Smart Agricultural Technology 4 (2023) 100199

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Towards a novel method for detecting atypical lying down and standing up behaviors in dairy cows using accelerometers and machine learning

Stijn P. Brouwers [a,b,*], Michael Simmler [c], Pascal Savary [a], Madeleine F. Scriba [a]

[a] *Centre for Proper Housing of Ruminants and Pigs, Federal Food Safety and Veterinary Office (FSVO), Agroscope, Tänikon 1, 8356 Ettenhausen, Switzerland*
[b] *Animal Physiology, Institute of Agricultural Sciences, ETH Zurich, Universitätstrasse 2, Zurich 8092, Switzerland*
[c] *Digital Production, Agroscope, Tänikon 1, 8356 Ettenhausen, Switzerland*

ARTICLE INFO

Edited by: Stephen Symons

Keywords:
Animal welfare assessment
Free-stall cubicle
Time series classification
Sensor
Behavioral classification

ABSTRACT

Free-stall cubicles are designed so that cows do not defecate in the bedding material, while still providing comfortable lying places. However, fixed cubicle elements, such as the partition and neck rail, restrict available movement space and may hinder cows from performing natural lying down and standing up movement patterns. Although there are various types of cubicle partitions that differ in shape, dimensions, or materials, there is no method other than visual observations to assess their effects on cow welfare. An automated detection system could improve the efficiency and promote objectivity of such assessments. Therefore, the aim of this research was to explore which atypical lying down and standing up behaviors could be detected using body-mounted accelerometers and machine learning. Three leg- and one head-mounted accelerometer set to record at 20 Hz were fitted to 48 lactating dairy cows (Brown Swiss and Holstein × Swiss Fleckvieh). Lying down and standing up events were simultaneously assessed through video observations, by assigning binary presence/absence labels for atypical behaviors, such as lunging the head sideways when standing up and pawing the bedding material before lying down. Different time series classification algorithms were employed for model development using a nested cross-validation strategy. The best performing classifiers were MiniRocket and the deep learning algorithm InceptionTime. Atypical behaviors performed during standing up events, namely *Hesitant head lunge* and *Crawling backwards*, were identified as most promising candidates for accelerometer-based detection. These behaviors were detected with balanced accuracies of 0.67 and 0.74, respectively, and their learning curves indicated that more training data might further improve model performance. Overall, achieved performances were not yet satisfactory for application in the evaluation of new dairy cow housing installations. Potentially, ethograms designed for human observers are not optimal for machine learning and adjustments with machine learnability in mind might be necessary. The behaviors identified as promising are good candidates for further development into an efficient and objective method that could complement human observations in the assessment of dairy cow housing installations.

1. Introduction

Lying cubicles in free-stall systems are designed to provide comfortable lying places for dairy cows while maintaining proper hygiene [1,2]. Free-stall cubicles are separated by partitions and fitted with a neck rail or band above the lying place. These fixed elements prevent cows from standing fully inside stalls and ensure that animals lie down near the end of the stall with their rear towards the walking alley. In these standing and lying positions, cows do not defecate in the bedding material but on the concrete surface of the walking alley which is regularly cleaned by a manure scraper. However, when fixed cubicle elements restrict available space and limit freedom of movement, stall cleanliness may come at the cost of cow welfare [3].

Limited movement space inside free-stall cubicles can specifically hinder cows from performing natural lying down and standing up movements. These posture transitions follow species-specific, innate

Abbreviations: TSC, time series classification; LHL, left hind leg; RFL, right front leg; LFL, left front leg; H, head; CV, cross-validation; MTSC, multivariate time series classification.

[*] Corresponding author at: Centre for Proper Housing of Ruminants and Pigs, Federal Food Safety and Veterinary Office (FSVO), Agroscope, Tänikon 1, 8356 Ettenhausen, Switzerland.

E-mail address: stijn.brouwers@agroscope.admin.ch (S.P. Brouwers).

https://doi.org/10.1016/j.atech.2023.100199
Received 28 November 2022; Received in revised form 7 February 2023; Accepted 13 February 2023
Available online 15 February 2023

movement patterns in dairy cows [4,5]. The movements are largely determined by skeletal and muscular structure, leaving cows with limited ability to adapt them to their environment [6]. The prevalence of specific atypical behaviors during posture transitions, such as lunging the head sideways when standing up and pawing the bedding material before lying down, can indicate inadequate movement space in the lying cubicles [7,8]. Insufficient space in cubicles combined with other risk factors can cause ulcers, bruises on the metacarpal/metatarsal joints, hock lesions, and lameness [9,10,11]. These health issues can result in economic losses for the farmer because cows might stay in the herd for a shorter time and milk production, fertility, and slaughter value may be reduced [12,13]. The breeding selection for a higher milk yield and the associated increase in body size of dairy cows in recent decades have accentuated these issues [14].

To optimise free-stall cubicles both from a hygienic and animal welfare perspective, manufacturers of housing systems have developed various types of cubicle partitions (e.g. [15]). Assessing the impact of each type of partition on animal welfare is challenging as differences in e.g. shape, dimensions, and materials of the partition have diverse effects on cow behavior. Currently, novel partition types are evaluated through visual analysis of the cows' behavior in the cubicles by a human observer using either direct observations or video recordings. This method is labour intensive and to a certain degree subjective [16]. Even experienced assessors can be influenced by contextual factors, such as the overall cleanliness of the stall. Thus, there is the need for a method to support assessments made by human observers in the evaluation of free-stall housing installations with regard to cow welfare.

In contrast to visual observations by a human observer, automatic detection systems can provide efficient and objective methods to monitor animal behavior. Sensors, such as accelerometers, are now frequently used to study and monitor cow behavior. From acceleration data, general activities [17–19], lying behavior [20,21], grazing and rumination behavior [22–24], and health problems such as lameness can be tracked [25,26]. This data gives valuable insight into the welfare and health of animals, and enables farmers, veterinarians, and researchers to make informed decisions [27]. Similarly, manufacturers of mass-produced housing installations and regulatory authorities could evaluate the effects of new products using automatic detection systems [28].

Particularly, detecting atypical behaviors performed during posture transitions would be of great value because these are considered informative for decreased cow welfare in a stall-based housing environment [4,8]. Lying down and standing up events per se can be detected reliably with leg-mounted accelerometers and rule-based algorithms [29]. However, models for the detection of atypical movements performed during these posture transitions are currently not available. The task of detecting specific atypical movements during these transitions is far more complex than recognising the posture transitions themselves. However, results from human research suggest that it is possible to detect specific characteristics of standing-to-sitting and sitting-to-standing transitions from accelerometer data [30–32].

In general, supervised classification models are trained to flag events of interest from 'features' obtained by selecting and transforming the raw data [33,34]. Accelerometers collect data in the form of time series, which are often manually transformed into tabular features (e.g. maximum, mean, variance) to be compatible with standard machine-learning algorithms. However, recent advances in the field of time series classification (TSC) have provided classifiers that can effectively learn from raw time series data, without the need of manual feature engineering [35,36]. Many of them are particularly good in exploiting the temporal structure of the data, whereas this information is often poorly preserved when doing manual feature engineering. Therefore, bespoke time series classifiers might be more effective to detect atypical behaviors from accelerometer data than standard machine learning classifiers with manual feature engineering.

The aim of this study was to explore which atypical lying down and standing up behaviors are the most promising candidates to be detected with body-mounted accelerometers and machine learning. For this purpose, eight atypical behaviors described by Zambelis et al. [8] and Dirksen et al. [7] were targeted with individual binary classifiers predicting the presence/absence of the behavior. Several recent, conceptually different TSC algorithms, including a deep learning approach, were employed. Additionally, practical aspects of data collection and modelling, such as the effect of accelerometer sampling frequency on classifier performance, were investigated.

2. Methods

Ethical approval for the study was obtained from the Veterinary Office of the Canton Thurgau (Switzerland; TG03/2021, Approval No. 33448).

2.1. Animals and housing conditions

The study was conducted during the summer and autumn of 2021 at the Agroscope research station in Ettenhausen, Switzerland. Data was collected from 48 lactating cows of the two most common dairy breeds in Switzerland (Table 1; 34 Brown Swiss and 14 Holstein × Swiss Fleckvieh). Withers height of the cows was 146.3 ± 4.8 cm (mean ± SD). Cows were selected opportunistically based on availability to participate in the study.

The cows were housed in a free-stall barn that consisted of an exercise yard and three sections, one of which was used in this study. This section consisted of two rows of wall-facing deep-bedded cubicles (eight and nine cubicles, respectively) with a walking alley in between and a feeding alley on the opposite side of the wall of one of the cubicle rows. The cows had constant access to the exercise yard and additionally to the adjacent pasture whenever the weather allowed it. Cubicle partitions were of type *Liberty* (Krieger AG, Ruswil, Switzerland), and more than one cubicle was available per cow. A flexible neck band was installed. Cubicles measured 125 cm in width and 265 cm in length with a brisket board 200 cm from the curb and 65 cm head lunge space (Fig. 1). For the cubicle row adjacent to the feeding alley, a wooden board was removed from the head-facing wall to increase the head lunge space to >65 cm.

