# Estrus Detection from Ear-tag Accelerometery in Grazing Cattle

Reza Arablouei[a,*], Brendan Do[a], Neil Bagnall[b], Jody McNally[b], Greg Bishop-Hurley[b], Aaron Ingham[b]

[a] *Data61, CSIRO, Pullenvale QLD 4069, Australia*
[b] *Agriculture and Food, CSIRO, St Lucia QLD 4067, Australia*

**Abstract**

Timely and accurate detection of estrus is essential for improving cattle reproductive efficiency and herd management. We present a lightweight, data-driven pipeline that infers estrus from daily behavior patterns captured by ear-tag triaxial accelerometer data. At its core is a compact deep learning model that estimates per-animal daily behavior profiles from raw accelerometry. To account for individual and temporal variation, we compute individualized baseline profiles via causal dynamic trimmed means over preceding days and express features as deviations of daily behavior durations from these baselines. We fit a binary regularized logistic regression model to these features for final estrus detection. Evaluated on data from 33 Angus cows fitted with smart ear-tags during a four-month grazing trial in temperate Australia, the approach delivers high discriminative performance at an ultra-low computational footprint. The results confirm that subtle yet systematic behavioral deviations, measured by low-cost accelerometers, are reliable predictors of estrus, directly enabling scalable, infrastructure-light, and autonomous reproductive monitoring in extensive grazing systems. Crucially, the pipeline is compatible with on-tag or edge deployment and integrates readily with existing digital-agriculture workflows for alerting and decision support.

*Keywords:*
accelerometer data, animal behavior classification, deep learning, embedded systems, estrus detection.

## 1. Introduction

Understanding and quantifying livestock behavior is foundational to herd health, welfare, and productivity, yet continuous direct observation is impractical for large, spatially dispersed herds. This has accelerated the use of wearable sensing for automated, remote behavior recognition. Among available options, micro-electromechanical systems (MEMS) integrating inertial measurement unit (IMU) sensors, especially low-power triaxial accelerometers embedded in ear-tags, collars, or leg devices, offer a compelling balance of cost, robustness, and battery life. Over the past decade, these devices have enabled reliable behavior

*Corresponding author

classification across diverse production systems using signal-processing and machine-learning pipelines (Kamminga et al., 2018; Fogarty et al., 2020; Brandes et al., 2021; Pavlovic et al., 2021; Mao et al., 2023; Hlimi et al., 2024; Dunford et al., 2024; Russel and Selvaraj, 2024).

When primary behaviors (typically grazing, ruminating, resting, walking, and drinking) are recognized with high fidelity, the resulting time series become a rich substrate for higher-level inference. In practice, behavior profiles reveal characteristic motifs that precede or accompany important physiological events. For example, estrus (heat) is typically accompanied by surges in locomotion, mounting, and restlessness with short-term declines in grazing/eating and rumination (Roelofs et al., 2010; Reith and Hoy, 2018), whereas impending parturition (calving in cattle) is often preceded by more frequent shorter lying bouts, heightened restlessness, and shifts in feed intake (Borchers et al., 2017; Krieger et al., 2019). Mapping day-to-day behavioral baselines per animal and detecting departures from those baselines allows these latent events to be inferred from IMU streams without intrusive instrumentation, enabling timely decision support for reproduction and peripartum management (e.g., Edwards et al., 2020; Schweinzer et al., 2019; Schilkowsky et al., 2021; Mayo et al., 2019).

Among downstream applications, accurate estrus detection is especially consequential for precision agriculture. Timely and reliable identification of estrus increases submissions for artificial insemination, raises the likelihood of insemination within the optimal fertility window, and reduces days open (calving-to-conception interval), leading to higher reproductive efficiency and more predictable calving intervals. In addition, continuous detection supports embryo-transfer programs and the assessment of return to estrus after calving, a key fertility trait, by revealing delayed postpartum resumption and triggering targeted nutritional or health interventions (Rorie et al., 2002). At the system level, better fertility translates into fewer non-productive days, improved lifetime productivity, and a lower emissions intensity per unit of milk or meat, thereby helping to reduce the environmental footprint of ruminant production. Operationally, automated sensor-driven detection reduces labor demands for visual heat checks, supports targeted rather than blanket interventions, and enables data-driven breeding and culling decisions aligned with welfare and sustainability objectives (Schweinzer et al., 2019; Schilkowsky et al., 2021; Mayo et al., 2019).

Prompt and consistent detection of estrus remains challenging as modern high-producing cows often display fewer and briefer external signs of estrus, particularly under heat stress and certain housing or management conditions (Reith and Hoy, 2018; Lopez et al., 2004; Wiltbank et al., 2006), and because visual heat checks are labor-intensive and error-prone. This has motivated continuous, automated monitoring using wearable devices, particularly triaxial accelerometers mounted on the ear, neck, or leg, that can quantify activity surges, events of mounting or being mounted, and related motion motifs around estrus. The literature on estrus detection with IMU-based wearables broadly agrees that: (i) accelerometer-derived activity systems are effective and widely adopted; (ii) detection performance varies

with production context (parity, housing, health, management); and (iii) reported accuracy depends on the evaluation reference (visual observation, ultrasonography, or endocrine profiles) (Roelofs et al., 2010; Løvendahl and Chagunda, 2010; Reith and Hoy, 2018; Saint-Dizier and Chastant-Maillard, 2018; Cerri et al., 2021; Saint-Dizier and Chastant-Maillard, 2012).

Ear-tags containing IMUs, especially accelerometers, now feature prominently in both research and commercial deployments. Using ear-tags with triaxial accelerometers, Schweinzer et al. (2019) evaluated a commercial automated estrus-detection system in indoor-housed cows, with the system generating alerts from accelerometer-derived activity using proprietary algorithms. A larger evaluation of the same system showed that alerts, generated from activity and rumination, have characteristic timing and duration distributions and analyzed how cow-level factors (e.g., milk yield) affect detection performance (Schilkowsky et al., 2021). Complementing these studies, Perez Marquez et al. (2023) compared ear-tag accelerometer-derived behavioral predictors and ear temperature between the luteal phase and the falling-progesterone (declining $P_4$) phase preceding estrus, as defined by in-line milk progesterone. They then quantified the discriminative ability of individual predictors and their combinations. Although these studies mainly involve housed dairy herds, pasture-based work has also shown that accelerometer-based activity monitoring systems can detect estrus in grazing cows. Evidence includes field evaluations of collar-mounted accelerometers and comparisons with on-animal mount detectors or endocrine/ovulatory references in large seasonal, pasture-grazed herds (Kamphuis et al., 2012; Moore et al., 2021; Hockey et al., 2010). Collectively, these studies and other related ones (e.g., At-Taras and Spahr, 2001; Bikker et al., 2014; Weinert-Nelson et al., 2025) demonstrate the practicality of ear-tag IMUs for automated estrus detection in housed dairy cattle and motivate extension to pasture-based systems. They also point to design needs such as per-cow baselines and robust processing to handle sensor-placement-specific noise and context effects (e.g., milk yield, housing, heat stress).

Neck collars and leg pedometers provide long-standing baselines for automated estrus detection. In a synchronized, head-to-head benchmark of multiple commercial precision-dairy systems (including leg pedometers, leg and ear accelerometers, and a neck-mounted tag) Mayo et al. (2019) evaluated estrus detection by comparing device alerts with a hormonal reference (temporal plasma progesterone patterns) and reported comparative sensitivity, specificity, and accuracy useful for method comparisons and cost-benefit discussions. Earlier collar- and pedometer-based studies demonstrated that estrus can be detected using thresholded activity indices, either universal (herd-level) or cow-specific, though performance is highly sensitive to herd management and contextual factors like feeding and milking routines, flooring, and lameness (Løvendahl and Chagunda, 2010; Valenza et al., 2012). These insights motivate models that look beyond simple "activity spikes."

Recent studies emphasize the utility and significance of sensor fusion and data-driven (model-learning-based) approaches. In housed environments, combining accelerometers with

ultra-wideband (UWB) indoor localization improves estrus and calving detection and supports decision-relevant alert timing (Benaissa et al., 2020). In herds housed in free-stall barns, fusing location with IMU acceleration and applying supervised learning (e.g., KNN, LDA, CART, BPNN) enhances estrus discrimination relative to simple thresholded indices (Wang et al., 2020). A complementary, threshold-based approach fuses acceleration and location into a per-cow-normalized activity index that accounts for each cow's recent baseline and performs well without supervised learning (Wang et al., 2022). In these studies, data is sampled at 1 Hz and aggregated over 0.5-1.5 h windows (Wang et al., 2020) or hourly windows with per-cow normalization (Wang et al., 2022), choices that affect detection latency and the practical insemination window. These approaches have also been extended to grazing systems, as show in pasture-based trials (Kamphuis et al., 2012; Moore et al., 2021; Hockey et al., 2010).

Across this literature, the choice of reference labels is pivotal. Many studies synchronize estrus and define ground truth from intensive visual observation (standing heat) or ovulation timing by ultrasonography (Kamphuis et al., 2012; Hockey et al., 2010), while others align device alerts with endocrine references such as temporal plasma progesterone patterns or in-line milk progesterone (Mayo et al., 2019; Perez Marquez et al., 2023). Because these references capture different biological moments, reported sensitivity/specificity and the apparent lead-time of alerts can differ. Broader syntheses recommend making label definitions explicit, reporting performance at decision-relevant horizons (e.g., 8-24 h before planned artificial insemination), and linking activity phenotypes to physiology to explain between-cow variation in expression and fertility (Cerri et al., 2021). Although most validations are dairy-centric and in housed conditions, a recent beef-focused review catalogues available sensors (accelerometers, pedometers, rumen boluses, biosensors, and infrared thermography), and highlights the need for further validation in beef production systems that differ in management and environmental conditions from intensive housed dairy operations, points especially relevant for extensive pasture-based beef systems (Merkelytė et al., 2025).

