

Field validation of a proximity-gated AI framework for mounting-based estrus detection in dairy cattle

Wattana Pongsena^{*1)}, Prakaidoy Ditsayabut²⁾ and Phakpoom Chinprutthiwong¹⁾

¹⁾School of Software Engineering and Artificial Intelligence, Faculty of Liberal Arts and Science, Sisaket Rajabhat University, Sisaket 33000, Thailand

²⁾Department of Product Research and Development, Hyperm Corporation, Co., Ltd., Nakhonratchasima 30000, Thailand

Received 31 March 2026

Revised 18 July 2026

Accepted 21 July 2026

Abstract

Accurate estrus detection remains challenging because activity-based monitoring systems may misinterpret non-reproductive movements as estrus-related activity. This study hypothesized that mounting-based estrus detection can be improved by conditioning accelerometer-based motion classification on inter-animal proximity, where mounting is physically feasible. We developed and field-tested a proximity-gated AI framework integrating Bluetooth Low Energy (BLE) proximity sensing with tri-axial accelerometer-based motion analysis. Unlike conventional activity-only ML/DL approaches, the proposed framework treats estrus detection as an interaction-dependent behavioral event by first validating close-contact feasibility and then applying motion classification and event-level aggregation. Controlled evaluation using 120 manually annotated mounting events achieved an event-level F1-score of 93.2% with the LSTM model. A retrospective ablation analysis showed that BLE proximity gating reduced false-positive event candidates from 48 to 6 compared with ungated motion-only inference, corresponding to an 87.5% reduction. Precision increased from 70.0% to 94.8%, while recall decreased slightly from 93.3% to 91.7%. The system was subsequently deployed across three smallholder dairy farms using 71 collar-mounted devices. During field deployment, 31 confirmed estrus events were recorded, of which 23 were detected, corresponding to a field-level recall of 74.2%; no false-positive alerts were observed within this limited dataset. Among 13 system-detected cows that underwent artificial insemination, 11 did not return to estrus during the first expected post-insemination cycle follow-up, giving a first-cycle non-return proportion of 84.6% (11/13), used as a field proxy rather than confirmed pregnancy. These findings support the hypothesis that proximity-gated behavioral validation improves the biological specificity of mounting-based estrus detection, although broader multi-site and longitudinal validation remains necessary.

Keywords: AI-Driven estrus detection, Bluetooth Low Energy (BLE) proximity sensing, Mounting event detection, Deep learning, Precision livestock farming

1. Introduction

Accurate estrus detection is a critical factor influencing reproductive efficiency and economic sustainability in dairy production systems. Failure to identify estrus at the appropriate time results in missed insemination opportunities, extended calving intervals, and reduced lifetime milk yield [1, 2]. Although smallholder herds are smaller than commercial herds, reliable estrus detection remains difficult because estrus behavior is short, intermittent, and often occurs at night or outside routine observation periods. Smallholder farms also commonly face constraints in access to advanced commercial monitoring systems, timely reproductive decision support, and routine labor scheduling. Therefore, a low-power and biologically specific monitoring system may help reduce missed insemination opportunities and unnecessary interventions in resource-constrained farm settings.

Traditionally, estrus detection relies on visual observation of behavioral signs, particularly standing heat behavior. However, manual observation is labor-intensive, subjective, and often restricted to limited observation periods. Moreover, estrus-related behaviors frequently occur at night or outside routine monitoring schedules [3]. As herd sizes increase and labor resources decline, the limitations of visual observation become more evident, motivating the development of automated estrus detection technologies.

To address these challenges, various sensor-based monitoring systems have been proposed, including pedometers, accelerometers, pressure-sensitive mounting detectors, and rumination monitoring devices [4-6]. Reported performance varies substantially depending on sensor type, reference standard, and farm conditions. For example, Valenza et al. [7] reported that an accelerometer-based activity system detected 71% of cows in estrus, while heat-mount detectors detected 66% in the same experimental context. In a comparative field study, Holman et al. [8] reported sensitivities of 58.9% for neck-collar activity monitoring, 63.3% for pedometers, and 35.9% for scratchcard heat-mount detectors, with positive predictive values of 93.5%, 73.5%, and 63.9%, respectively. These quantitative findings indicate that conventional tools can support estrus detection but may still miss events or generate uncertainty under practical farm conditions.