Cubicles were maintained twice a day, including removing faeces and levelling of bedding material. The walking and the feeding alley were scraped eight times per day by a manure scraper robot. The cows were milked in a milking parlour twice a day, between 0500 and 0600 h and between 1600 h and 1700 h, respectively. They were fed a mixed ration twice a day at approximately 0900 h and after the afternoon milking. The mixed ration contained maize, grass, and hay silages, as well as concentrate and minerals. Additional concentrate was offered according to animal-individual allowance in an automated feeding station. Water was available ad libitum from a self-filling water trough.

2.2. Data collection

The complete workflow for data collection, data pre-processing, and model development and evaluation is shown in Fig. 2.

Table 1
Withers height, age, and lactation number (mean and range) of cows summarised per breed.

	Withers height		Age		Lactation number	
Breed	*Mean*	*Range*	*Mean*	*Range*	*Mean*	*Range*
Brown Swiss	145	133–155	5.6	2–13	3.8	1–11
Holstein × Swiss Fleckvieh	148	140–155	4.0	2–6	2.2	1–4

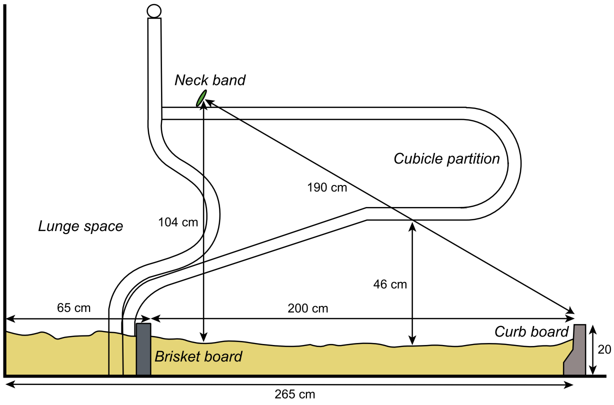

Fig. 1. Lying cubicle design and dimensions (partition type *Liberty*, Krieger AG, Ruswil, Switzerland).

2.2.1. Acceleration data

Acceleration data was recorded along three Cartesian axes (x, y, and z) at ~20 Hz (512 * 26^{-1} Hz) using MSR 145 data loggers with a tri-axial accelerometer (MSR Electronics GmbH, Seuzach, Switzerland; hereafter referred to as 'sensors'). Data was stored on the 30 MB internal memory of the sensor. With a sampling frequency of 20 Hz, the battery life of the sensors was longer than a week. The memory capacity was the limiting factor during data collection, since the memory filled up before battery life was depleted. The working range of the accelerometers was ±15 *g*. Each cow was equipped with four sensors (Fig. 3a): one on the left hind leg (LHL), one on each front leg (RFL and LFL), and one on the left side of the head (H). Sensors LHL, RFL, and LFL were mounted on the outward-facing side of the metatarsus using a piece of foam and self-adhesive bandage (Fig. 3b). Sensor H was placed inside a leather pouch attached to a halter. All sensors were attached when the cows were fixed in the feeding rack during the morning feeding. The recordings started at 0600 h the next morning, giving the cows one day to get used to the sensors and relieve potential stress effects of mounting the sensors. Recordings stopped at 1600 h on the same day, and sensors were removed the next morning, read out, and set up for a new recording. Sensors' batteries were recharged to full capacity during the data readout. Sensor recordings were made on 22 individual days, each time with 10 cows.

2.2.2. Video data

Two video cameras (Bascom 4XB40K, Bascom, Vianen, The Netherlands) were permanently installed at a height of four metres on opposite sides of the barn so that all cubicles were visible from both cameras. Cameras were connected to a well-accessible recorder (Bascom R4XK, Bascom, Vianen, The Netherlands). Continuous video recordings were made simultaneously to the collection of accelerometer data. The cows were marked with a number on the flank (RAIDEX animal marking spray) to identify individuals from the video footage (Fig. 3a).

2.3. Data pre-processing

2.3.1. Identifying and cutting out posture transitions

Data pre-processing was largely done in R (v.4.2.0; [37]). First, lying down and standing up events were identified from the signal of the accelerometer mounted on the left hind leg using the workflow of the *triact* R package (v.0.2.0; [38]) with default parameters, apart from employing the *add_lying* method with a window size of 25 s for the median filter (window_size = 25) and no minimum lying bout duration (minimum_duration_lying = 0). The *triact* R package uses a simple rule-based algorithm to differentiate between standing and lying posture based on which axis of the leg-mounted sensor gravitation loads. In total, 560 lying down events and 569 standing up events were identified. The number of lying down and standing up events recorded per cow ranged from 2 to 26 (median: 11) and 2 to 33 (median: 12), respectively.

Based on the maximal duration (87 s) observed in the videos, a time window from 60 s before to 30 s after (90 s in total) the time of posture transition as identified using *triact* was chosen to cut out lying down events. Standing up events (max. observed duration 54 s) were cut out from 45 s before to 15 s after (60 s in total) the time identified by *triact*. This resulted in time series lengths of 1772 and 1182 data points for lying down and standing up events, respectively.

2.3.2. Video labelling

Lying down and standing up events in the videos were located using the timestamps obtained with the *triact* R package (see Section 2.3.1) and labelled using the scoring system proposed by Zambelis et al. [8] and the behaviors observed by Dirksen et al. ([7]; Table 2). The observer (S.P.B.) had a background in observing animal behavior and was trained by an experienced scientist (P.S.) to assess the specific cow behaviors relevant in this study. To determine intra-observer reliability, the observer scored the same set of 40 videos (20 lying down and 20 standing up movements) once before and once after the video labelling (three months in between). The level of agreement was almost perfect (Cohen's Kappa κ = 0.96; [39]). If a behavior was not clearly identifiable owing to poor video footage (e.g. too dark, too far away from camera), it was not labelled but instead noted as missing value and later excluded from the dataset. The number of observations of each behavior, the class distributions, and the number of different cows that performed the behaviors in each class are given in Table 2. Lorenz curves indicating which percentage of cows performed which proportion of observations in each class are shown in Supp. Fig. 1B and C.

2.4. Model development and performance evaluation

Model development was done in Python 3.7, using the machine learning framework scikit-learn v.1.0.1 [40] together with its companion packages imbalanced-learn v.0.9.0 [41], sktime v.0.8.1 [42], and sktime-dl v.0.4.0 (github.com/sktime/sktime-dl). Models were trained via Microsoft Azure Machine Learning Studio on F72s_v2 compute instances, apart from the GPU-based training of InceptionTime, which was conducted on NC4as_T4_v3 compute instances equipped with an NVidia T4 GPU.

The detection of atypical lying down and standing up behaviors from tri-axial accelerometer data can in our case be described as a multivariate times series classification (MTSC) task with *n* dimensions equal to three (axes) times the number of sensors. Individual MTSC models were developed to detect each atypical behavior. Of the four sensors attached to different body parts, between one and three sensors were used for training the classifiers (Table 2). Sensors were selected based on a preliminary analysis where all possible sensor combinations were systematically tested (MiniRocket models without hyperparameter tuning and balanced accuracy as performance metric). Reducing the number of sensors was intended with regard to the applicability of the proposed automated detection system. Therefore, using a smaller subset of sensors with similar performance was preferred over using a larger subset or all sensors, respectively. For both lying down and standing up, class distributions of *Non-species-specific* and *Interruption* were imbalanced to such an extreme that classifiers could not be fitted for these atypical behaviors (Table 2).

2.4.1. Machine learning models

Three recently proposed MTSC algorithms that are among the best performing algorithms according to common domain-agnostic benchmarks were compared: MiniRocket [43], HIVE-COTE 2.0 [44], and InceptionTime [45].