In this study, we present a novel estrus-detection approach for grazing beef cattle using ear-tag accelerometer data. We first introduce a new lightweight deep-learning-based model that effectively maps triaxial acceleration to four core behaviors of grazing, ruminating, resting, and walking. After training the model on a behavior-annotated subset of ear-tag accelerometer data collected during a three-month-long trial with 33 heifers, we apply the learned model to the entire data and aggregate the predictions into daily behavior profiles for each animal. For each cow and behavior, we compute a robust rolling baseline as the 14-day trimmed mean of the preceding daily aggregate behavior predictions. We then derive predictive features as deviations from these baselines and fit a logistic regression classifier to the derived features to estimate the daily probability of estrus. The model's coefficients provide interpretable links between behavior changes and predicted estrus probability. We evaluate the developed approach against endocrine-derived ground truth from blood progesterone

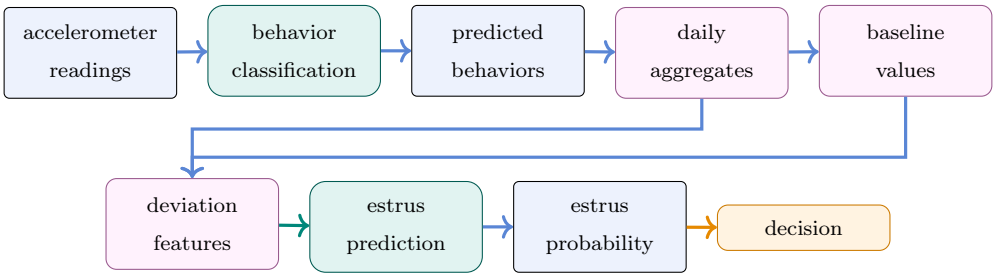



Figure 1: The proposed estrus detection pipeline.

($P_4$) assays, demonstrating an accurate and efficient solution that is readily deployable in extensive pasture-based systems. We summarize the end-to-end pipeline of the proposed approach in Figure 1.

## 2. Data

In this section, we describe the data-collection trial and the resultant dataset that we utilize to develop and evaluate our proposed estrus-detection approach, comprising ear-tag triaxial accelerometer data and manual behavior and estrus annotations. We ran the experiment at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Chiswick Research Station, Armidale NSW, Australia (30°36′28″S, 151°32′39″E; temperate climate) from January 30, 2025 to May 8, 2025 with 33 Angus cows. The CSIRO FD McMaster Laboratory Chiswick Animal Ethics Committee approved the experiment (authority number ARA 24/11).

We fitted all cattle with eGrazor collars[1] and CERES RANCHER ear-tags[2]. Figure 2 shows animals wearing both devices. Each collar carries a 32 GB micro-SD card, whereas each ear-tag contains 1 MB of onboard flash. Therefore, we streamed each ear-tag's accelerometer data packets to its paired collar over Bluetooth Low Energy (BLE) Generic Attribute Profile (GATT) and logged both collar and ear-tag accelerometer data on the collar's micro-SD card. At the end of the trial, we collected the cards and retrieved the logs.

We filmed cattle during daylight (typically 07:00-18:00 local time) to support behavior labeling. Coverage was not continuous and not all animals were visible at all times, hence we prioritized individuals with clear, unobstructed views. While reviewing the videos, trained annotators recorded behavior labels using a custom mobile application.

We align the video annotations to the accelerometer streams using shared UTC timestamps. We then segment the labeled accelerometer data into non-overlapping windows of consecutive 256 readings per axis (x, y, and z) to create fixed-length instances for training and evaluation of our behavior classification model. With an accelerometer sampling rate of 50 Hz, each window represents approximately 5.12 seconds of continuous animal behavior.

[1] https://www.csiro.au/research/animals/livestock/egrazor-measuring-cattle-pasture-intake
[2] https://www.cerestag.com/


Figure 2: Angus beef cattle wearing both collars and ear-tags.

Table 1: The number of labeled data instances for each behavior class.

| behavior | instances |
|---|---|
| grazing | 9,526 |
| ruminating | 4,363 |
| resting | 2,331 |
| walking | 526 |
| other | 464 |
| total | 17,210 |

We consider five mutually exclusive behavior classes: *grazing*, *ruminating*, *resting*, *walking*, and *other*, where *other* includes all behaviors not covered by the first four. Table 1 reports the number of labeled accelerometer data instances per class, each comprising 256 successive time steps and 3 axes ($256 \times 3$ samples per instance).

Throughout the trial, we fitted cows with tailhead-mounted Kamar Heatmount Detector devices[3] that signal standing heat (mounts), a hallmark of estrus. Figure 3 illustrates the placement of a Heatmount Detector on a cow's tailhead. Each time a detector flagged a mount, we collected a blood sample from that cow and assayed serum progesterone ($P_4$). We labeled the corresponding cow-day as estrus when $P_4 \leq 1\,\text{ng/mL}$ ($\approx 3.2\,\text{nmol/L}$). In total, we observed 121 confirmed estrus events with 55 having near-complete 24 h ear-tag accelerometer data and a reliable baseline with at least three day of data within the preceding 14 days. To assemble verified non-estrus days, we also collected mid-luteal (mid-cycle) blood samples and confirmed elevated $P_4$ (typically 30-100 nmol/L), yielding a final dataset of 927 non-estrus cow-days and 55 estrus cow-days. Blood was sampled into red-top serum tubes, and progesterone assays were performed by Vetnostics Pathology[4].

To complement naturally occurring estrus with hormone-induced episodes, on April 7, 2025, we administered an intramuscular $PGF_{2\alpha}$ analogue (cloprostenol) to all cows to induce luteolysis and synchronize estrus within approximately 48 h. Such synchronization practices are commonplace in commercial herds to align estrus with planned natural service or artificial insemination, thereby increasing submission and conception rates.

## 3. Behavior Classification

In this section, we present our deep-learning model for livestock behavior classification and analyze its efficiency and effectiveness. The model is specifically designed to process tri-axial accelerometer data in a lightweight, interpretable, and computationally efficient manner.

[3] `https://kamarinc.com/kamar-detectors/`

[4] `https://www.vetnostics.com.au/`


Figure 3: Kamar Heatmount Detector affixed to a cow's tailhead for detecting standing-to-be-mounted events indicative of estrus. The detector turns bright red after a mount, enabling easy identification during visual checks.

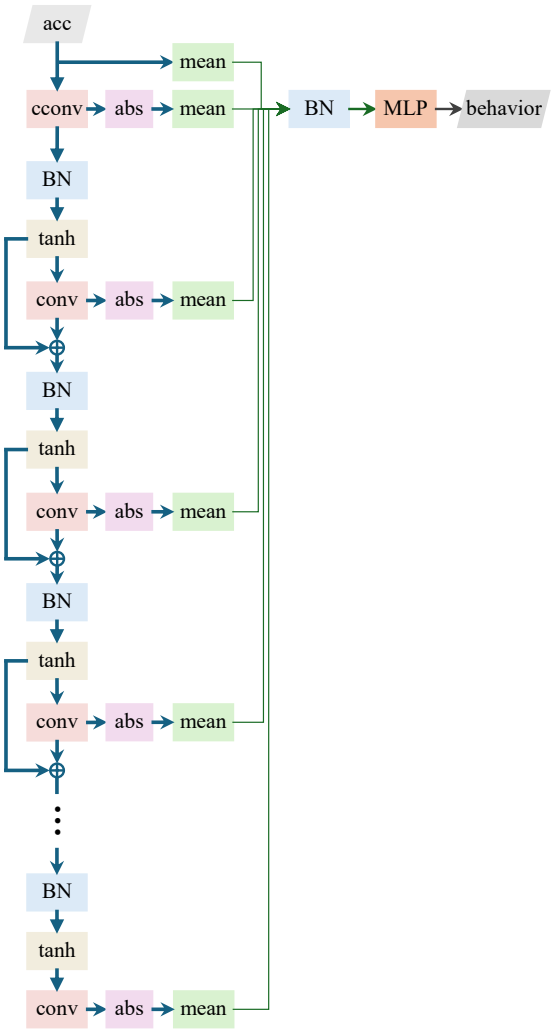



Figure 4: The proposed model architecture for classifying animal behavior using triaxial accelerometer data.

### *3.1. Model Architecture*

The proposed architecture combines depthwise temporal filtering, residual connections, and absolute-mean pooled features to balance accuracy, efficiency, and interpretability. We provide a schematic overview in Figure 4.

The network receives a stream of raw triaxial accelerometer readings and partitions them into contiguous windows of $L = 256$ samples per axis. The values in each window are treated as a data instance and are run through the network. The first three features are simply the per-axis means of the 256 values, representing the average sensor orientation over the window.

To suppress the constant gravitational component, the raw triaxial accelerometer readings are passed through a set of constrained convolutional layers (`cconv`), one or more per axis. During training, the kernel parameters of each filter in the `cconv` layer are constrained to sum to zero as proposed in (Arablouei et al., 2024). This nullifies the DC response of the `cconv` filters, thereby removing gravity's contribution. For each axis, we compute the mean absolute responses of these filters to form the next features.

