Prediction of body condition in Jersey dairy cattle from 3D images using machine learning techniques

R.B. Stephansen[1], C.I.V. Manzanilla-Pech[1], G. Gebreyesus[1], G. Sahana[1] and J. Lassen[1,2]

[1]Center for Quantitative Genetics and Genomics, Aarhus University, C. F. Møllers Alle, 8000 Aarhus, Denmark
[2]Viking Genetics, Ebeltoftvej 16, Assentoft, 8960 Randers, Denmark
Corresponding Author: rasmus.stephansen@qgg.au.dk

Abstract

The body condition of a dairy cow is one of the important indicators of the animal's welfare and health status. Maintaining optimal body condition in dairy cows is associated with more functional cows (healthy, fertile, etc.). Currently, the assessment of body condition in dairy cows is performed through manual scoring by trained classifiers, which is labor intensive and limits frequent application on farms. The use of computer-vision shows great potential as a high-throughput method for predicting the body condition score (**BCS**) of cows. However, despite its promise, no study has investigated the predictive ability of using 3D cameras to assess BCS in Jersey dairy cattle. Data from three commercial farms with 808 individual cows was obtained every second month from December 2021 to August 2022, with a total of 2,253 BCS observations. Body condition scores were scored by two trained classifiers from SEGES (Aarhus, Denmark). The feature data consisted of contours from top-down 3D images, generated when a cow leaves the milking area. The features represent the depth on specific points of the back. When a cow enters the image frame, the spine and circumference are identified, and a 3D cloud of the back is made within the circumference. The features used in this study, were the points on the back where there was a drop from the spine of 3, 5, 10, 15 cm each side. For each of these drops, 100 features were generated from the neck to the tail of the cow. Splitting the training and validation data was carried out as a random split of 7:3 clustered by cows and replicated 10 times. The clustering by cows ensured that cows could not appear in both the training and validation dataset. The H_2O AutoML algorithm was used to find the best performing classification and regression model. Furthermore, AutoML was used to tune input parameters for the machine learning model. Among classification and regression models, DeepLearning performed best. Additionally, a Partial Least Square (PLS) model was tested with the Proc PLS procedure in SAS software. Validating the classification model, showed accuracies with a weighted mean of 48.1% (range: 45.9-50.7%) on the exact phenotypic class. The accuracy increased to a weighted mean of 93.5% (range: 92.7-95.3 %) by adjusting a 0.5-unit deviation. The results from the regression models showed R^2 and RMSE at 0.67 and 0.31 for PLS and 0.66 and 0.29 for DeepLearning. The validation accuracies were comparable to reports for Holstein cows in the literature. The results indicate that we can predict BCS in Jersey cows with a 3D camera-based system, which potentially could be used to improve management decisions in Jersey dairy herds.

Keywords: Body condition, 3D-images, Jersey dairy cattle, machine learning.

Introduction

Body condition is a widely acknowledged and accepted indicator for dairy cows welfare (Welfare Quality®consortium, 2009). Maintaining good management of dairy cow's body condition is associated with more functional cows (healthy, fertile, etc.). The assessment of a dairy cow's body condition, is currently performed through manual scoring by trained classifiers through a body condition score (**BCS**) (Roche *et al.,* 2009). That is labor intensive and a frequent routine application on commercial farms is limited. Therefore, BCS at multiple times over the lactation is mostly recorded only in nucleus and research farms. From a dairy management perspective, frequent and precise BCS data on commercial dairy herds could improve animal welfare and functionality. In addition, genetic evaluation model for feed efficiency lacks a phenotype for BCS to distinguish between adipose and muscle tissue (Stephansen *et al.,* 2021a). Availability of accurate phenotypes for BCS could potentially improve modelling of feed efficiency especially in early lactation. Research in high-throughput methods to predict daily BCS in commercial farms has been applied with varying accuracy and level of automatization (Qiao *et al.,* 2021). Most studies have used 2D or 3D camera technology to develop machine learning (**ML**) algorithms to predict BCS in Holstein cows (Qiao *et al.,* 2021). However, despite its promise, no study has investigated the predictive ability of using 3D cameras to assess BCS in Jersey dairy cattle. The aim of the study was to establish a reliable prediction of body condition using 3D-images and ML techniques in Danish Jersey cows on commercial farms.