MiniRocket is a transform that generates features by transforming the input time series using a large number of convolutional kernels (10^4 kernels). These features are subsequently used to train a linear classifier such as a Ridge regression classifier [43]. MiniRocket has a remarkably

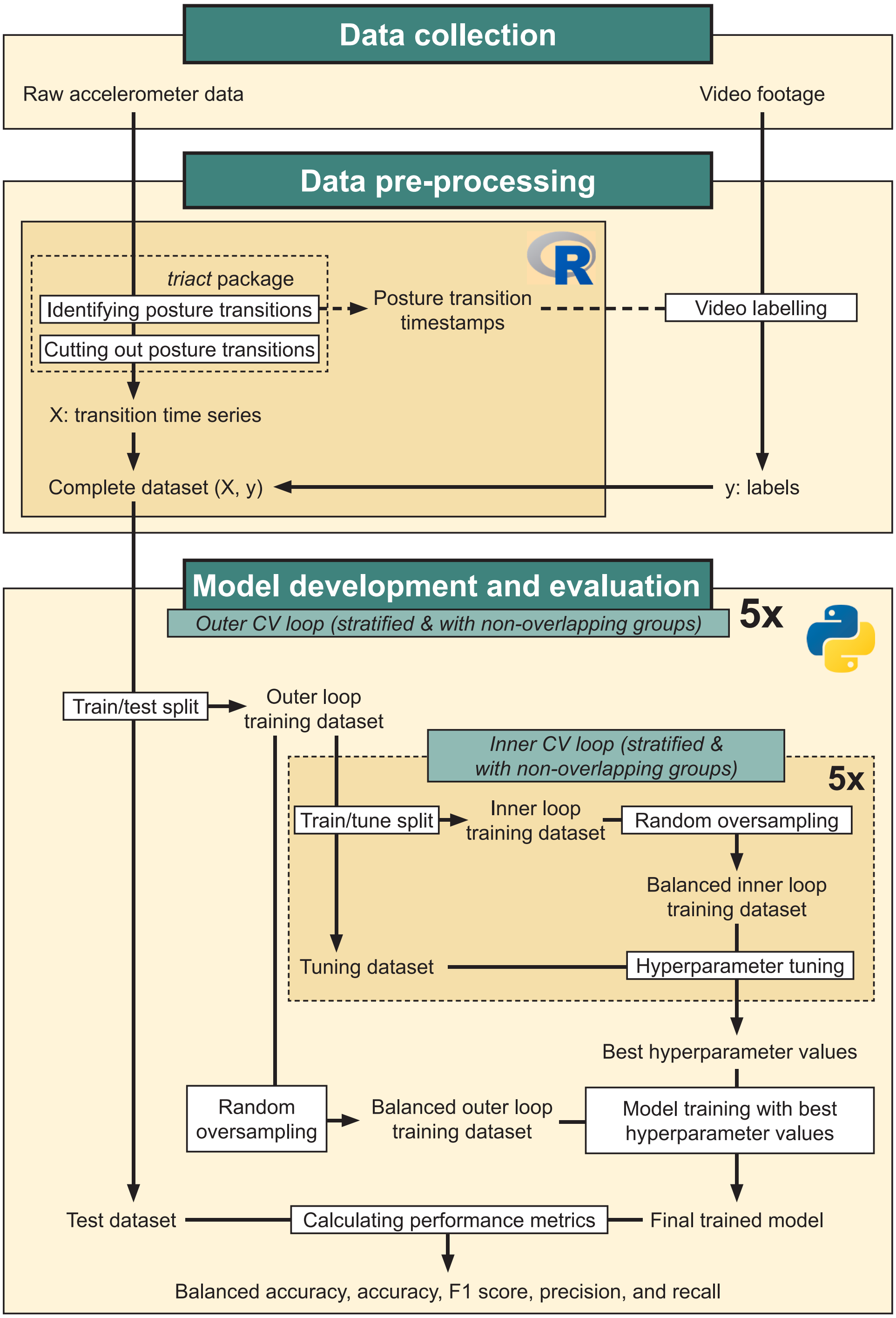
Data collection
Raw accelerometer data
Video footage
Data pre-processing
triact package
Identifying posture transitions
Cutting out posture transitions
Posture transition timestamps
Video labelling
X: transition time series
Complete dataset (X, y)
y: labels
Model development and evaluation
Outer CV loop (stratified & with non-overlapping groups)
5x
Train/test split
Outer loop training dataset
Inner CV loop (stratified & with non-overlapping groups)
5x
Train/tune split
Inner loop training dataset
Random oversampling
Balanced inner loop training dataset
Tuning dataset
Hyperparameter tuning
Best hyperparameter values
Random oversampling
Balanced outer loop training dataset
Model training with best hyperparameter values
Test dataset
Calculating performance metrics
Final trained model
Balanced accuracy, accuracy, F1 score, precision, and recall

Fig. 2. Workflow used to develop classification models for detecting atypical behaviors performed by dairy cows during lying down and standing up events from accelerometer data. CV = cross-validation.

low computational cost, while achieving high classification performances. HIVE-COTE 2.0 is a meta ensemble classifier that combines four TSC models that use features from different data domains: the shapelet-based Shapelet Transform Classifier, the interval-based Diverse Representation Canonical Interval Forest Classifier, the convolution-based Arsenal (ensemble of Rocket transforms, the predecessor of MiniRocket), and the dictionary-based Temporal Dictionary Ensemble. Each classifier is trained independently and makes a prediction as probability estimate. The probabilities are then combined and weighted by an estimate of the quality of each model to make a final prediction [44]. HIVE-COTE 2.0 is one of the most accurate TSC models currently available, but its computational costs are excessive. InceptionTime is a recent deep learning model for TSC. It is an ensemble of five deep Convolutional Neural Network models, each with the same architecture but different randomly initialised weights [45]. Each network is composed of a stack of multiple Inception modules. The core idea of these modules is that multiple filters of varying lengths are applied simultaneously. By stacking several of them, the network is able to extract latent hierarchical features of multiple resolutions [45]. InceptionTime has been applied to detect arm motor impairment from accelerometer bracelets worn by humans [46].

In addition to these three recently proposed algorithms, K-Nearest

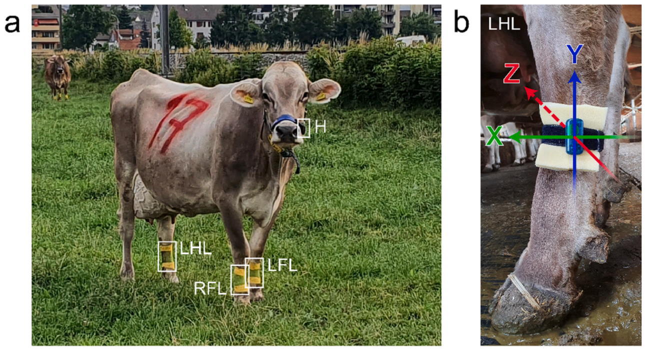

Fig. 3. (a) Sensors were attached to the left hind leg (LHL) and both front legs (RFL and LFL) with self-adhesive bandage, and to the left side of the head using a halter (H). (b) Detailed attachment of sensor LHL and orientation of the recording axes.

Neighbours with Dynamic Time Warping as distance measure (DTW + KNN) was applied [47]. This has long been considered as 'gold-standard' and is now a popular TSC benchmark [36]. Lastly, a Dummy classifier that ignores the input and always predicts the majority class in the training dataset was used as a baseline to contrast with the more complex classifiers.

2.4.2. Model development

A nested cross-validation (CV) strategy was used with an outer loop exclusively serving the purpose of evaluating the models' generalization performance and an inner loop for hyperparameter tuning (Fig. 2; [48]). This strategy prevents any leakage of information from the test dataset into the model and thus allows an unbiased estimate of the model's generalization performance [49]. In the inner CV loop, hyperparameter values (grid search) were evaluated using *stratified group 5-fold* CV. Each individual cow was considered as a group to ensure that data from the same cow were present exclusively in either the training or the tuning dataset. Additionally, the stratification ensured that percentages of observations from each class were preserved in each fold. Because behaviors were imbalanced to different extents, each training dataset was randomly oversampled to equalise the number of observations of the different classes.

Hyperparameter values explored in hyperparameter tuning are listed in Supp. Table 1. Due to its excessive computational cost, HIVE-COTE 2.0 was used with default hyperparameters without tuning. For these models, the inner loop was therefore obsolete and the strategy reduced to non-nested CV. Because behaviors were imbalanced to different extents and all classes were equally important, *balanced accuracy* was used as metric for model performance during tuning as well as during model evaluation (see next section). Balanced accuracy is insensitive to imbalanced class distributions, because it is the arithmetic mean of

Table 2
Ethogram of atypical behaviors performed by dairy cows during lying down and standing up events (adapted from [7,8]). Number of observations, class distributions, number of different cows that performed the behaviors in the classes, the number of observations that were not clearly identifiable and therefore excluded from the dataset ('NA'), and the sensors selected to detect the behaviors.