The gravity-suppressed outputs are fed through a batch-normalization (`BN`) layer followed by tangent hyperbolic (`tanh`) activation. The post-activation values are then filtered by a depthwise convolutional layer (`conv`), whose output are used to calculate the next set of per-axis features by averaging their absolute values. Afterwords, the post-activation values are added to the convolution layer output to form the input to the next `BN-tanh-conv` block. The absolute values of the outputs of each block are averaged to calculated the next features and the input of each `conv` layer is added to its output as a skip connection as shown in Fig. 4. This residual design improves gradient flow, stabilizes training, and ensures that features are refined rather than overwritten. In addition, the modularity of the architecture offers flexibility, allowing the model to be adapted to tasks of varying complexity.

To maintain the zero-mean property introduced by the constrained filters, all subsequent convolutional layers have their biases fixed to zero, and the affine parameters of their following `BN` layers are also set to zero. The choice of tanh activation is consistent with this design, as it preserves the symmetry of the input.

Once all behavior-classification features are calculated, we concatenate them and apply another `BN` (with affine parameters enabled) before passing through a multiplayer perceptron (`MLP`) classifier with one hidden layer and ReLU nonlinearity to predict the animal behavior associated with the input accelerometer data.

The proposed architecture reflects the following principles:

- *Efficiency:* Depthwise convolutions and global pooling keep computational cost low, enabling embedded deployment.
- *Residual learning:* Skip connections improve gradient flow, thereby enhancing training stability and accelerating convergence.

- *Progressive factorization:* Successive `BN-tanh-conv` blocks tackle simpler learning subtasks as downstream blocks operate on more refined, higher-SNR, standardized representations produced upstream, easing optimization and improving data efficiency.
- *Interpretability:* Mean-absolute features provide an intuitive link to signal energy in different spectral bands.
- *Compactness:* Each feature serves as a compact per-channel descriptor, yielding a low-dimensional yet informative representation.
- *Flexibility:* The modular design allows straightforward adaptation to different problem settings or sensing modes.

### *3.2. Relation to Prior Work*

Deep-learning methods have been widely explored for accelerometer-based livestock behavior classification, frequently using recurrent neural networks (RNNs) (Peng et al., 2019; Wang et al., 2023), convolutional neural networks (CNNs) (Pavlovic et al., 2021; Bloch et al., 2023), or convolutional-recurrent hybrids (Arablouei et al., 2023b; Jin et al., 2024), with transformer-based variants appearing more recently (Eckhardt et al., 2025).

RNNs, specifically the most common long short-term memory (LSTM) and gated recurrent unit (GRU) variants, model temporal dynamics by propagating a hidden state across samples, which helps capture long-range dependencies in tri-axial sequences. Stacked and bidirectional variants further improve accuracy but increase latency and memory due to larger hidden states and backpropagation through time. While robust to variable-length windows and sampling-rate changes, their sequential nature limits parallelism and can complicate low-power, on-animal inference. Residual 1D CNNs adapted to tri-axial signals stack temporal convolutional blocks with skip connections, improving gradient flow and enabling multi-scale receptive fields. However, these models are often computationally heavy for embedded, animal-borne sensors and may reduce interpretability for ethological analyses. Hybrid CNN-RNN designs typically apply convolutional layers for local feature extraction and downsampling, followed by recurrent units to aggregate temporal context. These models achieve competitive accuracy but incur nontrivial latency and memory usage.

Overall, while accurate, general-purpose deep models repurposed for livestock behavior classification often incur substantial parameter counts and large intermediate-activation footprints, hindering deployment on animal-borne devices (e.g., ear-tags and collars) where power and compute are tightly constrained. Consequently, recent work underscores bespoke, resource-aware designs that enable *in-situ* inference (Arablouei et al., 2021, 2023a). To meet these embedded constraints, lightweight sensor-based recognition commonly leverages depthwise/separable convolutions and global/temporal pooling, or pairs compact handcrafted/statistical descriptors with shallow heads (Howard et al., 2017; Chollet, 2017; Ignatov, 2018). Field deployments further highlight the need for compact, interpretable mod-

els that remain robust to attachment variability and environmental noise, particularly for on-collar and ear-tag systems. Complementary strategies, including multimodal fusion with GNSS and cross-device model learning, improve robustness, behavioral resolution, and transfer across heterogeneous tag hardware (Arablouei et al., 2023c, 2024).

The proposed animal behavior classifier aligns with embedded requirements by combining (i) constrained temporal filters whose kernel weights sum to zero to suppress the gravitational (DC) component Arablouei et al. (2024), (ii) depthwise temporal convolutions with residual connections for efficient multi-scale representation, and (iii) block-wise global $\ell_1$ temporal pooling (mean of absolute values across time) that yields compact, per-channel descriptors. Replacing recurrent stacks with pooled descriptors and a shallow MLP head reduces both memory footprint and inference latency while preserving the benefits of residual learning and multi-scale feature extraction. This design is well suited to long-term, battery-powered, on-animal deployments.

### *3.3. Implementation Details*

We implement the model in PyTorch and train it end-to-end on the labeled accelerometer dataset described in Section 2. For on-device inference, we re-implement the network in C using the CMSIS-NN and CMSIS-DSP libraries.

In this work, we use $N_b = 3$ `BN-tanh-conv` blocks and $N_f = 7$ filters per axis across $N_a = 3$ axes, yielding $N_c = N_a N_f = 21$ depthwise-convolution channels and a total of $N_t = N_a + (1 + N_b)\, N_c = 87$ scalar features. The feature vector comprises the three per-axis means plus mean-absolute descriptors taken after the constrained-filter stage (`cconv`) and after each of the $N_b$ residual blocks. Each temporal filter uses an odd kernel size $K = 9$ to preserve symmetric context and residual additions use centered cropping to align valid-length outputs. Convolution layers are bias-free and use valid (no-padding) convolutions. We initialize the `cconv` filters with discrete cosine transform (DCT) basis functions of appropriate orders (matched to the kernel length), following (Arablouei et al., 2024). The MLP head includes a single hidden layer with $N_h = 46$ neurons.

To train the model, we run the AdamW optimization algorithm with learning rate $10^{-4}$ and weight decay $10^{-3}$ for $10^4$ iterations on randomly drawn batches of 1024 windows, optimizing the cross-entropy loss augmented with a confidence penalty ($\lambda = 1$) as in (Arablouei et al., 2024).

In the CMSIS-C inference build, we fold `BN` applied to the feature vector into the first `MLP` layer, eliminating a standalone `BN` operation and its parameters. For `BN` layers that precede the `tanh-conv` blocks, we retain only per-channel running statistics, specifically, mean and inverse standard deviation and no batch counter. We implement `tanh` via a single 257-entry lookup table with linear interpolation (256 intervals), requiring no separate slope table.

Table 2: Parameter counts and FP32 memory (bytes) for each component of the proposed model in the CMSIS-C inference build.

| component | parameters | bytes |
|---|---|---|
| depthwise convolution weights (4 layers) | $4 \times (21 \times 9) = 756$ | 3,024 |
| `BN` running statistics (3 layers) | $3 \times (2 \times 21) = 126$ | 504 |
| 1st `MLP` layer weights and biases ($87 \to 46$) | $87 \times 46 + 46 = 4{,}048$ | 16,192 |
| 2nd `MLP` layer weights and biases ($46 \to 5$) | $46 \times 5 + 5 = 235$ | 940 |
| `tanh` lookup table (257 entries) | 257 | 1,028 |
| total | 5,422 | 21,688 |

### *3.4. Complexity Analysis*

In Table 2, we summarize the number of inference-time parameters for each component of the proposed model. The inference artifact comprises 5,422 single-precision floating-point (FP32) scalars in total, corresponding to a memory footprint of 21,688 bytes.

In Table 3, we list the floating-point operations (FLOPs) for each stage of the proposed model during inference. The FLOP counts reflect valid (no-padding) convolutions, include per-axis means, depthwise convolutions at each stage, mean-absolute pooling, three non-affine `BN` layers, `tanh` calculation via a lookup table with linear interpolation, residual additions, and the two-layer `MLP` head (with its preceding `BN` fused). The resulting inference cost is 486,566 FLOPs per triaxial 256-sample window, with most of the work lying in simple multiply-accumulate operations within the depthwise convolutions.

Depthwise convolutions and block-wise global $\ell_1$ temporal pooling keep both the parameter count and FLOPs low. On a 64 MHz ARM Cortex-M4F microcontroller, specifically the nRF52840 used in CERES RANCHER ear-tags, the end-to-end inference latency for a triaxial 256-sample window typically falls in the 25-35 ms range when all model parameters listed in Table 2 reside in SRAM and kernels stream per channel. The static (flash) memory footprint is $\sim 21.7$ KB, and the peak RAM working set for activations in a streaming, in-place implementation is $\sim 44$ KB, yielding an overall on-chip footprint of $\sim 66$ KB. Q15/Q7 quantization with CMSIS-NN can provide a further $1.5-2\times$ speedup. Combined with interpretable block-wise pooled features, this efficiency supports continuous, long-duration field monitoring under tight power, bandwidth, and storage constraints.

### *3.5. Evaluation*

We evaluate the proposed behavior classification model on the labeled dataset described in Section 2 using five-fold stratified cross-validation (CV) without shuffling. In each split, we train on four folds and test on the held-out fold. Stratification keeps class proportions consistent across train-test splits, which stabilizes estimates and reduces fold-to-fold variance, especially given the severe class imbalance in our dataset. We avoid shuffling to respect temporal autocorrelation in the accelerometer stream and to reduce leakage from adjacent windows, yielding more realistic generalization estimates.