*Corresponding author.

Email address: w.pongsena@sskru.ac.th

<https://doi.org/10.64960/easr.2026.266990>

Despite their advantages, activity-based estrus detection approaches often suffer from limited behavioral specificity. Increased locomotor activity can arise from various non-reproductive factors, including social competition, feeding disturbances, lameness, or environmental stressors [5, 7]. As a result, these systems may generate false-positive alerts that reduce farmer confidence and increase unnecessary insemination interventions, especially when the system lacks contextual evidence that a reproductive social interaction is physically feasible.

Among behavioral indicators of estrus, mounting behavior is widely recognized as one of the most biologically meaningful signs because it reflects sexually motivated social interaction during the peri-estrus period [9, 10]. Mounting events are closely associated with hormonal changes during the estrous cycle and can provide guidance for optimal insemination timing [11, 12]. Unlike generalized locomotor activity, mounting requires close physical interaction between animals and therefore provides stronger behavioral specificity for estrus expression. However, mounting behavior is typically short in duration, lasting approximately 3-5 seconds, and occurs sporadically within complex social interactions, making reliable automated detection technically challenging [8].

Recent advances in artificial intelligence (AI) and machine learning have significantly improved the ability to recognize complex behavioral patterns from high-frequency sensor data and wearable devices [13-15]. Computer-vision approaches have also been investigated for automated livestock behavior analysis [16]. Deep learning architectures such as recurrent neural networks and convolutional neural networks enable automated extraction of temporal motion features from wearable sensor signals [17]. AI-driven approaches have therefore been increasingly explored for livestock behavior recognition and health monitoring applications [18, 19].

However, most existing AI-based estrus detection systems still rely primarily on single-animal activity signals without considering the interaction context required for mounting behavior. From a biological perspective, mounting events inherently involve the spatial interaction of at least two animals. Ignoring this interaction constraint may lead to misinterpretation of unrelated motion patterns as potential estrus signals. The central hypothesis of this study was therefore that mounting-based estrus detection can be made more biologically specific by evaluating accelerometer-derived motion patterns only during inter-animal proximity intervals in which mounting is physically feasible.

From a smart farming systems perspective, reliable decision support requires not only accurate signal classification but also contextual validation to ensure biological plausibility [20]. In the context of estrus detection, this implies that motion signals should be evaluated within interaction-feasible situations rather than analyzed in isolation.

To address this limitation, this study proposes a proximity-gated AI framework for mounting-based estrus detection. The novelty of the proposed framework lies not in the use of ML/DL alone, but in the integration of interaction-aware BLE proximity validation before motion classification. This design formalizes estrus detection as a biologically constrained social interaction event rather than as an isolated increase in locomotor activity. BLE signals are used to identify short-range interaction intervals between animals, and accelerometer signals within these intervals are subsequently analyzed by machine learning and deep learning models to detect mounting behavior.

To establish reliable training data, controlled dummy-mounting experiments were conducted at a bull stud farm where breeding bulls repeatedly mount a stationary artificial dummy cow during semen collection procedures. This setup enables accurate annotation of mounting events and provides high-quality ground-truth data for model development. Following the controlled calibration phase, the system was deployed across multiple smallholder dairy farms to evaluate operational performance under real farm conditions.

Unlike studies that focus primarily on algorithmic accuracy under laboratory conditions, this research emphasizes system-level validation of an integrated estrus monitoring platform. The evaluation considers mounting detection accuracy, the independent contribution of proximity gating, field deployability, operational robustness, and reproductive management outcomes.