Behavior	Definition	No. of observations (distribution in %) [No. of cows]	Selected sensor(s)[1,2]
Lying down			
Non-species-specific (yes/no)[3]	Cow first lowers the hindquarters and then the forequarters ('dog-sitting')	*Yes*: 0 (0%) [0] *No*: 559 (100%) [48] NA: 1 [1]	–
Interruption (yes/no)	Carpal joints touch the ground, but the lying down movement is then interrupted by raising from the carpal joints	*Yes*: 2 (0.4%) [2] *No*: 557 (99.6%) [48] NA: 1 [1]	–
Repeated stepping with front legs (yes/no)	Stepping in place with front legs more than two times before the lying down movement	*Yes*: 61 (11.2%) [24] *No*: 483 (88.8%) [47] NA: 16 [13]	LHL, RFL, LFL
Extensive inspection (yes/no)	Head lowered and sweeping sideways (while sniffing the bed surface) more than two times before the lying down movement	*Yes*: 137 (24.9%) [45] *No*: 414 (75.1%) [48] NA: 9 [6]	H
Pawing (yes/no)	Pawing with front leg (possibly displacing bedding material) before the lying down movement	*Yes*: 47 (8.6%) [20] *No*: 498 (91.4%) [48] NA: 15 [13]	LHL, RFL, LFL
Standing up			
Non-species-specific (yes/no)[4]	Cow first raises the forequarters and then the hindquarters ('horse-like rising')	*Yes*: 1 (0.2%) [1] *No*: 568 (99.8%) [48] NA: 0 [0]	–
Interruption (yes/no)	Hindquarters lifted from the ground, but standing up movement is then interrupted by lowering the hindquarters (to the same or other side of the body)	*Yes*: 2 (0.4%) [2] *No*: 567 (99.6%) [48] NA: 0 [0]	–
Hesitant head lunge (yes/no)	Hesitant, interrupted, or repeated motion of the head during the head lunge movement	*Yes*: 149 (26.2%) [37] *No*: 419 (73.8%) [45] NA: 1 [1]	LHL, H
Sideways head lunge (yes/no)[5]	Head lunge movement is directed sideways by bending the head and neck to the side	*Right*: 200 (35.5%) [46] *Left*: 185 (32.8%) [44] *Straight*: 179 (31.7%) [43] NA: 5 [5]	LHL, H
Crawling backwards (yes/no)	Backwards movement on carpal joints after the head lunge	*Yes*: 71 (13.0%) [21] *No*: 477 (87.0%) [48] NA: 21 [16]	LHL, RFL, LFL

[1] Dash indicates that no models were developed for the atypical behavior owing to too imbalanced class distribution.
[2] LHL = left hind leg, RFL = right front leg, LFL = left front leg, *H* = head.
[3] In species-specific lying down posture transitions, cows first drop onto their carpal joints and then lower their hindquarters.
[4] In species-specific standing up posture transitions, cows first lift their hindquarters during the head lunge and then rise from their carpal joints.
[5] Models were trained on the three classes *left*, *right*, and *straight*, and predictions were reclassified to *yes* (right and left) and *no* (straight; see Section 2.4.3).

sensitivity and specificity (Eq. 1, [50]).

$$Balanced\ Accuracy = \frac{sensitivity\ +\ specificity}{2} \quad (1)$$

2.4.3. Performance evaluation

The outer CV loop served the purpose of evaluating model generalization performance (Fig. 2). As with the inner loop, folds were obtained using *stratified group 5-fold* CV with cows as groups. To ensure unbiased comparison, the same set of folds were used when comparing models. For each outer fold, the model with the best hyperparameter values, as determined in the corresponding inner loop, was fitted to the entire training dataset and evaluated on the test dataset using balanced accuracy, and additionally accuracy, F1 score, precision, and recall. Generalization performance and model robustness were then determined as mean and standard deviation, respectively, of these metrics across the five outer folds. Model robustness refers to the degree to which performance is affected when changes are made to the training dataset. *Sideways head lunge* was trained on three classes (*straight, left*, and *right*; see Table 2), but predictions were reclassified as *yes* or *no*. Preliminary tests revealed that exploiting this additional information on the direction increased performance for the final prediction of the binary labels as compared with using the binary labels for training directly. Finally, differences in performance between classifiers were evaluated using Bayesian correlated *t*-tests (hereafter simply refererred to as 'correlated *t*-tests'; [51]) with the *two_on_single* function of the *baycomp* Python package (v.1.0.2). Only differences of Bayes factor larger than $10^{1/2}$ were considered as substantial evidence for a performance difference between models [52,53].

2.5. Effect of training dataset size and accelerometer sampling frequency

MiniRocket was selected for further analysis because it was found to be comparatively well performing and remarkably fast to train. To assess whether more data would increase model performance, learning curves were generated by fitting MiniRocket models on subsets of the training dataset of varying sizes (10% to 100% of the full dataset in 20 steps). Additionally, the trade-off between accelerometer sampling frequency and classifier performance was investigated by fitting MiniRocket models on down-sampled datasets. This trade-off is of interest because a lower sampling frequency implies a lower power consumption and less data to be stored on the device, which in turn eases limitations of memory and battery capacity. The investigated frequencies were 20 Hz (original), 10, 5, 1 and 0.5 Hz. Here too, correlated *t*-tests were used to compare the generalization performance of models trained on data with different sampling frequencies. With a sampling frequency of 0.5 Hz, time series approached the minimal length for MiniRocket (nine time points). For both, studying the effect of training dataset size and that of accelerometer sampling frequency, models were fitted according to the outer CV loop as described above (Fig. 2), with hyperparameter values as found to be best for the model developed at 20 Hz with the full dataset (see Section 2.4.2).

3. Results

3.1. Performance

Best values for hyperparameters as identified in hyperparameter tuning (inner CV loop) are listed in Supp. Table 3. Generalization performances (outer CV loop) as described by balanced accuracies are shown in Fig. 4. All performance metrics (balanced accuracy, accuracy, F1 score, precision, and recall) are listed in Supp. Table 2. Results of correlated *t*-tests for comparison between the balanced accuracies achieved by the different classifiers are shown in Fig. 5 (comparison of means across folds; Bayes factor $> 10^{1/2}$ regarded as substantial evidence for a true difference in performance).

Crawling backwards was the best detected atypical behavior overall, with a balanced accuracy of 0.74 ± 0.02 (mean ± SD) with InceptionTime (Fig. 4). Correlated *t*-tests showed substantial evidence for a true difference in performance between InceptionTime and all other classifiers for detecting this behavior (Fig. 5). *Hesitant head lunge* was detected with a balanced accuracy of 0.67 ± 0.06 using MiniRocket. However, correlated *t*-tests showed no substantial evidence for a true difference in performance between MiniRocket and HIVE-COTE 2.0 for detecting this behavior *Extensive inspection* was also detected with a balanced accuracy of 0.67 ± 0.06 using MiniRocket. Correlated *t*-tests showed substantial evidence that MiniRocket outperformed the other classifiers in detecting this behavior. *Sideways head lunge* was detected with a balanced accuracy of 0.65 ± 0.06 using InceptionTime. Based on correlated *t*-tests, there is substantial evidence that InceptionTime performed better than all other classifiers for detecting *Sideways head lunge. Pawing* was detected with a balanced accuracy of 0.57 ± 0.05 by MiniRocket. Correlated *t*-tests showed substantial evidence that MiniRocket

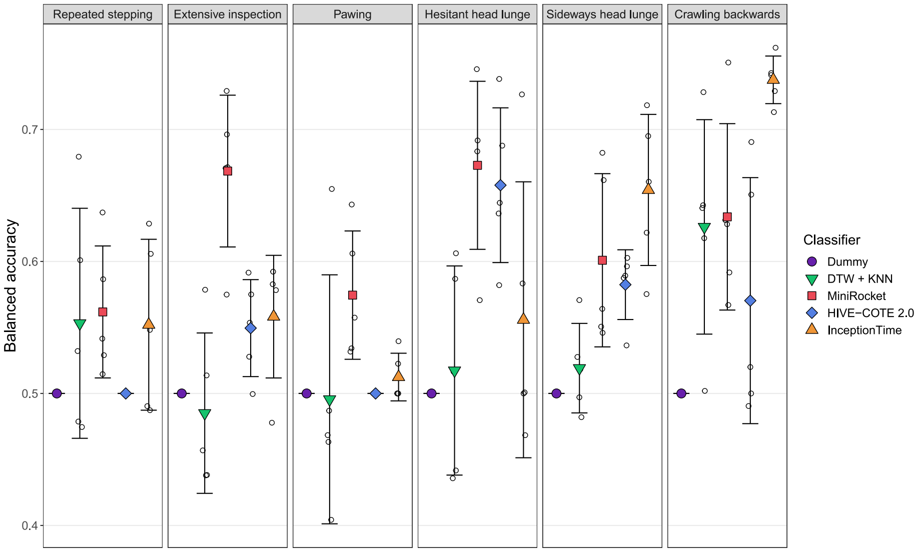

Fig. 4. Balanced accuracies (mean ± SD across outer cross-validation folds) in detecting the six atypical behaviors performed by dairy cows during lying down and standing up events with the different classifiers. Open circles show performance observed for the individual outer cross-validation folds. DTW + KNN = K-Nearest Neighbours model with Dynamic Time Warping as distance measure.