Table 3: Computational cost (in terms of FLOPs) at inference time for each stage of the proposed model.

| stage | description | channels | window | FLOPs |
|---|---|---|---|---|
| mean pooling | global average over time (per axis) | 3 | 256 | 768 |
| `cconv` | 1D depthwise convolution (no bias or padding) | 21 | 248 | 88,536 |
| mean-abs pooling | global average of absolute values (per channel) | 21 | 248 | 10,416 |
| `BN` (1st block) | batch norm (no learnable scale or shift parameter) | 21 | 248 | 10,416 |
| `tanh` (1st block) | activation (lookup table and linear interpolation) | 21 | 248 | 20,832 |
| `conv` (1st block) | 1D depthwise convolution (no bias or padding) | 21 | 240 | 85,680 |
| mean-abs pooling | global average of absolute values (per channel) | 21 | 240 | 10,080 |
| skip connection | residual addition (sequences matched via cropping) | 21 | 240 | 5,040 |
| `BN` (2nd block) | batch norm (no learnable scale or shift parameter) | 21 | 240 | 10,080 |
| `tanh` (2nd block) | activation (lookup table and linear interpolation) | 21 | 240 | 20,160 |
| `conv` (2nd block) | 1D depthwise convolution (no bias or padding) | 21 | 232 | 82,824 |
| mean-abs pooling | global average of absolute values (per channel) | 21 | 232 | 9,744 |
| skip connection | residual addition (sequences matched via cropping) | 21 | 232 | 4,872 |
| `BN` (3rd block) | batch norm (no learnable scale or shift parameter) | 21 | 232 | 9,744 |
| `tanh` (3rd block) | activation (lookup table and linear interpolation) | 21 | 232 | 19,488 |
| `conv` (3rd block) | 1D depthwise convolution (no bias or padding) | 21 | 224 | 79,968 |
| mean-abs pooling | global average of absolute values (per channel) | 21 | 224 | 9,408 |
| 1st `MLP` layer | feedforward neural network ($87 \to 46$) | 87 | – | 8,004 |
| ReLU | rectified linear unit activation | 46 | – | 46 |
| 2nd `MLP` layer | feedforward neural network ($46 \to 5$) | 46 | – | 460 |
| total | | | | 486,566 |

Table 4: Behavior classification performance (MCC ×100) of the proposed model, reported as mean ± SD over 1,000 independent runs.

| behavior | MCC |
|---|---|
| grazing | 95.60 ± 0.11 |
| ruminating | 87.78 ± 0.25 |
| resting | 82.44 ± 0.37 |
| walking | 82.97 ± 0.63 |
| other | 55.06 ± 1.43 |
| overall | 88.47 ± 0.18 |

Table 5: Mean ± SD confusion matrix over 1,000 independent runs for behavior classification, green: correct predictions, red: errors.

| | | predicted | | | | |
|---|---|---|---|---|---|---|
| | | grazing | ruminating | resting | walking | other |
| true | grazing | 9411.04 ± 5.14 | 42.46 ± 3.32 | 29.74 ± 1.72 | 18.62 ± 3.24 | 10.15 ± 2.63 |
| | ruminating | 62.65 ± 4.21 | 3894.63 ± 9.45 | 374.87 ± 8.47 | 4.64 ± 1.55 | 15.21 ± 3.34 |
| | resting | 25.11 ± 2.06 | 202.74 ± 13.35 | 2071.40 ± 13.72 | 2.90 ± 0.85 | 27.85 ± 3.61 |
| | walking | 94.00 ± 4.43 | 3.21 ± 1.09 | 6.39 ± 1.20 | 411.55 ± 4.74 | 9.84 ± 2.85 |
| | other | 92.14 ± 3.91 | 81.84 ± 6.18 | 71.34 ± 4.69 | 25.50 ± 2.81 | 192.18 ± 7.09 |

We use the Matthews correlation coefficient (MCC) (Matthews, 1975) as the primary measure of behavior classification performance. MCC summarizes performance using all entries of the confusion matrix and is robust to class imbalance. For the binary case, it is defined as

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \in [-1, 1].$$

An MCC of 1 denotes perfect predictions, 0 indicates chance-level performance, and −1 signifies complete disagreement. For multi-class evaluation, we use the standard multiclass MCC computed from the full confusion matrix, which preserves the interpretability and imbalance robustness of the binary form.

We repeat the five-fold stratified CV experiment 1,000 times with independent random initializations (using fixed, unshuffled folds). For each run, we compute the out-of-fold (OOF) MCC by concatenating predictions for the held-out fold across all five splits and then evaluating MCC on the pooled confusion matrix. We report the mean and standard deviation (SD) of MCC across all runs in Table 4 (both per behavior and overall), and present the mean and SD of the corresponding confusion matrices in Table 5.

## 4. Estrus Detection

In this section, we describe and evaluate our proposed estrus detection algorithm.

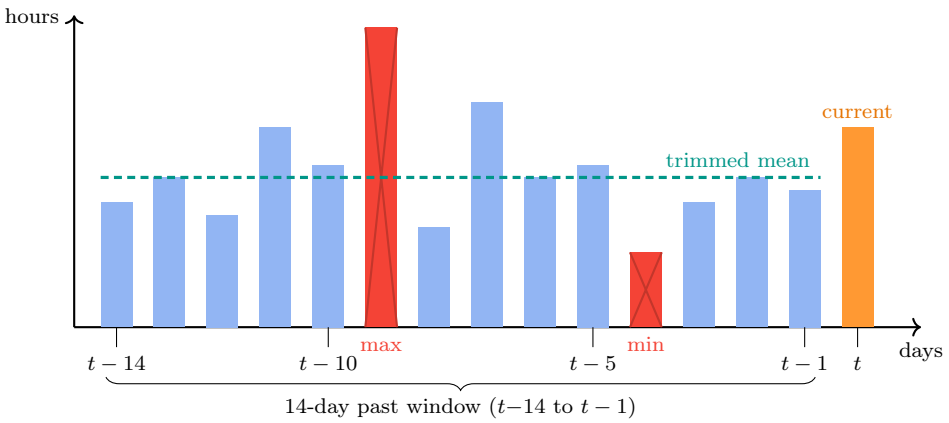


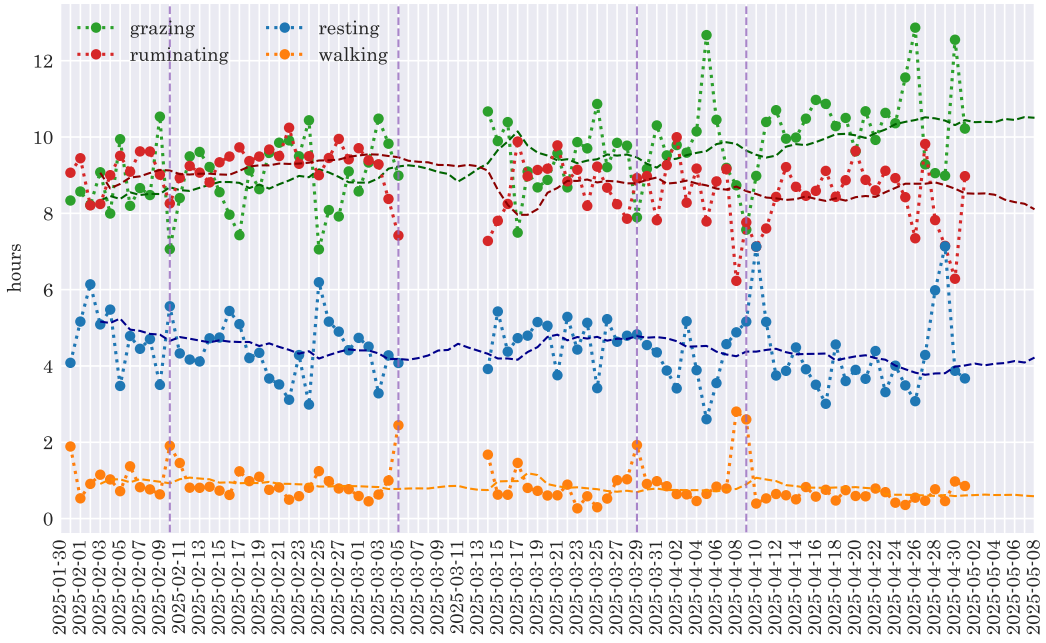

Figure 5: Example dynamic baseline computation. Daily durations over the past 14 days (blue) form a causal window. The minimum and maximum values (red) are discarded, and the mean of the remaining values defines the baseline (dashed line).

