condition scoring chart for Holstein dairy cows. J. Dairy Sci. 72:68–78. https://doi.org/10.3168/jds.S0022-0302(89)79081-0.

Esposito, G., P. C. Irons, E. C. Webb, and A. Chapwanya. 2014. Interactions between negative energy balance, metabolic diseases, uterine health and immune response in transition dairy cows. Anim. Reprod. Sci. 144:60–71. https://doi.org/10.1016/j.anireprosci.2013.11.007.

Ferguson, J. D., G. Azzaro, and G. Licitra. 2006. Body condition assessment using digital images. J. Dairy Sci. 89:3833–3841. https://doi.org/10.3168/jds.S0022-0302(06)72425-0.

Ferguson, J. D., D. T. Galligan, and N. Thomsen. 1994. Principal descriptors of body condition score in Holstein cows. J. Dairy Sci. 77:2695–2703. https://doi.org/10.3168/jds.S0022-0302(94)77212-X.

Fischer, A., T. Luginbühl, L. Delattre, J. M. Delouard, and P. Faverdin. 2015. Rear shape in 3 dimensions summarized by principal component analysis is a good predictor of body condition score in Holstein dairy cows. J. Dairy Sci. 98:4465–4476. https://doi.org/10.3168/jds.2014-8969.

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.

Grummer, R. R., D. G. Mashek, and A. Hayirli. 2004. Dry matter intake and energy balance in the transition period. Vet. Clin. North Am. Food Anim. Pract. 20:447–470. https://doi.org/10.1016/j.cvfa.2004.06.013.

Hady, P. J., J. J. Domecq, and J. B. Kaneene. 1994. Frequency and precision of body condition scoring in dairy cattle. J. Dairy Sci. 77:1543–1547. https://doi.org/10.3168/jds.S0022-0302(94)77095-8.

Halachmi, I., M. Klopčič, P. Polak, D. J. Roberts, and J. M. Bewley. 2013. Automatic assessment of dairy cattle body condition score using thermal imaging. Comput. Electron. Agric. 99:35–40. https://doi.org/10.1016/j.compag.2013.08.012.

Halachmi, I., P. Polak, D. J. Roberts, and M. Klopcic. 2008. Cow body shape and automation of condition scoring. J. Dairy Sci. 91:4444–4451. https://doi.org/10.3168/jds.2007-0785.

Hansen, M. F., M. L. Smith, L. N. Smith, K. Abdul Jabbar, and D. Forbes. 2018. Automated monitoring of dairy cow body condition, mobility and weight using a single 3D video capture device. Comput. Ind. 98:14–22. https://doi.org/10.1016/j.compind.2018.02.011.

Herdt, T. H. 2000. Ruminant adaptation to negative energy balance: Influences on the etiology of ketosis and fatty liver. Vet. Clin. North Am. Food Anim. Pract. 16:215–230. https://doi.org/10.1016/S0749-0720(15)30102-X.

Hussein, H. A., A. Westphal, and R. Staufenbiel. 2013. Relationship between body condition score and ultrasound measurement of backfat thickness in multiparous Holstein dairy cows at different production phases. Aust. Vet. J. 91:185–189. https://doi.org/10.1111/avj.12033.

Komaragiri, M. V., and R. A. Erdman. 1997. Factors affecting body tissue mobilization in early lactation dairy cows. 1. Effect of dietary protein on mobilization of body fat and protein. J. Dairy Sci. 80:929–937. https://doi.org/10.3168/jds.S0022-0302(97)76016-8.

Kristensen, E., L. Dueholm, D. Vink, J. E. Andersen, E. B. Jakobsen, S. Illum-Nielsen, F. A. Petersen, and C. Enevoldsen. 2006. Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. J. Dairy Sci. 89:3721–3728. https://doi.org/10.3168/jds.S0022-0302(06)72413-4.

Kuzuhara, Y., K. Kawamura, R. Yoshitoshi, T. Tamaki, S. Sugai, M. Ikegami, Y. Kurokawa, T. Obitsu, M. Okita, T. Sugino, and T. Yasuda. 2015. A preliminarily study for predicting body weight and milk properties in lactating Holstein cows using a three-dimensional camera system. Comput. Electron. Agric. 111:186–193. https://doi.org/10.1016/j.compag.2014.12.020.

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.

Lin, L. I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255. https://doi.org/10.2307/2532051.