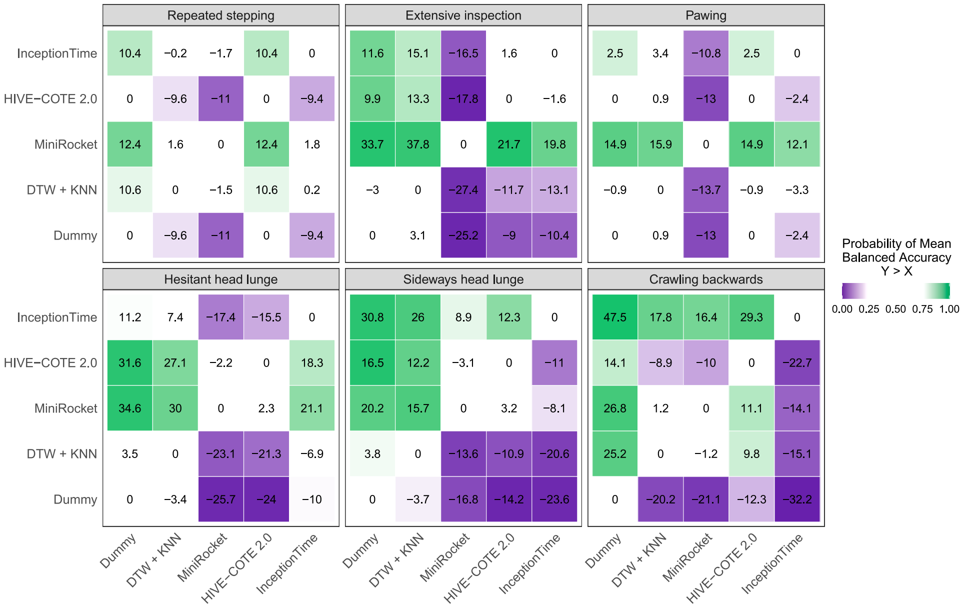

Fig. 5. Bayesian correlated *t*-tests for comparison of classifier performance shown in Fig. 4 (i.e. comparison of the means across the outer cross-validation folds). The colour gradient shows the probability that the classifier on the Y-axis outperforms the classifier on the X-axis. Probabilities corresponding to substantial evidence for true under- and outperformance (Bayes factor $> 10^{1/2}$) are coloured purple and green, respectively. Numbers represent the relative differences (%) in classifier performance with respect to X. DTW + KNN = K-Nearest Neighbours model with Dynamic Time Warping as distance measure.

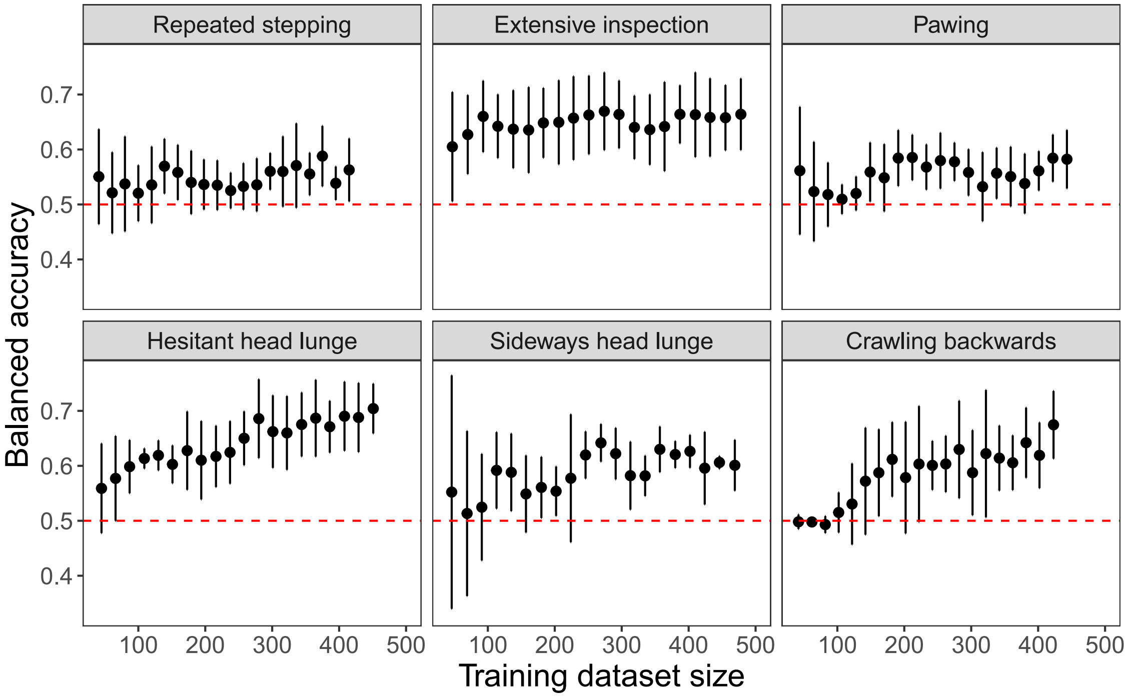

Fig. 6. Dependency of performance of MiniRocket models (mean ± SD across cross-validation folds) on training dataset size. Red dashed line indicates performance of the Dummy classifier.

performed better than all other classifiers for detecting this behavior. Lastly, *Repeated stepping* was the most poorly detected of all behaviors, with a balanced accuracy of 0.56 ± 0.05 achieved with MiniRocket. Correlated *t*-tests showed no substantial evidence for a true performance difference between DTW + KNN, MiniRocket, and InceptionTime for *Repeated stepping*. For both *Repeated stepping* and *Pawing*, the performance of all five outer CV folds of the HIVE COTE 2.0 models was the same as that of the Dummy classifier.

3.2. *Effect of dataset size on performance*

The dependency of performance of MiniRocket models on training dataset size is shown for each atypical behavior in Fig. 6. For *Hesitant head lunge* and *Crawling backwards*, performance increased with increasing training dataset size up to the maximum available dataset size. For *Sideways head lunge*, performance generally increased, but this increase was more erratic. For *Extensive inspection, Pawing*, and *Repeated stepping*, there was little performance increase from using 10% of the dataset to using all available training data.

3.3. *Effect of accelerometer sampling frequency on performance*

Dependency of performance of MiniRocket models on the accelerometer sampling frequency is shown in Fig. 7. The results of correlated *t*-

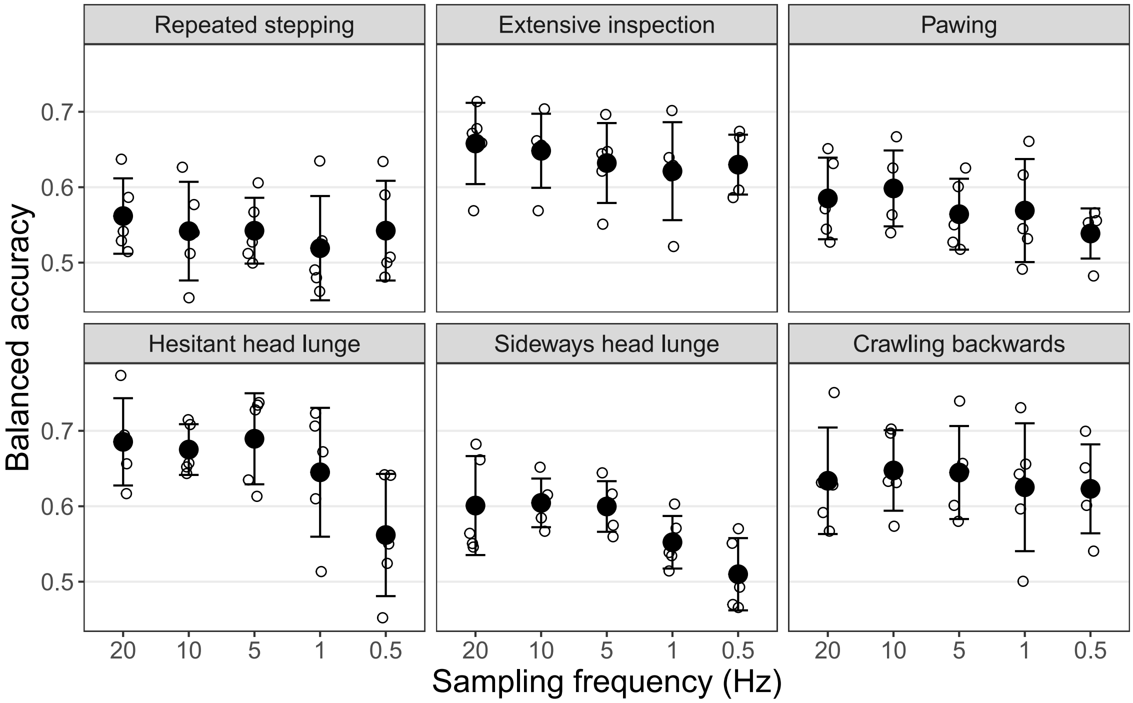

Fig. 7. Balanced accuracies (mean ± SD across cross-validation folds) of MiniRocket models trained on accelerometer data with different sampling frequencies. Open circles show the performance observed for the individual cross-validation folds.