### *4.1. Feature Construction*

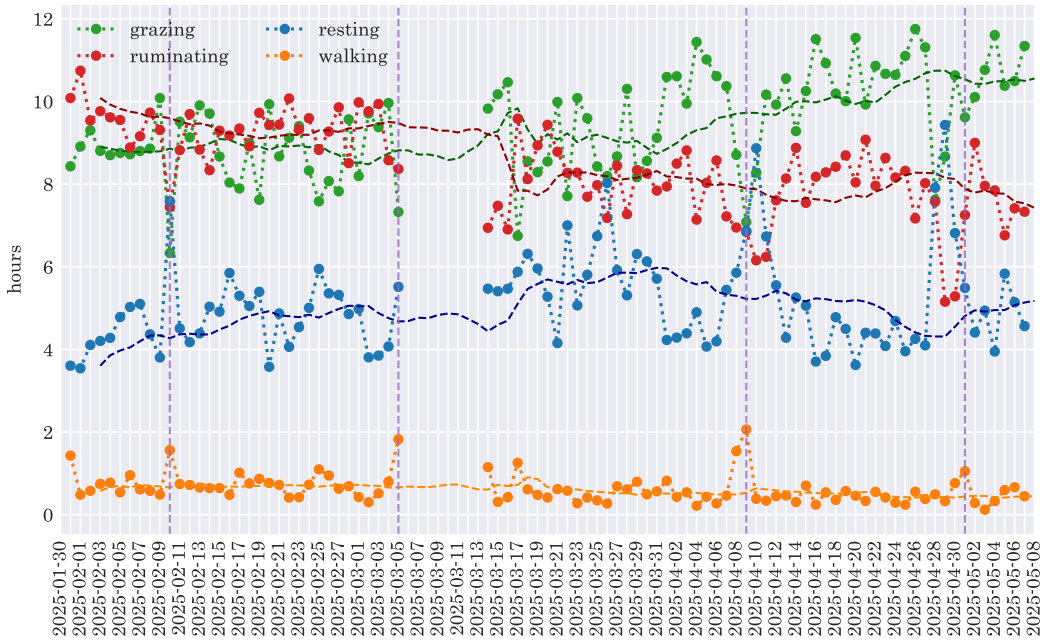
We pose estrus detection as a binary classification task that distinguishes estrus from non-estrus days. For each animal, we first run the behavior classifier on raw accelerometer streams to produce predictions on 5.12 s (256 sample) segments, then aggregate these to daily durations (hours) of four mutually exclusive behaviors: grazing, ruminating, resting, and walking. We exclude the "other" category to avoid collinearity, as the daily durations sum to 24 hours. Accordingly, we compute a per-animal behavior profile for each calendar day (00:00-24:00 local time) with near-complete accelerometer coverage.

To address individual variability and slow temporal trends, we establish a dynamic baseline for each animal and behavior using a 14-day causal sliding window. We calculate the baseline for a given day as the trimmed mean of the preceding 14 calendar days' durations. If fewer than three valid days (with near-complete accelerometer data) fall within that window, we deem the baseline to be unreliable and do not compute any feature for that day. To calculate the trimmed mean, we remove the minimum and maximum values within the window to ensure robustness against outliers. This procedure creates a stable daily reference against which we can reliably measure deviations. In Figure 5, we visually depict the procedure, representing the raw daily values as blue bars, the trimmed extrema as red bars, and the calculated baseline as a dashed line. In Figure 6, we show daily behavior durations and their baselines (dashed) together with ground-truth estrus days (vertical dashed lines) for two representative animals over the entire trial.

We define the estrus-detection features as per-animal, baseline-normalized daily durations. Specifically, for each animal and calendar day with a valid baseline, we compute the deviation of each behavior's observed duration from its 14-day trimmed-mean baseline. This normalization is intended to isolate short-term departures from each animal's typical pattern that are indicative of estrus. After quality filtering and baseline-availability checks, the dataset comprises 55 estrus cow-days and 927 non-estrus cow-days. Figure 7 illustrates these




(a) animal 37




(b) animal 44

Figure 6: Predicted daily behavior durations and their respective baselines (dashed lines) for two cattle during the entire trial period. Estrus days are marked by vertical dashed lines.

features for two representative animals with vertical dashed lines indicating ground-truth estrus days.

In Figure 8, we summarize the statistics of the estrus-detection features (baseline-normalized daily behavior deviations) for estrus and non-estrus days. Figure 8(a) shows the mean for each behavior with 95% confidence intervals. Figure 8(b) shows the corresponding distributions as violin plots, where short horizontal tick marks indicate the respective medians. Together, these plots demonstrate that, on average, cattle graze and ruminate less, and walk and rest more, on estrus days relative to non-estrus days.

### *4.2. Model*

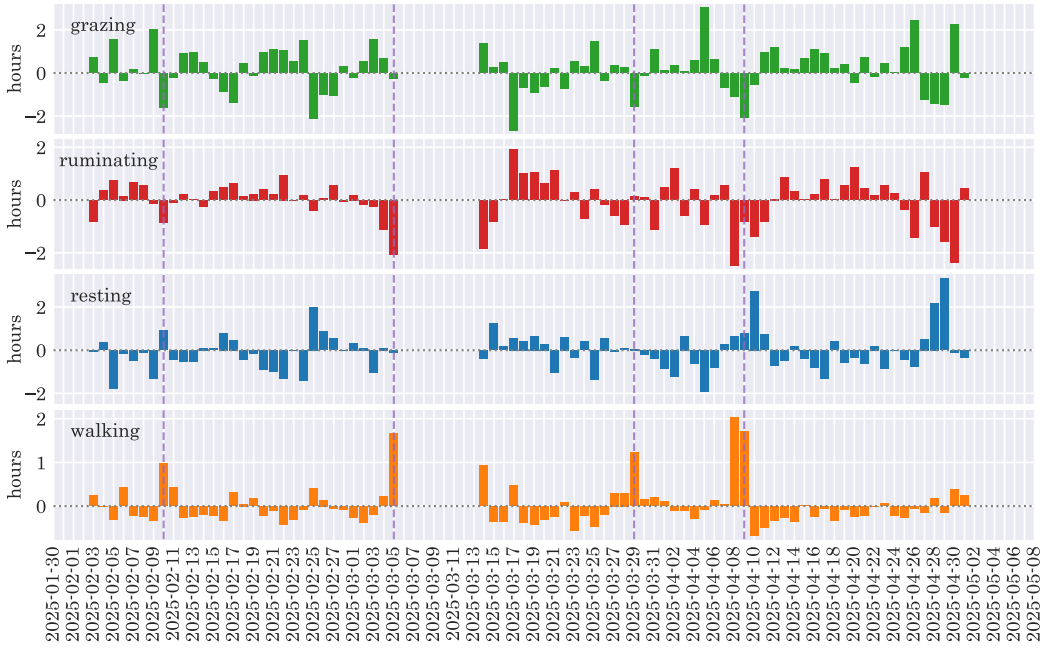
To recognize estrus events from the calculated daily features, we employ an $\ell_2$-regularized logistic regression (LogReg) model on standardized inputs. We implement the model as a pipeline ($z$-score standardization $\rightarrow$ LogReg) to ensure that feature scaling is fit exclusively within each CV split. We determine the optimal regularization strength $C$ once in a preliminary tuning step. In particular, over a log-spaced grid $C \in [10^{-2}, 10^{3}]$, we execute $1,000$ independent repetitions of five-fold stratified CV, each utilizing a unique random seed. The value that minimizes the mean OOF negative log-likelihood loss (log-loss), aggregated across all runs, is $C \approx 48.63$. We fix this value for all subsequent evaluation and reporting.

### *4.3. Evaluation*

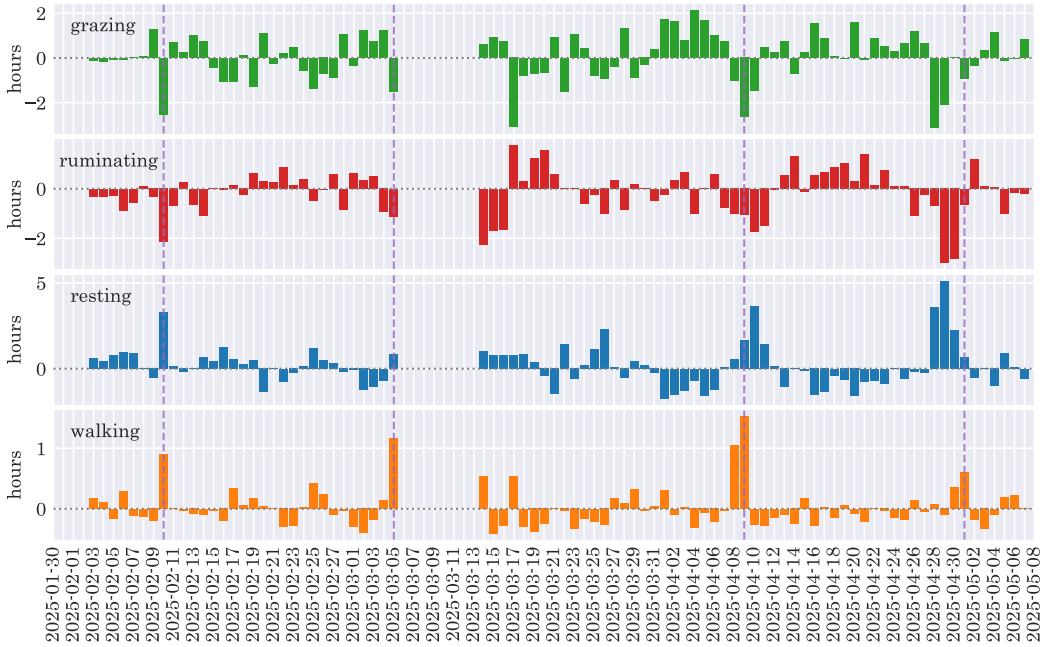
We evaluate our estrus-detection model (i.e., binary classifier for estrus vs. non-estrus) using five-fold stratified CV, repeated 1,000 times with independent shuffles and solver seeds. For each of the 1,000 runs, we obtain OOF predictions by concatenating the held-out predictions across the five splits. We compute the following threshold-free performance measures directly from the OOF probabilities: area under the receiver operating characteristic curve (ROC-AUC), area under the precision-recall curve (PR-AUC), log-loss, and Brier score. We calculate the threshold-dependent performance measures, precision, recall, F1, and MCC, from the pooled OOF confusion matrix at any given decision threshold ($\tau$). We report the mean $\pm$ SD across the 1,000 repeats for all measures. For the threshold-dependent measures, we provide results at the conventional threshold $\tau = 0.5$ and at a single recommended threshold $\tau = 0.6232$, chosen as the maximizer of the mean OOF MCC over a threshold sweep aggregated across repetitions.