Material and methods

Body condition score recording

Three Danish commercial Jersey farms participated in the project, with an annual herd size of 150, 260 and 280 cows. The scoring was performed by two trained classifiers from SEGES (Skejby, Denmark, *https://www.seges.dk/*) every second month from December 2021 to August 2022. The classifiers took rotation to visit the project herds during this recording period and classified all cows in the herds. In total 2,253 BCS phenotypes were recorded on 808 Jersey cows. The cows were scored on a scale from 1 to 9 following ICAR (2022). As most studies (Qiao *et al.,* 2021) and farms use the 1 to 5 scale, the score were transformed to the 1 to 5 scale following Garnsworthy (2006):

$$BCS = 0.5 \times score + 0.5 \quad (1)$$

Basic information, such as calving date (December 2020 to August 2022) and lactation data (parity range 1-9, average parity 2.65; days in milk in the range 10-401, average days in milk 142.4 days), were extracted from the Danish Cattle database.

Feature extraction and quality control

Feature data of the animals within ±3 days from the day of BCS scoring were provided by VikingGenetics (Randers, Denmark). Detailed description of the hardware and software used in capturing and processing 3D-images into contour features can be found in Gebreyesus *et al.* (2023), Lassen and Borchersen (2022) and Lassen *et al.* (2023). Briefly, the hardware used was a 3D camera using time-of-flight technology (Microsoft Xbox One Kinect v2), placed in a narrow corridor through which cows leave the milking area. The feature data were generated as contours from top-down 3D camera-images and represent the depth on specific points of the back. The camera is triggered by an electronic identification of the animal. In this case an electronic ear tag. When a cow enters the image frame, the spine and circumference are identified, and a 3D cloud

on the back is made within the circumference. The features used in this study were the points on the back where there was a drop from the spine of 3, 5, 10, 15 cm each side. For each of these drops and the spine, 190 features were generated from the neck to the tail of the cow. In total 950 features across the four contours and spine.

Quality control was undertaken on the feature data using the SAS software version 9.4 (SAS Institute Inc, 2013), to remove outlier values. Features were set missing for values out of the range of mean ± 3SD. This was done twice by cow and date of evaluation. Hereafter features with a missing rate higher than 25% were discarded. Cows has on average 32.8 pictures (SD of 11.9) per round of classification. Animals with fewer than five pictures per classification round were removed. This resulted in a total of 700 features used as predictors for training the models. We calculated a mean feature per round of scoring to give the most stable prediction of BCS. A mean feature was calculated for all individual features by cow and classification date for the individual features and weighted by

$$\frac{1}{1+|date\ of\ feature - date\ of\ classification|} \quad (2)$$

The weighting was used to put emphasis on features from the day of classification, assigning more weight to closer days apart between visual classification and image data capture.

Data split, learning algorithms and evaluation metrics

Splitting training and validation datasets for model development is commonly done with a 7:3 random split of the data (Rodríguez Alvarez *et al.,* 2019, Yukun *et al.,* 2019). The 7:3 random split was performed using Proc Survey procedure in SAS version 9.4, and clustered by cow ID to ensure individual (cows) only appeared in either the training or validation dataset. Ten replicates of training and validation datasets were created for the model development. The two most extreme BCS classes (1.0 and 5.0) were grouped with the immediate next class due to very low observations (three in each) and to ensure adequate observations were available for the learning step.

We used the *AutoML* algorithm from H_2O package in R (LeDell *et al.,* 2022) for testing best-performing classification and regression algorithms. We used the first training dataset replicate in the *AutoML,* to test which ML algorithm performed best. The non-default parameters in *AutoML* were set to test maximum 2,000 models for classification or regression and had seed set to 1 and *nfolds* to 10. Common class predictors including classifier, parity number, round of classification and herd were considered across all the ML methods. Predictors were features from 3D-images, which were standardized to a mean of 0 and SD of 1, and Legendre polynomials fitted on weeks of lactation up to 5th order. Tuning parameters for the various classification and regression models in the *AutoML* algorithm were optimized based on cross-validation with "*logloss*" and mean squared error (**MSE**) as optimizing metrics for the classification and regression models, respectively.

The best performing algorithm for both classification and regression were DeepLearning (**DL**) which is a multi-layer feedforward artificial neural network algorithm in H_2O. In addition, we tested a Partial Least Square (**PLS**) model, as it works well on correlated predictors (James *et al.,* 2013). The PLS model was tested in SAS with the Proc PLS procedure (SAS Institute Inc, 2013) fitting the same features and class variables as in the DL algorithm. The first training and validation dataset was used to fine-tune the PLS model and to define the optimum number of components. The tunning process of PLS showed 20 components were the optimum.

Output from the validation process of classification models were grouped into four individual classes based on confusion matrices between observed and predicted BCS: True Positives (**TP**), True Negative (**TN**), False Positives (**FP**) and False Negatives (**FN**).