The main contributions of this study are summarized as follows:

1. Development of an AI-driven proximity-gated sensing architecture that integrates BLE-based proximity validation with motion-based mounting classification for biologically plausible estrus detection.
2. Systematic comparison of classical machine learning models and deep learning architectures for mounting-event recognition under controlled experimental conditions.
3. Formal event-level ablation analysis comparing ungated motion-only inference with BLE proximity-gated inference to quantify the contribution of proximity gating to false-positive suppression.
4. Field deployment of 71 collar-mounted sensing devices across three smallholder dairy farms to evaluate operational feasibility under real farm conditions.
5. Field-based validation of estrus detection performance under real farm conditions, including behavioral detection accuracy and reproductive management outcomes.

By combining interaction-aware sensing with AI-based behavioral recognition and real-world deployment, this study suggests a scalable and biologically grounded framework for automated estrus monitoring in smallholder dairy production systems.

2. Materials and methods

This section describes the design, implementation, and evaluation methodology of the proposed AI-driven IoT framework for mounting behavior detection targeting automated estrus detection in dairy cattle. The emphasis is placed on system architecture, proximity-gated inference design, data collection and annotation, temporal signal processing, event-level decision logic, and experimental evaluation under controlled calibration prior to real-farm deployment. Hardware components are described only to the extent necessary to support reliable and scalable AI-based inference.

2.1 System architecture for automated estrus detection

The overall system architecture is illustrated in Figure 1, presenting an end-to-end IoT-AI pipeline for proximity-gated mounting detection targeting automated estrus detection in dairy cattle.

Each animal is equipped with a collar-mounted sensing unit that continuously acquires proximity and motion data. These multimodal data streams are transmitted to a centralized backend server where AI-based inference is performed. The architecture is designed to decouple sensing, preprocessing, and decision making, allowing biological and physical constraints to be embedded directly into the AI pipeline rather than enforced through device-level rules.

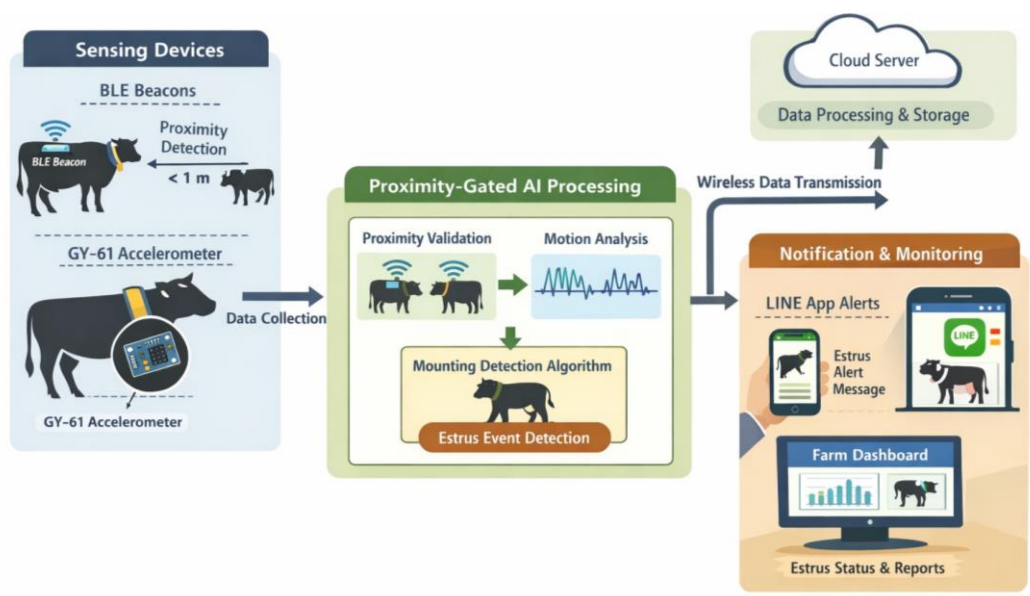


Figure 1 End-to-end IoT system architecture for proximity-gated mounting detection

At the sensing layer, Bluetooth Low Energy (BLE)-based proximity information is used to identify candidate interaction intervals during which mounting behavior is physically plausible. BLE proximity sensing provides a low-power and scalable mechanism for validating short-range interactions in indoor and semi-indoor environments [21, 22]. In the proposed framework, proximity information functions as a hard gating mechanism that constrains the temporal segments forwarded to AI models.