In Tables 6 and 7, we present the mean $\pm$ SD values for the threshold-dependent performance measures and the confusion matrices at the conventional threshold $\tau = 0.5$ and the optimal threshold $\tau = 0.6232$ while, in Table 8, we provide the mean $\pm$ SD values for the threshold-free performance measures.

The results in Tables 6-8 demonstrate that using the four baseline-normalized features, the classifier achieves high precision and recall for the estrus class while maintaining excellent performance for the non-estrus class, indicating that daily deviations from behavioral

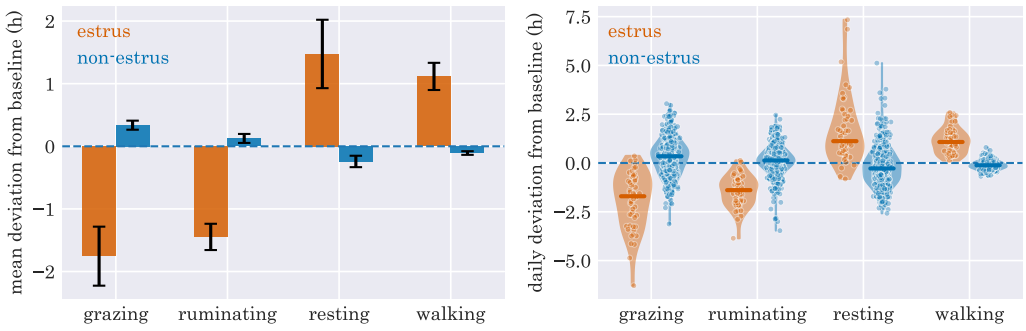



(a) animal 37




(b) animal 44

Figure 7: Baseline-normalized features for two cattle during the entire trial period. Estrus days are marked by vertical dashed lines.




(a) means with CIs

(b) violin plots and medians

Figure 8: (a) Mean (together with 95% CI) of baseline-normalized daily features (corresponding to grazing, ruminating, resting, and walking behaviors) across the entire dataset for estrus and non-estrus days. CIs are $t$-based and computed across cattle after averaging per animal to avoid day-count bias. (b) Corresponding distributions shown as violin plots with one dot per datapoint. The horizontal tick marks indicate the respective medians.

Table 6: Mean ± SD across 1,000 independent runs for precision, recall, F1, and MCC (×100) at the default ($\tau = 0.5$) and optimal ($\tau = 0.6232$) thresholds.

| threshold | class | precision | recall | F1 score | MCC |
|---|---|---|---|---|---|
| 0.5 | estrus | $94.92 \pm 1.46$ | $92.63 \pm 0.65$ | $93.76 \pm 0.80$ | $93.40 \pm 0.85$ |
| | non-estrus | $99.56 \pm 0.04$ | $99.70 \pm 0.09$ | $99.63 \pm 0.05$ | |
| 0.6232 | estrus | $96.70 \pm 1.71$ | $91.72 \pm 1.34$ | $94.14 \pm 1.18$ | $93.84 \pm 1.25$ |
| | non-estrus | $99.51 \pm 0.08$ | $99.81 \pm 0.10$ | $99.66 \pm 0.07$ | |

Table 7: Mean ± SD confusion matrices over 1,000 independent runs for estrus detection at the default ($\tau = 0.5$) and recommended ($\tau = 0.6232$) thresholds, green: correct predictions, red: errors.

| | | predicted ($\tau = 0.5$) | | predicted ($\tau = 0.6232$) | |
|---|---|---|---|---|---|
| | | estrus | non-estrus | estrus | non-estrus |
| true | estrus | $50.95 \pm 0.36$ | $4.05 \pm 0.36$ | $50.45 \pm 0.74$ | $4.55 \pm 0.74$ |
| | non-estrus | $2.74 \pm 0.82$ | $924.26 \pm 0.82$ | $1.73 \pm 0.92$ | $925.27 \pm 0.92$ |

Table 8: Mean ± SD across 1,000 independent runs for the considered threshold-free performance measures.

| measure | ROC-AUC | PR-AUC | log-loss | Brier score |
|---|---|---|---|---|
| value | $0.9985 \pm 0.0004$ | $0.9812 \pm 0.0035$ | $0.0205 \pm 0.0023$ | $0.0058 \pm 0.0006$ |

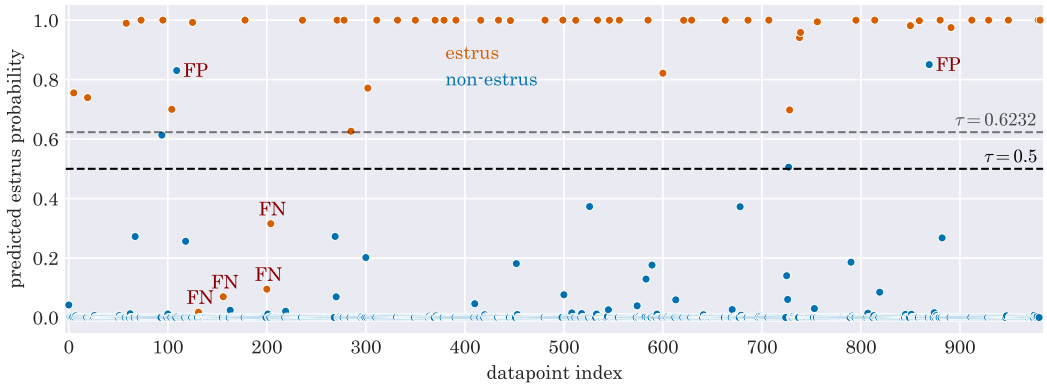



Figure 9: Predicted estrus probabilities (OOF) for all estrus-annotated cow-days in a representative CV split.

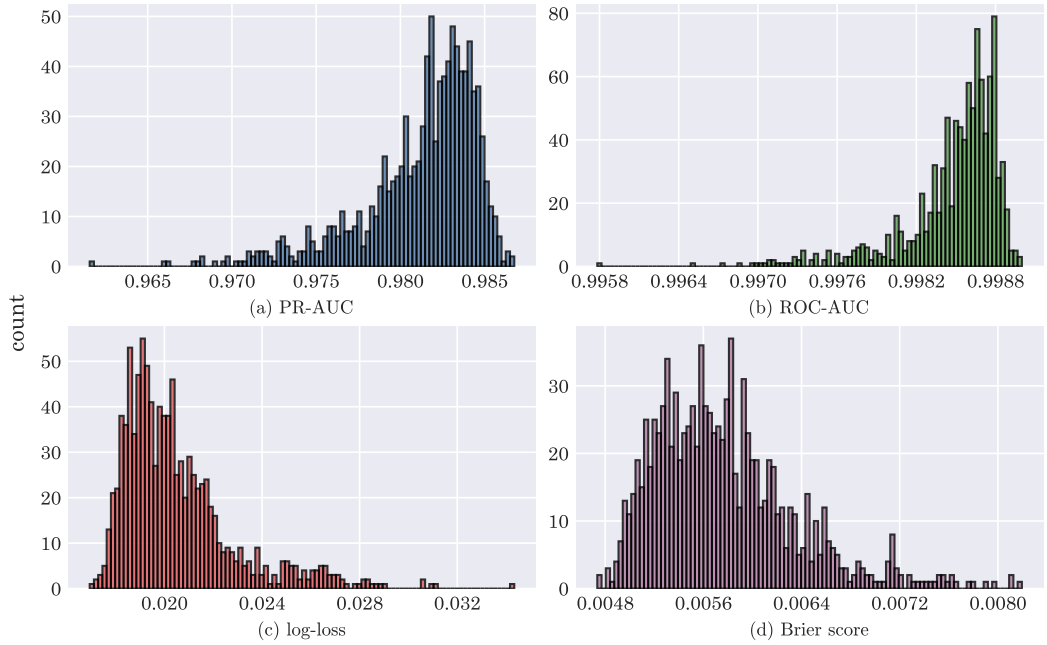
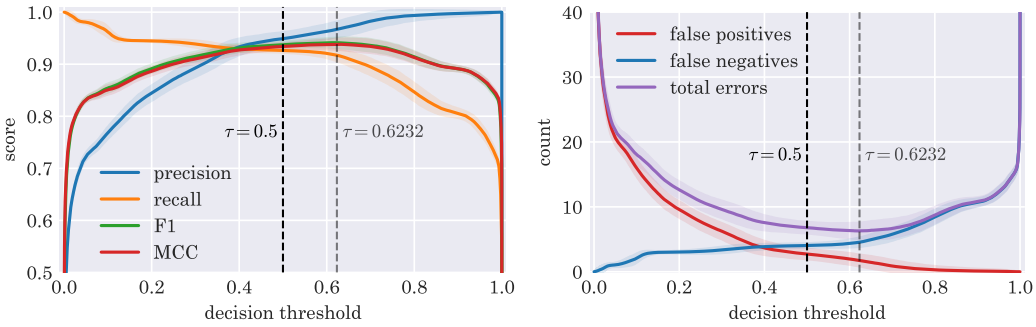



Figure 10: Histograms of OOF threshold-independent performance measures, PR-AUC, ROC-AUC, log-loss, and Brier score, across 1,000 repeated CV splits.

baselines are highly predictive of estrus, both naturally occurring and synchronized. Setting the operating point at $\tau = 0.6232$ yields a modest improvement in MCC and F1 with minimal trade-offs.