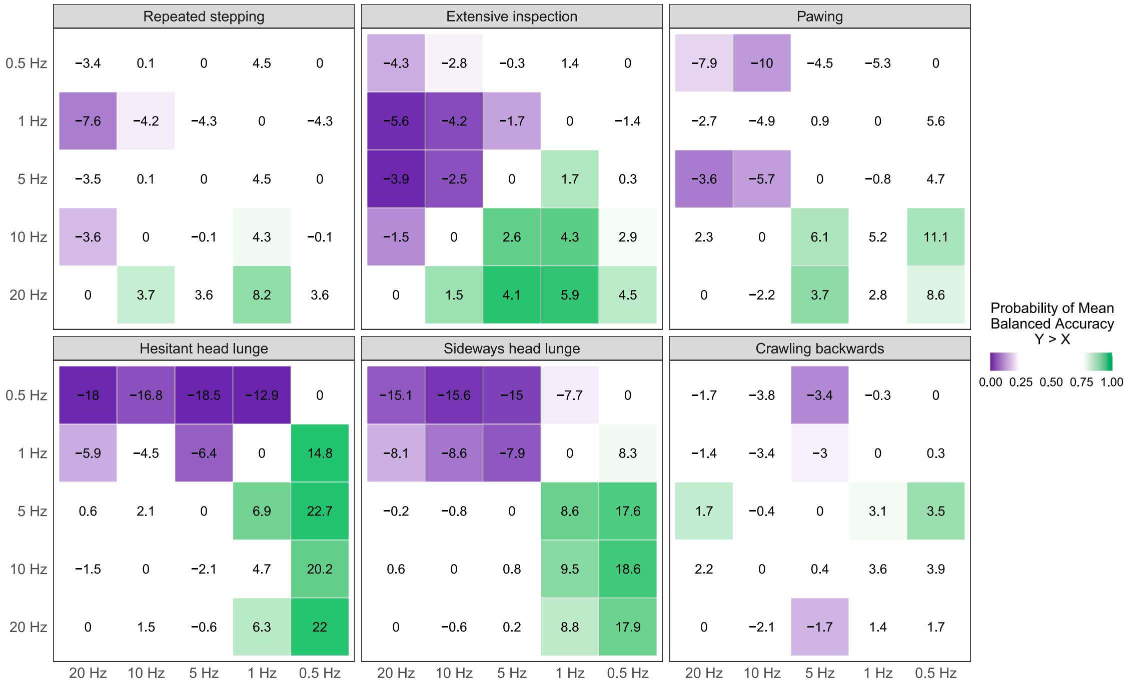

Fig. 8. Bayesian correlated *t*-test comparing performance of MiniRocket models trained on accelerometer data with different sampling frequencies shown in Fig. 7 (i.e. comparison of the means across the cross-validation folds). The colour gradient shows the probability that performances for the frequency on the Y-axis were higher than for the frequency on the X-axis. Probabilities corresponding to substantial evidence (Bayes factor > $10^{1/2}$) for true under- and outperformance are coloured purple and green, respectively. Numbers represent the relative differences (%) in classifier performance with respect to X.

tests comparing the model performances are shown in Fig. 8, together with the relative differences in performance. In general, lower performance was observed for lower accelerometer sampling frequency. However, correlated *t*-tests showed only for *Repeated stepping* and *Extensive inspection* substantial evidence for a true performance decrease when using data sampled at 10 Hz as compared with 20 Hz. For *Hesitant head lunge* and *Sideways head lunge*, no substantial evidence for a true performance decrease was found even of when using a sampling frequency of as low as 5 Hz (Fig. 8).

4. Discussion

The aim of this research was to investigate which atypical behaviors performed by dairy cows during lying down and standing up are promising candidates to be detected using body-mounted accelerometers and machine learning. A two-step approach was used: lying down and standing up events were first cut out of the raw accelerometer data using a reliable rule-based approach. Subsequently, the presence or absence of atypical behaviors was classified using machine learning models based on recent MTSC algorithms.

Achieved balanced accuracies for the detection of the atypical lying down and standing up behaviors ranged from 0.56 ± 0.05 to 0.74 ± 0.02. Best performances were obtained using the MiniRocket time series transform and the deep learning algorithm InceptionTime, which both have not previously been applied to classify time series data from sensors worn by dairy cows. The obtained performances are not yet satisfactory for application in the evaluation of new housing installations with regard to cow welfare. Nonetheless, *Hesitant head lunge* and *Crawling backwards*, appear to be the most promising candidates for accelerometer-based detection. These atypical standing up behaviors were detected with balanced accuracies of 0.67 and 0.74, respectively, and their learning curves indicated that more training data could further increase this performance (Fig. 6).

The two atypical behaviors identified as promising for accelerometer-based detection have the most direct welfare implications for dairy cows. *Hesitant head lunges* are indicators of inadequate head lunge space, because the available lunge space and the shape of the cubicle partition primarily determine how cows perform the head lunge [4]. Inability to use the head properly as counterweight directly affects animal welfare by limiting the cow's ability to rise up [5]. *Crawling backwards* occurs when cows lie to far forwards in cubicles. It exerts great force on the knee joints, which can result in discomfort and leg injuries [4,54]. Because this behavior is usually caused by inadequate stall dimensions or poor placement of the neck rail, it is highly relevant to farmers and producers of dairy cow housing installations [55].

The prevalence of the poorly detected behaviors, *Repeated stepping* and *Pawing*, is not only related to the design and configuration of fixed cubicle elements, but also to the quality of the lying place. Moreover, these behaviors are mainly indicators of hesitance to lie down and are even occasionally observed on pasture [4,8,56]. Although hesitation does not directly cause physical discomfort or injuries, it could lead to cows lying down less often and for less time [4]. Therefore, it is still highly relevant to assess indicators of hesitation with regard to dairy cow welfare. *Extensive inspection*, also indicative of hesitance [57], was better detected than *Pawing* and *Repeated stepping* and could be a more suitable behavior to assess hesitation before lying down.

The generally not satisfactory detection performances raise the question whether behaviors as defined in an ethogram designed for human observers are suitable to be detected by machine learning algorithms from accelerometer or other sensor data. For example, *Pawing* is described as pawing the ground with a front leg before the lying down movement [7]. However, this behavior was also labelled as present when the ground was pawed multiple times within one lying down event, possibly with both front legs. Even though one and multiple occurrences performed with either of the front legs are all evident cases of *Pawing* to a human observer, the variety in actual movements performed by the cow might have prevented the machine learning algorithm from effectively learning the generalizable patterns related to the target behavior. Surprisingly, *Sideways head lunge* was not among the best detected behaviors. Here, superior model performance was expected compared to detection models for other, seemingly more complex behaviors, such as *Extensive inspection*. This may be due to the lack of a clear boundary between a straight neck and a slightly bent neck in the ethogram (Table 2), which compromised label quality. However, *Sideways head lunge* was better detected when the model was trained on three classes specifying the actual direction of the head movement than when trained on the binary labels (see Table 2). This illustrates that redefining parts of the ethogram may help to increase detection performance for certain atypical behaviors.

In addition, it was occasionally impossible for the human observer to determine the presence or absence of atypical lying down and standing up behaviors beyond doubt. For example, cows occasionally performed multiple head lunge attempts within one standing up event, some of which straight and some directed sideways, leaving the observer puzzled how to record one value for the event. Ambiguous cases like these introduce a degree of error in the labels. Again, redefining the ethogram might alleviate the problem – classifying behaviors in more detail, for example with additional categories, could improve the quality of the labels and model performance.

Differences in class imbalance might partly explain why some atypical behaviors were better detected than others. The least well detected behaviors, *Repeated stepping* and *Pawing*, were the least often performed ones, with class distributions being around 1:10 (presence: absence, Table 2). This imbalance in combination with a limited amount of data leaves only few instances of the presence class for the machine-learning models to learn generalizable patterns in the data related to the behavior. The random oversampling used in the study (Fig. 2) does not discard any potentially useful information from the already limited dataset (as opposed to random undersampling). However, as it balances the class distribution by duplicating instances from the minority class, it may lead to overfitting if there are very few observations of the minority class [58].