Liu, D., D. He, and T. Norton. 2020. Automatic estimation of dairy cattle body condition score from depth image using ensemble model. Biosyst. Eng. 194:16–27. https://doi.org/10.1016/j.biosystemseng.2020.03.011.

Martins, B. M., A. L. C. Mendes, L. F. Silva, T. R. Moreira, J. H. C. Costa, P. P. Rotta, M. L. Chizzotti, and M. I. Marcondes. 2020. Estimating body weight, body condition score, and type traits in dairy cows using three dimensional cameras and manual body measurements. Livest. Sci. 236:104054. https://doi.org/10.1016/j.livsci.2020.104054.

McHugh, M. L. 2012. Interrater reliability: The kappa statistic. Biochem. Med. (Zagreb) 22:276–282. https://doi.org/10.11613/BM.2012.031.

Morin, P. A., Y. Chorfi, J. Dubuc, J. P. Roy, D. Santschi, and S. Dufour. 2017. Short communication: An observational study investigating inter-observer agreement for variation over time of body condition score in dairy cows. J. Dairy Sci. 100:3086–3090. https://doi.org/10.3168/jds.2016-11872.

Mullins, I. L., C. M. Truman, M. R. Campler, J. M. Bewley, and J. H. C. Costa. 2019. Validation of a commercial automated body condi-