In Figure 9, we depict OOF predicted estrus probabilities for all cow-days annotated as estrus or non-estrus under a representative CV split. Dashed horizontal lines mark the decision thresholds $\tau = 0.5$ and $\tau = 0.6232$. We also mark the false positives and false negatives as FP and FN, respectively. In Figure 10, we present the histograms of the OOF threshold-independent performance measures across 1,000 repeated CV splits. Finally, in Figure 11, we plot the OOF threshold-dependent measures together with the counts of false positives, false negatives, and total errors as functions of the decision threshold, aggregated over 1,000 CV repeats. Shaded bands denote $\pm 1$ SD.




(a) performance measures

(b) error counts

Figure 11: OOF threshold-dependent performance measures, precision, recall, F1, and MCC, together with the counts of false positives, false negatives, and total errors as functions of the decision threshold, aggregated over 1,000 CV repeats.

## 5. Concluding Remarks

*Summary.* We presented a lightweight, end-to-end pipeline for in-situ estrus detection designed to run directly on a smart ear-tag's embedded system, mapping raw triaxial accelerometer data to an estrus probability. The foundation is a compact deep-learning-based behavior classifier that employs constrained temporal filters, depthwise residual blocks, and block-wise mean-absolute pooling to distill accelerometer streams into compact, interpretable descriptors. Our CMSIS-C implementation of the model fits comfortably within the tight memory budget of a low-power Cortex-M4F-class microcontroller, enabling low-latency on-tag inference. Building on the daily aggregate behavior predictions, we fit a regularized LogReg model to baseline-normalized features that capture deviations from individual-specific baseline daily behavior durations. The resulting LogReg classifier provides strong discrimination between estrus and non-estrus days and maintains excellent calibration under repeated CV, with further performance improvements achieved through careful optimization of the decision threshold.

*Cycle-aware filtering.* We explored post-processing strategies that enforce biological constraints on estrus recurrence by requiring an inter-estrus interval of at least 17 days and at most 24 days. In principle, such constraints can suppress spurious detections and align predictions with known physiology. However, with our dataset, neither basic nor advanced cycle filters improved upon the raw LogReg detector. Two factors likely explain this: (i) hormone-induced estrus synchronization on April 9 disrupted natural cyclicity, making cycle-based spacing rules less informative, and sometimes counterproductive, around the synchronization window, and (ii) the cycle-agnostic LogReg classifier already achieves near-ceiling performance on a relatively small set of 55 estrus events, leaving limited headroom for post-hoc gains. Nevertheless, synchronization protocols are common in commercial settings and need be considered in practical deployments. From this perspective, it is reassuring that our detector performs well for both naturally occurring and synchronized estrus without relying

on rigid cycle spacing. Therefore, we view cycle-aware post-processing as offering limited benefit for synchronized, data-limited cohorts, but as promising for larger herds managed without synchronization, where multiple natural cycles can be observed and temporal consistency is informative. Where synchronization programs are used, cycle filters can be gated by schedule-aware masks that temporarily relax spacing rules during the programmed window.

*Choice of estrus-detection model.* We evaluated alternative binary classifiers including random forest, extreme gradient boosting (XGBoost), and MLP, but selected LogReg for three main reasons: (i) interpretability: coefficients (and odds ratios) directly quantify how each baseline-normalized feature (daily behavior deviation) changes the odds of estrus, providing valuable insights to domain experts, (ii) computational efficiency: LogReg relies on linear operations and is well suited for deployment on embedded devices with limited computational and energy resources, and (iii) small-sample robustness: with 982 cow-day datapoints, LogReg generalizes reliably with minimal tuning and avoids overfitting issue of more complex models such as MLPs. We choose the regularization strength once by minimizing mean OOF log-loss under repeated stratified CV and keep it fixed thereafter. We report the performance evaluation results for the conventional decision threshold $\tau = 0.5$ and a single recommended threshold $\tau = 0.6232$ that maximizes mean OOF MCC over a threshold sweep.

*Limitations.* Our current estrus detection pipeline has several key limitations that guide future work:

(i) The system operates solely on daily, baseline-normalized behavior-deviation features derived from a single accelerometry-processing pipeline and evaluated on a finite and partially synchronized cohort.

(ii) The proposed approach is developed and evaluated in a single deployment configuration, which limits immediate applicability. Generalization to other management regimes, breeds, production systems, and hardware (ear-tag vs. collar) remains to be established.

(iii) Despite the use of constrained convolutions (the `cconv` layer) to attenuate the quasi-static gravity component, the behavior-classification features can remain sensitive to sensor orientation and device-specific characteristics. Robust deployment will likely require orientation-invariant features and cross-device calibration.

(iv) Our approach uses 24-hour-scale estrus annotations, meaning within-day estrus onset and duration are not estimated. In addition, we defined days using the civil calendar (midnight-midnight), yet alternative boundaries (e.g., dawn-to-dawn or sliding 24 h windows) may better align with the animals' diurnal activity and event timing.

(v) The trial included a hormone-induced synchronization event, which may limit the representativeness of natural cyclicity and the utility of spacing-based post-processing

techniques.

(vi) The decision thresholds were selected using OOF performance and may require retuning when deployed in populations with different estrus prevalences or varying decision costs.

*Practical implications and outlook.* The developed pipeline's efficiency and transparency enable continuous field monitoring under tight power, memory, and bandwidth constraints, with straightforward embedded (on ear-tag) or edge deployment and easy integration into existing workflows. Looking ahead, we will pursue:

(i) using repeated-cycle information accumulated during deployment to map each animal's cycle on a calendar and forecast likely estrus windows for proactive scheduling of artificial insemination or embryo transfer,

(ii) multimodal fusion (e.g., GNSS and proximity) for finer temporal resolution and higher accuracy and robustness,

(iii) cross-device and cross-domain adaptation to accommodate heterogeneous devices, herds, and production environments,

(iv) per-animal adaptive thresholds with online calibration to track drift,

(v) integer quantization (Q15/Q7) via CMSIS-NN kernels for additional latency and energy gains,

(vi) semi- and self-supervised pretraining on unlabeled streams to reduce reliance on curated labels,

(vii) explicit differentiation between natural and synchronized estrus as our initial observations indicate stronger behavioral changes during synchronized episodes, and

(viii) extension of the framework to related reproductive or health endpoints, e.g., gestation status, parturition timing, dystocia risk, and early-postpartum complications.

**Acknowledgment**

We thank the CSIRO Chiswick research farm staff, particularly Dan Driscoll, Graham Acton, and Dom Niemeyer, for assistance with hardware deployment, trial operations, day-to-day animal care, and logistics. We are grateful to CERES TAG LTD for financial support and to Embeint Pty Ltd for technical assistance.

**References**

Arablouei, R., Bishop-Hurley, G.J., Bagnall, N., Ingham, A., 2024. Cattle behavior recognition from accelerometer data: Leveraging in-situ cross-device model learning. Computers and Electronics in Agriculture 227, 109546. doi:https://doi.org/10.1016/j.compag.2024.109546.

Arablouei, R., Currie, L., Kusy, B., Ingham, A., Greenwood, P.L., Bishop-Hurley, G.J., 2021. In-situ classification of cattle behavior using accelerometry data. Computers and Electronics in Agriculture 183, 106045.

Arablouei, R., Wang, L., Currie, L., Alvarenga, F.A.P., Bishop-Hurley, G.J., 2023a. Animal behavior classification via deep learning on embedded systems. Computers and Electronics in Agriculture 207, 107707.

Arablouei, R., Wang, L., Phillips, C., Currie, L., Yates, J., Bishop-Hurley, G.J., 2023b. In-situ animal behavior classification using knowledge distillation and fixed-point quantization. Smart Agricultural Technology 4, 100159.

Arablouei, R., Wang, Z., Bishop-Hurley, G.J., Liu, J., 2023c. Multi-modal sensor data fusion for in-situ classification of animal behavior using accelerometry and gnss data. Smart Agricultural Technology 4, 100163.

At-Taras, E., Spahr, S., 2001. Detection and characterization of estrus in dairy cattle with an electronic heatmount detector and an electronic activity tag. Journal of Dairy Science 84, 792–798. doi:10.3168/jds.S0022-0302(01)74535-3.

Benaissa, S., Tuyttens, F., Plets, D., Trogh, J., Martens, L., Vandaele, L., Joseph, W., Sonck, B., 2020. Calving and estrus detection in dairy cattle using a combination of indoor localization and accelerometer sensors. Computers and Electronics in Agriculture 168, 105153. doi:https://doi.org/10.1016/j.compag.2019.105153.

Bikker, J., van Laar, H., Rump, P., Doorenbos, J., van Meurs, K., Griffioen, G., Dijkstra, J., 2014. Technical note: Evaluation of an ear-attached movement sensor to record cow feeding behavior and activity. Journal of Dairy Science 97, 2974–02979. doi:10.3168/jds.2013-7560.

Bloch, V., Frondelius, L., Arcidiacono, C., Mancino, M., Pastell, M., 2023. Development and analysis of a cnn- and transfer-learning-based classification model for automated dairy cow feeding behavior recognition from accelerometer data. Sensors 23, 2611.

Borchers, M., Chang, Y., Proudfoot, K., Wadsworth, B., Stone, A., Bewley, J., 2017. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. Journal of Dairy Science 100, 5664–5674. doi:10.3168/jds.2016-11526.

Brandes, S., Sicks, F., Berger, A., 2021. Behaviour classification on giraffes (giraffa camelopardalis) using machine learning algorithms on triaxial acceleration data of two commonly used gps devices and its possible application for their management and conservation. Sensors 21, 2229.