Accuracy of classification (**AOC**) defined as (Rodríguez Alvarez *et al.,* 2019):

$$AOC, \% = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100 \quad (3)$$

F1-score is a measure that combined the trade-offs of precision and recall and defined as:

$$F1, \% = \left(2 \times \frac{TP \times FP}{TP + FP}\right) \times 100 \quad (4)$$

Accuracy of classification and F1-score were evaluated for their ability to predict on the exact phenotype and with a 0.50-unit deviation (**DEV**) to account for the human error judgement. For regression models, R-square ($\mathbf{R^2}$) and Root Mean Squared Error (**RMSE**) were used and estimated with the Proc ANOVA procedure in SAS. Another evaluation parameter for the regression methods, was to evaluate the percentage of predicted BCS phenotypes that were equal to the observed phenotype and on the exact phenotype and with a 0.5-unit DEV. This was implemented by rounding the predicted BCS phenotype from a regression model to the nearest 0.5-unit. The percentage of correctly assigned phenotypes were then reported for each class of observed BCS, but also a weighted average based on frequency was reported.

Results and discussion

The AOC of DL models were 48.1% for the exact phenotype (range 45.9 to 50.7%). With a 0.5-unit DEV the AOC of DL models increased to 93.5 (range 92.7 to 95.3%). Rodríguez Alvarez *et al.* (2019) estimated a lower AOC on the exact phenotype 41.2%, compared to this study. However, they found a higher AOC of 97.4% with a 0.5-unit DEV. Rodríguez Alvarez *et al.* (2019) developed an ensemble model from Convolutional Neural Network (**CNN**) models, trained on 1,661 Holstein cows in Argentina. A study on 512 Chinese Holstein cows by Shi *et al.* (2023) reported an AOC of 49% on the exact phenotype and 96% with a 0.5-unit DEV. Both studies of Rodríguez Alvarez *et al.* (2019) and Shi *et al.* (2023) used complex CNN models, which have high computational requirements, compared to simpler models (regression).

For the F1-score, a tradeoff metric between precision and recall, we found a weighted average of 46% on the exact phenotype (Table 1). A lower level was reported by Rodríguez Alvarez *et al.* (2019) at 38% for the best CNN model. Shi *et al.* (2023) estimated an average F1-score for the exact phenotype at 44%. With a 0.5-unit DEV in this study, the weighted average of the F1-score increased to 91%. That was lower than the 97% in Rodríguez Alvarez *et al.* (2019) and 95% in Shi *et al.* (2023).

The approximated AOC from regression models (Table 2), showed that the choice among the PLS and DL algorithms in this study were limited. On the exact phenotype both regression models performed better on AOC (51.2-52.0%) than the DL classification model (48.1%), but also higher than Rodríguez Alvarez *et al.* (2019) and Shi *et al.* (2023). Allowing a 0.5-unit DEV increased the weighted average of AOC to 95.5 and 96.1% for

Table 1. Validation results for sensitivity, precision, and F1-score in percentage for DL, using. The parenthesis represents the range among replicates. DL = DeepLearning, BCS = Body Condition Score, Exact = exact score, DEV = deviation, WAvg = weighted average by frequency.

	F1-Score	
BCS	Exact	0.5-unit DEV
1.5	3 (0-14)	39 (0-100)
2.0	59 (51-63)	98 (97-99)
2.5	55 (52-57)	96 (95-97)
3.0	36 (27-43)	94 (93-97)
3.5	42 (32-48)	85 (80-92)
4.0	9 (0-34)	81 (73-91)
4.5	4 (0-25)	13 (0-57)
WAvg	46 (44-49)	91 (89-94)

Table 2. Validation results from regression models. The parenthesis represents the range among replicates. PLS = Partial Least Square, DL = DeepLearning, BCS = Body Condition Score, R2 = R-square, RMSE = Root Mean Square Error, Exact = exact score, DEV = deviation, WAvg = weighted average by frequency.

	PLS		DL	
BCS	Exact	0.5-unit DEV	Exact	0.5-unit DEV
1.5	33 (19-52)	91 (80-100)	16 (6-29)	89 (72-100)
2.0	49 (46-52)	97 (94-99)	50 (45-54)	97 (95-99)
2.5	60 (56-69)	98 (96-99)	67 (61-73)	98 (97-99)
3.0	55 (51-58)	98 (97-100)	52 (44-60)	98 (97-99)
3.5	45 (36-51)	94 (91-99)	41 (34-48)	91 (86-98)
4.0	23 (10-35)	86 (80-96)	23 (7-29)	78 (71-83)
4.5	9 (0-20)	65 (42-78)	9 (0-20)	64 (40-78)
WAvg	51.2	96.1	52.0	95.5
R^2	0.67 (0.65-0.68)		0.66 (0.64-0.68)	
RMSE	0.31 (0.29-0.33)		0.29 (0.26-0.32)	

DL and PLS respectively. This was higher than the DL classification model (Table 1) and similar level as reported in Rodríguez Alvarez *et al.* (2019) and Shi *et al.* (2023).