tion scoring system on a commercial dairy farm. Animals (Basel) 9:287. https://doi.org/10.3390/ani9060287.
Nagy, S. Á., O. Kilim, I. Csabai, G. Gábor, and N. Solymosi. 2023. Impact evaluation of score classes and annotation regions in deep learning-based dairy cow body condition prediction. Animals (Basel) 13:194. https://doi.org/10.3390/ani13020194.
O'Leary, N., L. Leso, F. Buckley, J. Kenneally, D. McSweeney, and L. Shalloo. 2020. Validation of an automated body condition scoring system using 3D imaging. Agriculture 10:246. https://doi.org/10.3390/agriculture10060246.
O'Mahony, N., L. Krpalkova, G. Sayers, L. Krump, J. Walsh, and D. Riordan. 2022. Two- and three-dimensional computer vision techniques for more reliable body condition scoring. Dairy 4:1–25. https://doi.org/10.3390/dairy4010001.
Otto, K. L., J. D. Ferguson, D. G. Fox, and C. J. Sniffen. 1991. Relationship between body condition score and composition of ninth to eleventh rib tissue in Holstein dairy cows. J. Dairy Sci. 74:852–859. https://doi.org/10.3168/jds.S0022-0302(91)78234-9.
Passing, H., and W. Bablok. 1983. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. Clin. Chem. Lab. Med. 21:709–720. https://doi.org/10.1515/cclm.1983.21.11.709.
Randall, L. V., M. J. Green, M. G. G. Chagunda, C. Mason, S. C. Archer, L. E. Green, and J. N. Huxley. 2015. Low body condition predisposes cattle to lameness: An 8-year study of one dairy herd. J. Dairy Sci. 98:3766–3777. https://doi.org/10.3168/jds.2014-8863.
Roche, J. R., P. G. Dillon, C. R. Stockdale, L. H. Baumgard, and M. J. VanBaale. 2004. Relationships among international body condition scoring systems. J. Dairy Sci. 87:3076–3079. https://doi.org/10.3168/jds.S0022-0302(04)73441-4.
Roche, J. R., N. C. Friggens, J. K. Kay, M. W. Fisher, K. J. Stafford, and D. P. Berry. 2009. Body condition score and its association with dairy cow productivity, health, and welfare. J. Dairy Sci. 92:5769–5801. https://doi.org/10.3168/jds.2009-2431.
Roche, J. R., J. K. Kay, N. C. Friggens, J. J. Loor, and D. P. Berry. 2013. Assessing and managing body condition score for the prevention of metabolic disease in dairy cows. Vet. Clin. North Am. Food Anim. Pract. 29:323–336. https://doi.org/10.1016/j.cvfa.2013.03.003.
Rodríguez Alvarez, J., M. Arroqui, P. Mangudo, J. Toloza, D. Jatip, J. M. Rodríguez, A. Teyseyre, C. Sanz, A. Zunino, C. Machado, and C. Mateos. 2018. Body condition estimation on cows from depth images using convolutional neural networks. Comput. Electron. Agric. 155:12–22. https://doi.org/10.1016/j.compag.2018.09.039.
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. Bokkers, P. W. Groot Koerkamp, T. Van Hertem, S. Viazzi, C. E. 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.
Schneider, C. A., W. S. Rasband, and K. W. Eliceiri. 2012. NIH Image to ImageJ: 25 years of Image Analysis. Nat. Methods 9:671–675. https://doi.org/10.1038/nmeth.2089.
Schröder, U. J., and R. Staufenbiel. 2006. Invited review: Methods to determine body fat reserves in the dairy cow with special regard to ultrasonographic measurement of backfat thickness. J. Dairy Sci. 89:1–14. https://doi.org/10.3168/jds.S0022-0302(06)72064-1.
Schuster, C. 2004. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. Educ. Psychol. Meas. 64:243–253. https://doi.org/10.1177/0013164403260197.
Shi, W., B. Dai, W. Shen, Y. Sun, K. Zhao, and Y. Zhang. 2023. Automatic estimation of dairy cow body condition score based on attention-guided 3D point cloud feature extraction. Comput. Electron. Agric. 206:107666. https://doi.org/10.1016/j.compag.2023.107666.
Siachos, N., G. Oikonomou, N. Panousis, G. Banos, G. Arsenos, and G. E. Valergakis. 2021. Association of body condition score with ultrasound measurements of backfat and longissimus dorsi muscle thickness in periparturient Holstein cows. Animals (Basel) 11:818. https://doi.org/10.3390/ani11030818.
Song, X., E. A. M. Bokkers, S. van Mourik, P. W. G. Groot Koerkamp, and P. P. J. van der Tol. 2019. Automated body condition scoring of dairy cows using 3-dimensional feature extraction from multiple body regions. J. Dairy Sci. 102:4294–4308. https://doi.org/10.3168/jds.2018-15238.
Spoliansky, R., Y. Edan, Y. Parmet, and I. Halachmi. 2016. Development of automatic body condition scoring using a low-cost 3-dimensional Kinect camera. J. Dairy Sci. 99:7714–7725. https://doi.org/10.3168/jds.2015-10607.
Strieder-Barboza, C., A. Zondlak, J. Kayitsinga, A. F. A. Pires, and G. A. Contreras. 2015. Lipid mobilization assessment in transition dairy cattle using ultrasound image biomarkers. Livest. Sci. 177:159–164. https://doi.org/10.1016/j.livsci.2015.04.020.
Tan, M., and Q. Le. 2021. EfficientNetV2: Smaller models and faster training. Pages 10096–10106 in Proc. 38th Int. Conf. on Machine Learning. Virtual Event. Proceedings of Machine Learning Research.
Vasseur, E., J. Gibbons, J. Rushen, and A. M. de Passillé. 2013. Development and implementation of a training program to ensure high repeatability of body condition scoring of dairy cows. J. Dairy Sci. 96:4725–4737. https://doi.org/10.3168/jds.2012-6359.
Weber, A., J. Salau, J. H. Haas, W. Junge, U. Bauer, J. Harms, O. Suhr, K. Schönrock, H. Rothfuß, S. Bieletzki, and G. Thaller. 2014. Estimation of backfat thickness using extracted traits from an automatic 3D optical system in lactating Holstein-Friesian cows. Livest. Sci. 165:129–137. https://doi.org/10.1016/j.livsci.2014.03.022.
Wright, I. A., and A. J. F. Russel. 1984. Partition of fat, body composition and body condition score in mature cows. Anim. Sci. 38:23–32. https://doi.org/10.1017/S0003356100041313.
Yukun, S., H. Pengju, W. Yujie, C. Ziqi, L. Yang, D. Baisheng, L. Runze, and Z. Yonggen. 2019. Automatic monitoring system for individual dairy cows based on a deep learning framework that provides identification via body parts and estimation of body condition score. J. Dairy Sci. 102:10140–10151. https://doi.org/10.3168/jds.2018-16164.
Zhao, K., M. Zhang, W. Shen, X. Liu, J. Ji, B. Dai, and R. Zhang. 2023. Automatic body condition scoring for dairy cows based on efficient net and convex hull features of point clouds. Comput. Electron. Agric. 205:107588. https://doi.org/10.1016/j.compag.2022.107588.

ORCIDS

N. Siachos https://orcid.org/0000-0001-7670-4950
A. Anagnostopoulos https://orcid.org/0000-0002-5193-858X
B. E. Griffiths https://orcid.org/0000-0002-2698-9561
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