Additionally, there were large inter-individual differences in class balances between cows. For example, *Pawing* and *Repeated stepping* were never performed by approximately half of the cows included in the study, whereas other individuals performed these behaviors often (Supp. Fig. 1B). These differences could have caused the machine-learning models to learn patterns specific to the individual cows and not related to the behavior in a generalizable manner. Consistent with this rationale, for the better-detected behaviors related to the head lunge, inter-individual distribution was more favourable because at least 75% of all cows performed these behaviors at least once. However, the best-detected behavior overall, *Crawling backwards*, had the largest inter-individual imbalances (Supp. Fig. 1B).

Substantially more data and employing random undersampling to balance class distributions, potentially even per cow, could improve detection performance. A sound strategy would be to collect more data on different farms (with different cubicle partitions and dimensions), where the atypical behaviors may be more common than in the research barn in this study.

In the further development of this method, the sampling frequency of the sensors could likely be reduced, saving battery life and storage capacity, without substantially sacrificing model performance. Subsampling the time series data from 20 to 5 Hz did not decrease the performance of MiniRocket models for the identified most promising behaviors, *Hesitant head lunge* and *Crawling backwards*. Moreover, for behaviors where lowering the sampling frequency to 10 Hz negatively affected performance, it decreased only by 4% as compared with 20 Hz.

5. Conclusion

A novel method was investigated to detect atypical lying down and standing up behaviours in dairy cows using accelerometers and machine learning. Overall, achieved detection performances for the atypical lying down and standing up behaviours were not yet satisfactory for application in the evaluation of new housing installations with regard to cow welfare. However, two behaviours associated with a hindered standing up movement were identified as promising candidates for accelerometer-based detection using machine learning. *Hesitant head lunge* and *Crawling backwards* were detected by balanced accuracies of 0.67 and 0.74, respectively, and their learning curves indicated that more training data might further increase model performance.

Therefore, these behaviours should be considered in the further development of an accelerometer-based method to assess standing up behaviours of dairy cows. The generally rather poor detection performance of atypical lying down behaviours might indicate that behaviours, as described in an ethogram designed for human observers, might often not be suitable for detection by machine learning. Detection performances may be improved by adjusting the ethogram with machine learnability in mind. Issues with imbalanced class distributions and inter-individual differences could potentially be alleviated by collecting substantially more data in stables with different lying cubicles. When developed further, the proposed method could improve efficiency and promote objectivity in the evaluation procedure of dairy cow housing installations by complementing human observations.

Data availability statement

The data presented in this study are available on request from the corresponding author.

CRediT authorship contribution statement

Stijn P. Brouwers: Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Michael Simmler:** Methodology, Software, Data curation, Writing – original draft, Writing – review & editing. **Pascal Savary:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Madeleine F. Scriba:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgments

This publication has emanated from research conducted with the financial support of the Federal Food Safety and Veterinary Office on behalf of the Swiss government (grant number: 2.21.01). The authors would like to thank Urs Marolf, Markus Keller, Matthias Hatt, Manon Delalande, Clémence Nanchen, and Diana Schönberger for help with mounting and removing the sensors. We would also like to thank Christoph Bühler and the farm staff of the Swiss Future Farm for assisting in the management of the study animals. Lastly, we are grateful to Roland Nasser for the interesting exchange on machine learning methods, and to Roxanne Berthel and Ralph Stoop for fruitful feedback that led to substantial improvements of the storyline of this publication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.atech.2023.100199.

References

[1] J. Fregonesi, M. Von Keyserlingk, C. Tucker, D. Veira, D Weary, Neck-rail position in the free stall affects standing behavior and udder and stall cleanliness, J. Dairy Sci. 92 (5) (2009) 1979–1985, https://doi.org/10.3168/jds.2008-1604.

[2] C.B. Tucker, D.M. Weary, D Fraser, Influence of neck-rail placement on free-stall preference, use, and cleanliness, J. Dairy Sci. 88 (8) (2005) 2730–2737, https://doi.org/10.3168/jds.S0022-0302(05)72952-0.

[3] F. Bernardi, J. Fregonesi, C. Winckler, D. Veira, M. Von Keyserlingk, D Weary, The stall-design paradox: neck rails increase lameness but improve udder and stall hygiene, J. Dairy Sci. 92 (7) (2009) 3074–3080, https://doi.org/10.3168/jds.2008-1166.

[4] L. Lidfors, The use of getting up and lying down movements in the evaluation of cattle environments, Vet. Res. Commun. 13 (4) (1989) 307–324, https://doi.org/10.1007/BF00420838.

[5] U. Schnitzer, Abliegen, Liegestellungen und Aufstehen beim Rind im Hinblick auf die Entwicklung von Stalleinrichtungen f..r Milchvieh, Kuratorium für Technik und Bauwesen in der Landwirtschaft 10 (1971) 43.

[6] S. Österman, I. Redbo, Effects of milking frequency on lying down and getting up behaviour in dairy cows, Appl. Anim. Behav. Sci. 70 (3) (2001) 167–176, https://doi.org/10.1016/S0168-1591(00)00159-3.

[7] N. Dirksen, L. Gygax, I. Traulsen, B. Wechsler, J.B. Burla, Body size in relation to cubicle dimensions affects lying behavior and joint lesions in dairy cows, J. Dairy Sci. 103 (10) (2020) 9407–9417, https://doi.org/10.3168/jds.2019-16464.

[8] A. Zambelis, M. Gagnon-Barbin, J. St John, E. Vasseur, Development of scoring systems for abnormal rising and lying down by dairy cattle, and their relationship with other welfare outcome measures, Appl. Anim. Behav. Sci. 220 (2019), 104858, https://doi.org/10.1016/j.applanim.2019.104858.

[9] J. Blom, S. Konggaard, J. Larsson, K. Nielsen, A. Northeved, P. Solfjeld, Electronic recording of pressure exerted by cows against structures in free-stall housing, Appl. Anim. Behav. Sci. 13 (1–2) (1984) 41–46, https://doi.org/10.1016/0168-1591(84)90050-9.

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

[11] D. Weary, I. Taszkun, Hock lesions and free-stall design, J. Dairy Sci. 83 (4) (2000) 697–702, https://doi.org/10.3168/jds.S0022-0302(00)74931-9.

[12] L. Green, V. Hedges, Y. Schukken, R. Blowey, A. Packington, The impact of clinical lameness on the milk yield of dairy cows, J. Dairy Sci. 85 (9) (2002) 2250–2256, https://doi.org/10.3168/jds.S0022-0302(02)74304-X.

[13] H.R. Whay, J.K. Shearer, The impact of lameness on welfare of the dairy cow, Vet. Clin. Food Anim. Pract. 33 (2) (2017) 153–164, https://doi.org/10.1016/j.cvfa.2017.02.008.

[14] P.A. Oltenacu, D.M. Broom, The impact of genetic selection for increased milk yield on the welfare of dairy cows, Anim. Welfare 19 (1) (2010) 39–49.

[15] I. Veissier, J. Capdeville, E. Delval, Cubicle housing systems for cattle: comfort of dairy cows depends on cubicle adjustment, J. Anim. Sci. 82 (11) (2004) 3321–3337, https://doi.org/10.2527/2004.82113321x.

[16] E. Vasseur, Animal behavior and well-being symposium: optimizing outcome measures of welfare in dairy cattle assessment, J. Anim. Sci. 95 (3) (2017) 1365–1371, https://doi.org/10.2527/jas.2016.0880.

[17] P. Martiskainen, M. Järvinen, J.P. Skön, J. Tiirikainen, M. Kolehmainen, J. Mononen, Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines, Appl. Anim. Behav. Sci. 119 (1–2) (2009) 32–38, https://doi.org/10.1016/j.applanim.2009.03.005.

[18] L. Riaboff, S. Poggi, A. Madouasse, S. Couvreur, S. Aubin, N. Bédère, E. Goumand, A. Chauvin, G. Plantier, Development of a methodological framework for a robust prediction of the main behaviours of dairy cows using a combination of machine learning algorithms on accelerometer data, Comput. Electron. Agric. 169 (2020), 105179, https://doi.org/10.1016/j.compag.2019.105179.

[19] J.A.V. Diosdado, Z.E. Barker, H.R. Hodges, J.R. Amory, D.P. Croft, N.J. Bell, E. A. Codling, Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system, Anim. Biotelem. 3 (1) (2015) 1–14, https://doi.org/10.1186/s40317-015-0045-8.

[20] G. Finney, A. Gordon, G. Scoley, S. Morrison, Validating the IceRobotics IceQube tri-axial accelerometer for measuring daily lying duration in dairy calves, Livest. Sci. 214 (2018) 83–87, https://doi.org/10.1016/j.livsci.2018.05.014.

[21] L. Schmeling, G. Elmamooz, P.T. Hoang, A. Kozar, D. Nicklas, M. Sünkel, S. Thurner, E. Rauch, Training and validating a machine learning model for the sensor-based monitoring of lying behavior in dairy cows on pasture and in the barn, Animals 11 (9) (2021) 2660, https://doi.org/10.3390/ani11092660.