Cerri, R.L.A., Burnett, T.A., Madureira, A.M.L., Silper, B.F., Denis-Robichaud, J., LeBlanc, S., Cooke, R.F., Vasconcelos, J.L.M., 2021. Symposium review: Linking activity-

sensor data and physiology to improve dairy cow fertility. Journal of Dairy Science 104, 1220–1231. doi:10.3168/jds.2019-17893.

Chollet, F., 2017. Xception: Deep Learning with Depthwise Separable Convolutions , in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA. pp. 1800–1807. doi:10.1109/CVPR.2017.195.

Dunford, C.E., Marks, N.J., Wilson, R.P., Scantlebury, D.M., 2024. Identifying animal behaviours from accelerometers: Improving predictive accuracy of machine learning by refining the variables selected, data frequency, and sample duration. Ecology and Evolution 14, e11380.

Eckhardt, R., Arablouei, R., Ingham, A., McCosker, K., Bernhardt, H., 2025. Livestock behaviour forecasting via generative artificial intelligence. Smart Agricultural Technology 11, 100987. doi:https://doi.org/10.1016/j.atech.2025.100987.

Edwards, E., Krawczel, P., Dann, H., Schneider, L., Whitlock, B., Proudfoot, K., 2020. Calving location preference and changes in lying and exploratory behavior of preparturient dairy cattle with access to pasture. Journal of Dairy Science 103, 5455–5465. doi:10.3168/jds.2019-17218.

Fogarty, E.S., Swain, D.L., Cronin, G.M., Moraes, L.E., Trotter, M., 2020. Behaviour classification of extensively grazed sheep using machine learning. Computers and Electronics in Agriculture 169, 105175.

Hlimi, A., El Otmani, S., Elame, F., Chentouf, M., El Halimi, R., Chebli, Y., 2024. Application of precision technologies to characterize animal behavior: A review. Animals 14.

Hockey, C., Morton, J., Norman, S., McGowan, M., 2010. Evaluation of a neck mounted 2-hourly activity meter system for detecting cows about to ovulate in two paddock-based australian dairy herds. Reproduction in Domestic Animals 45, e107–e117. doi:10.1111/j.1439-0531.2009.01531.x.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv abs/1704.04861. URL: https://api.semanticscholar.org/CorpusID:12670695.

Ignatov, A., 2018. Real-time human activity recognition from accelerometer data using convolutional neural networks. Applied Soft Computing 62, 915–922. doi:https://doi.org/10.1016/j.asoc.2017.09.027.

Jin, Z., Shu, H., Hu, T., Jiang, C., Yan, R., Qi, J., Wang, W., Guo, L., 2024. Behavior classification and spatiotemporal analysis of grazing sheep using deep learning. Computers

and Electronics in Agriculture 220, 108894. doi:https://doi.org/10.1016/j.compag.2024.108894.

Kamminga, J.W., Le, D.V., Meijers, J.P., Bisby, H., Meratnia, N., Havinga, P.J.M., 2018. Robust sensor-orientation-independent feature selection for animal activity recognition on collar tags. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, 15.

Kamphuis, C., DelaRue, B., Burke, C., Jago, J., 2012. Field evaluation of 2 collar-mounted activity meters for detecting cows in estrus on a large pasture-grazed dairy farm. Journal of Dairy Science 95, 3045–3056. doi:10.3168/jds.2011-4934.

Krieger, S., Oczak, M., Lidauer, L., Berger, A., Kickinger, F., Öhlschuster, M., Auer, W., Drillich, M., Iwersen, M., 2019. An ear-attached accelerometer as an on-farm device to predict the onset of calving in dairy cows. Biosystems Engineering 184, 190–199. doi:10.1016/j.biosystemseng.2019.06.011.

Lopez, H., Satter, L., Wiltbank, M., 2004. Relationship between level of milk production and estrous behavior of lactating dairy cows. Animal Reproduction Science 81, 209–223. doi:https://doi.org/10.1016/j.anireprosci.2003.10.009.

Løvendahl, P., Chagunda, M.G.G., 2010. On the use of physical activity monitoring for estrus detection in dairy cows. Journal of Dairy Science 93, 249–259. doi:10.3168/jds.2008-1721.

Mao, A., Huang, E., Wang, X., Liu, K., 2023. Deep learning-based animal activity recognition with wearable sensors: Overview, challenges, and future directions. Computers and Electronics in Agriculture 211, 108043.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405, 442–451.

Mayo, L., Silvia, W., Ray, D., Jones, B., Stone, A., Tsai, I., Clark, J., Bewley, J., Heersche, G., 2019. Automated estrous detection using multiple commercial precision dairy monitoring technologies in synchronized dairy cows. Journal of Dairy Science 102, 2645–2656. doi:https://doi.org/10.3168/jds.2018-14738.

Merkelytė, I., Šiukščius, A., Nainienė, R., 2025. The role of sensor technologies in estrus detection in beef cattle: A review of current applications. Animals 15, 2313. doi:10.3390/ani15152313.

Moore, S., Aublet, V., Butler, S., 2021. Monitoring estrous activity in pasture-based dairy cows. Theriogenology 160, 90–94. doi:10.1016/j.theriogenology.2020.11.002.

Pavlovic, D., Davison, C., Hamilton, A., Marko, O., Atkinson, R., Michie, C., Crnojević, V., Andonovic, I., Bellekens, X., Tachtatzis, C., 2021. Classification of cattle behaviours using

neck-mounted accelerometer-equipped collars and convolutional neural networks. Sensors 21, 4050.

Peng, Y., Kondo, N., Fujiura, T., Suzuki, T., Wulandari, Yoshioka, H., Itoyama, E., 2019. Classification of multiple cattle behavior patterns using a recurrent neural network with long short-term memory and inertial measurement units. Computers and Electronics in Agriculture 157, 247–253.

Perez Marquez, H.J., Ambrose, D.J., Bench, C.J., 2023. Behavioral changes to detect estrus using ear-sensor accelerometer compared to in-line milk progesterone in a commercial dairy herd. Frontiers in Animal Science 4, 1149085. doi:10.3389/fanim.2023.1149085.

Reith, S., Hoy, S., 2018. Review: Behavioral signs of estrus and the potential of fully automated systems for detection of estrus in dairy cattle. Animal 12, 398–407.

Roelofs, J., López-Gatius, F., Hunter, R.H., van Eerdenburg, F.J., Hanzen, C., 2010. When is a cow in estrus? clinical and practical aspects. Theriogenology 74, 327–344. doi:10.1016/j.theriogenology.2010.02.016.

Rorie, R., Bilby, T., Lester, T., 2002. Application of electronic estrus detection technologies to reproductive management of cattle. Theriogenology 57, 137–148. doi:https://doi.org/10.1016/S0093-691X(01)00663-X.

Russel, N.S., Selvaraj, A., 2024. Decoding cow behavior patterns from accelerometer data using deep learning. Journal of Veterinary Behavior 74, 68–78.

Saint-Dizier, M., Chastant-Maillard, S., 2012. Towards an automated detection of oestrus in dairy cattle. Reproduction in Domestic Animals 47, 1056–1061.

Saint-Dizier, M., Chastant-Maillard, S., 2018. Potential of connected devices to optimize cattle reproduction. Theriogenology 112, 53–62. doi:10.1016/j.theriogenology.2018.01.009.

Schilkowsky, E., Granados, G., Sitko, E., Masello, M., Perez, M., Giordano, J., 2021. Evaluation and characterization of estrus alerts and behavioral parameters generated by an ear-attached accelerometer-based system for automated detection of estrus. Journal of Dairy Science 104, 6222–6237.

Schweinzer, V., Gusterer, E., Kanz, P., Krieger, S., Süß, D., Lidauer, L., Berger, A., Kickinger, F., Öhlschuster, M., Auer, W., Drillich, M., Iwersen, M., 2019. Evaluation of an ear-attached accelerometer for detecting estrus events in indoor housed dairy cows. Theriogenology 130, 19–25.

Valenza, A., Giordano, J., Lopes, G., Vincenti, L., Amundson, M., Fricke, P., 2012. Assessment of an accelerometer system for detection of estrus and treatment with gonadotropin-

releasing hormone at the time of insemination in lactating dairy cows. Journal of Dairy Science 95, 7115–7127. doi:10.3168/jds.2012-5639.

Wang, J., Bell, M., Liu, X., Liu, G., 2020. Machine-learning techniques can enhance dairy cow estrus detection using location and acceleration data. Animals 10, 1160. doi:10.3390/ani10071160.

Wang, J., Zhang, Y., Bell, M., Liu, G., 2022. Potential of an activity index combining acceleration and location for automated estrus detection in dairy cows. Information Processing in Agriculture 9, 288–299. doi:10.1016/j.inpa.2021.04.003.

Wang, L., Arablouei, R., Alvarenga, F.A.P., Bishop-Hurley, G.J., 2023. Animal behavior classification via accelerometry data and recurrent neural networks. Computers and Electronics in Agriculture 206, 107647.

Weinert-Nelson, J.R., Werner, J., Jacobs, A.A., Anderson, L., Williams, C.A., Davis, B.E., 2025. Effects of heat stress on the accuracy of an ear-tag accelerometer for monitoring rumination and eating behavior in dairy-beef cross cattle using an automated gold standard. Journal of Dairy Science 108, 735–749. doi:10.3168/jds.2024-24858.

Wiltbank, M., Lopez, H., Sartori, R., Sangsritavong, S., Gümen, A., 2006. Changes in reproductive physiology of lactating dairy cows due to elevated steroid metabolism. Theriogenology 65, 17–29. doi:10.1016/j.theriogenology.2005.10.003.