Conclusions

The aim was to build a reliable prediction algorithm of BCS using 3D-images and ML techniques in Danish Jersey cows on commercial farms. Among classification and regression models, DL performed best. Additionally, a PLS model was tested. Validating the classification model, showed an accuracy of 48.1% (range: 45.9-50.7%) on the exact phenotype. The accuracy increased to 93.5% (range: 92.7-95.3 %) with a 0.5-unit DEV. The results from the regression models showed R^2 and RMSE at 0.67 and 0.31 for PLS and 0.66 and 0.29 for DL. The approximated AOC for regression models showed for PLS 51.2 and 96.1% and for DL 52.0 and 95.5% on the exact and 0.5-unit DEV, respectively. The results indicate that we can predict BCS in Jersey cows with contour features from a 3D camera-based system in ML models. This can potentially improve management decisions on Jersey dairy herds.

Acknowledgements

The research leading to the study results was funded by the CFIT project (9090-00083B), funded by the Innovation Fund Denmark. The authors thank the Danish classifiers Jørgen Knudsen and Villy Nicolajsen for an excellent job on scoring cows' body conditions and Mogens H. Madsen for administrating the classification. The authors also thank the farmers Palle Bjerggaard Hansen, Hans Ulvsbjerg Rasmusssen and Viacheslav Kreitor for participating with their herds in the project.

References

Garnsworthy, P. C. 2006. Body condition score in dairy cows: targets for production and fertility. Rec. Adv. An. 40:61.

Gebreyesus, G., V. Milkevych, J. Lassen, and G. Sahana. 2023. Supervised learning techniques for dairy cattle body weight prediction from 3D digital images. Front. Genet.https://doi.org/10.3389/fgene.2022.947176.

ICAR. 2022. Section 5 - ICAR Guidelines for Conformation Recording of Dairy Cattle, Beef Cattle, Dual Purpose Cattle and Dairy Goats. Accessed Jan. 16., 2023. https://www.icar.org/Guidelines/05-Conformation-Recording.pdf

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An introduction to statistical learning. Vol. 112. Springer.

Lassen, J. and S. Borchersen. 2022. Weight determination of an animal based on 3d imaging. Accessed 25. April, 2023. https://patents.google.com/patent/US20220221325A1/en

Lassen, J., J. R. Thomasen, and S. Borchersen. 2023. Repeatabilities of individual measure of feed intake and body weight on in-house commercial dairy cattle using a 3D camera system. J. Dairy. Sci. Accepted.

LeDell, E., N. Gill, S. Aiello, A. Fu, A. Candel, C. Click, T. Kraljevic, T. Nykodym, P. Aboyoun, M. Kurka, and M. Malohlava. 2022. R Interface for the 'H2O' Scalable Machine Learning Platform. Accessed https://CRAN.R-project.org/package=h2o

Qiao, Y., H. Kong, C. Clark, S. Lomax, D. Su, S. Eiffert, and S. Sukkarieh. 2021. Intelligent perception for cattle monitoring: A review for cattle identification, body condition score evaluation, and weight estimation. Comput. Electron. Agr. 185:106143.https://doi.org/10.1016/j.compag.2021.106143.

Roche, J. R., N. C. Friggens, J. K. Kay, M. W. Fisher, K. J. Stafford, and D. P. Berry. 2009. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. J. Dairy Sci. 92(12):5769-5801.https://doi.org/10.3168/jds.2009-2431.

Rodríguez Alvarez, J., M. Arroqui, P. Mangudo, J. Toloza, D. Jatip, J. M. Rodriguez, A. Teyseyre, C. Sanz, A. Zunino, and C. Machado. 2019. Estimating body condition score in dairy cows from depth images using convolutional neural networks, transfer learning and model ensembling techniques. Agronomy 9(2):90.

SAS Institute Inc. 2013. The PLS Procedure. Accessed 25. February, 2023. https://support.sas.com/documentation/onlinedoc/stat/131/pls.pdf

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. Agr. 206:107666 .https://doi.org/10.1016/j.compag.2023.107666.

Stephansen, R. B., J. Lassen, J. F. Ettema, L. P. Sørensen, and M. Kargo. 2021a. Economic value of residual feed intake in dairy cattle breeding goals. Livest. Sci. 253:104696.https://doi.org/10.1016/j.livsci.2021.104696.

Welfare Quality®consortium. 2009. Assessment protocol for cattle Accessed 24. February, 2023. https://edepot.wur.nl/233467

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(11):10140-10151.