[22] M.W. Iqbal, I. Draganova, P.C. Morel, S.T. Morris, Validation of an accelerometer sensor-based collar for monitoring grazing and rumination behaviours in grazing dairy cows, Animals 11 (9) (2021) 2724, https://doi.org/10.3390/ani11092724.

[23] A.A. Rayas-Amor, E. Morales-Almaráz, G. Licona-Velázquez, R. Vieyra-Alberto, A. García-Martínez, C.G. Martínez-García, R.G. Cruz-Monterrosa, G.C. Miranda-de la Lama, Triaxial accelerometers for recording grazing and ruminating time in dairy cows: an alternative to visual observations, J. Vet. Behav. 20 (2017) 102–108, https://doi.org/10.1016/j.jveb.2017.04.003.

[24] S. Reiter, G. Sattlecker, L. Lidauer, F. Kickinger, M. Öhlschuster, W. Auer, V. Schweinzer, D. Klein-Jöbstl, M. Drillich, M. Iwersen, Evaluation of an ear-tag-based accelerometer for monitoring rumination in dairy cows, J. Dairy Sci. 101 (4) (2018) 3398–3411, https://doi.org/10.3168/jds.2017-12686.

[25] J. Haladjian, J. Haug, S. Nüske, B. Bruegge, A wearable sensor system for lameness detection in dairy cattle, Multimodal Technol. Interact. 2 (2) (2018) 27, https://doi.org/10.3390/mti2020027.

[26] V.M. Thorup, L. Munksgaard, P.E. Robert, H. Erhard, P. Thomsen, N. Friggens, Lameness detection via leg-mounted accelerometers on dairy cows on four commercial farms, Animal 9 (10) (2015) 1704–1712, https://doi.org/10.1017/S1751731115000890.

[27] D. Lovarelli, J. Bacenetti, M. Guarino, A review on dairy cattle farming: is precision livestock farming the compromise for an environmental, economic and social sustainable production? J. Clean. Prod. 262 (2020), 121409 https://doi.org/10.1016/j.jclepro.2020.121409.

[28] B. Wechsler, An authorisation procedure for mass-produced farm animal housing systems with regard to animal welfare, Livest. Prod. Sci. 94 (1–2) (2005) 71–79, https://doi.org/10.1016/j.livprodsci.2004.11.034.

[29] S. Hendriks, C. Phyn, J. Huzzey, K. Mueller, S. Turner, D. Donaghy, J. Roche, Graduate student literature review: evaluating the appropriate use of wearable accelerometers in research to monitor lying behaviors of dairy cows, J. Dairy Sci. 103 (12) (2020) 12140–12157, https://doi.org/10.3168/jds.2019-17887.

[30] E.P. Doheny, C. Walsh, T. Foran, B.R. Greene, C.W. Fan, C. Cunningham, R. A. Kenny, Falls classification using tri-axial accelerometers during the five-times-sit-to-stand test, Gait Posture 38 (4) (2013) 1021–1025, https://doi.org/10.1016/j.gaitpost.2013.05.013.

[31] M. Lipperts, S. van Laarhoven, R. Senden, I. Heyligers, B. Grimm, Clinical validation of a body-fixed 3D accelerometer and algorithm for activity monitoring in orthopaedic patients, J. Orthop. Translat. 11 (2017) 19–29, https://doi.org/10.1016/j.jot.2017.02.003.

[32] R. Van Lummel, E. Ainsworth, U. Lindemann, W. Zijlstra, L. Chiari, P. Van Campen, J. Hausdorff, Automated approach for quantifying the repeated sit-to-stand using one body fixed sensor in young and older adults, Gait Posture 38 (1) (2013) 153–156, https://doi.org/10.1016/j.gaitpost.2012.10.008.

[33] N. O'Leary, D. Byrne, A. O'Connor, L. Shalloo, Invited review: cattle lameness detection with accelerometers, J. Dairy Sci. 103 (5) (2020) 3895–3911, https://doi.org/10.3168/jds.2019-17123.

[34] L. Riaboff, L. Shalloo, A. Smeaton, S. Couvreur, A. Madouasse, M. Keane, Predicting livestock behaviour using accelerometers: a systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data, Comput. Electron. Agric. 192 (2022), 106610, https://doi.org/10.1016/j.compag.2021.106610.

[35] A. Bagnall, J. Lines, A. Bostrom, J. Large, E. Keogh, The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances, Data Min. Knowl. Discov. 31 (3) (2017) 606–660, https://doi.org/10.1007/s10618-016-0483-9.

[36] A.P. Ruiz, M. Flynn, J. Large, M. Middlehurst, A. Bagnall, The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances, Data Min. Knowl. Discov. 35 (2) (2021) 401–449, https://doi.org/10.1007/s10618-020-00727-3.

[37] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2022. http://www.r-project.org.

[38] M. Simmler, S.P. Brouwers, triact package for R: analyzing the lying behavior of cows from accelerometer data (2023), https://cran.r-project.org/web/packages/triact.

[39] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174, https://doi.org/10.2307/2529310.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[41] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (1) (2017) 559–563.

[42] M. Löning, A. Bagnall, S. Ganesh, V. Kazakov, J. Lines, F.J. Király, sktime: a unified interface for machine learning with time series. arXiv:1909.07872 (2019), https://doi.org/10.48550/arXiv.1909.07872.

[43] A. Dempster, D.F. Schmidt, G.I. Webb, Minirocket: a very fast (almost) deterministic transform for time series classification, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021, pp. 248–257, https://doi.org/10.1145/3447548.3467231.

[44] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, A. Bagnall, HIVE-COTE 2.0: a new meta ensemble for time series classification, Mach. Learn. 110 (11) (2021) 3211–3243, https://doi.org/10.1007/s10994-021-06057-9.

[45] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D.F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.A. Muller, F. Petitjean, Inceptiontime: finding alexnet for time series classification, Data Min. Knowl. Discov. 34 (6) (2020) 1936–1962, https://doi.org/10.1007/s10618-020-00710-y.

[46] J. Wasselius, E.L. Finn, E. Persson, P. Ericson, C. Brogårdh, A.G. Lindgren, T. Ullberg, K. Åström, Detection of unilateral arm paresis after stroke by wearable accelerometers and machine learning, Sensors 21 (23) (2021) 7784, https://doi.org/10.3390/s21237784.

[47] D.J. Berndt, J. Clifford, in: Using dynamic time warping to find patterns in time series, 1994, pp. 359–370. https://dl.acm.org/doi/10.5555/3000850.3000887.

[48] J. Wainer, G.C. Cawley, Nested cross-validation when selecting classifiers is overzealous for most practical applications, Expert Syst. Appl. 182 (2021), 115222, https://doi.org/10.1016/j.eswa.2021.115222.

[49] G.C. Cawley, N.L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107.

[50] M.Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, arXiv:2008.05756 (2020), https://doi.org/10.48550/arXiv.2008.05756.

[51] A. Benavoli, G. Corani, J. Demšar, M. Zaffalon, Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis, J. Mach. Learn. Res. 18 (1) (2017) 2653–2688, https://doi.org/10.48550/arXiv.1606.04316.

[52] H. Jeffreys, The Theory of Probability, OUP Oxford, 1998.

[53] R.E. Kass, A.E. Raftery, Bayes factors, J. Am. Stat. Assoc. 90 (430) (1995) 773–795, https://doi.org/10.1080/01621459.1995.10476572.

[54] B. Wechsler, J. Schaub, K. Friedli, R. Hauser, Behaviour and leg injuries in dairy cows kept in cubicle systems with straw bedding or soft lying mats, Appl. Anim. Behav. Sci. 69 (3) (2000) 189–197, https://doi.org/10.1016/S0168-1591(00)00134-9.

[55] H. Hoffman, M. Rist, Tiergerechte und arbeitswirtschaftlich gunstige Anbindevorrichtungen fur Kuhe, Schweiz. Landwirtsch. Monatshefe 53 (1975) 119–126.

[56] S.P. Brouwers, M.F. Scriba, P. Savary, Assessment of lying down and standing up movements of dairy cows on pasture and in free-stall cubicles, Kuratorium für Technik und Bauwesen in der Landwirtschaft 524 (2022) 54–56, https://doi.org/10.13140/RG.2.2.33207.39847.

[57] C.C. Krohn, L. Munksgaard, Behaviour of dairy cows kept in extensive (loose housing/pasture) or intensive (tie stall) environments II. Lying and lying-down behaviour, Appl. Anim. Behav. Sci. 37 (1) (1993) 1–16, https://doi.org/10.1016/0168-1591(93)90066-X.

[58] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data Min. Knowl. Discov. 28 (2014) 92–122, https://doi.org/10.1007/s10618-012-0295-5.