Within these proximity-gated intervals, motion signals acquired from a tri-axial accelerometer are analyzed by machine learning (ML) and deep learning (DL) models to infer mounting behavior. Centralized inference allows the AI framework to capture discriminative short-duration motion patterns associated with mounting behavior while suppressing false positives caused by unrelated movements [10, 23]. Event-level decisions are generated by aggregating window-level predictions in accordance with the biological duration of mounting behavior.

2.2 Collar-Mounted sensing device and IoT gateway

Each animal was equipped with a collar-mounted sensing device that provided multimodal input for AI-based mounting detection without performing on-device classification. As shown in Figure 2, the collar integrated a Bluetooth Low Energy (BLE) beacon for proximity estimation and a GY-61 module with a tri-axial accelerometer.

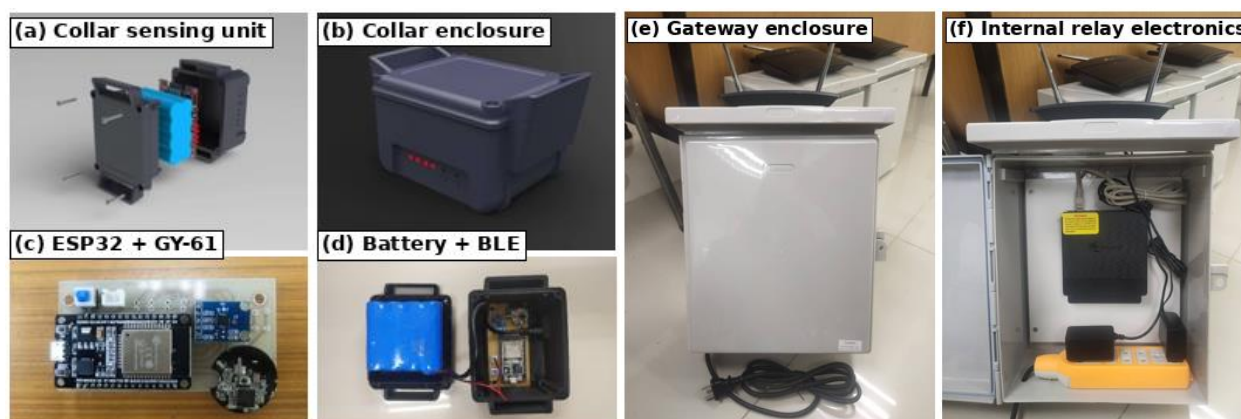


Figure 2 Collar-mounted sensing and relay hardware used in the field deployment: (a) collar sensing unit, (b) collar enclosure, (c) ESP32 microcontroller integrated with the GY-61 tri-axial accelerometer, (d) battery and BLE module inside the collar enclosure, (e) external gateway enclosure, and (f) internal relay electronics used to forward collar data to the server.

The BLE beacon periodically broadcast and scanned wireless signals to estimate inter-animal proximity using the received signal strength indicator (RSSI), while the tri-axial accelerometer captured posture-related and dynamic motion patterns associated with mounting behavior lasting approximately 3-5 s. BLE close-contact gating was implemented as an operational criterion for interaction-feasible proximity before motion classification.

The IoT gateway aggregated data streams from multiple collars, aligned BLE and accelerometer timestamps, performed lightweight buffering and packet validation, and forwarded synchronized data for backend inference through Wi-Fi or cellular connectivity. Figure 2 also includes an example signal relay unit used to transmit collar data to the server, while Figure 3 presents the data transmission and synchronization pipeline from collar-level sensing to backend AI inference.

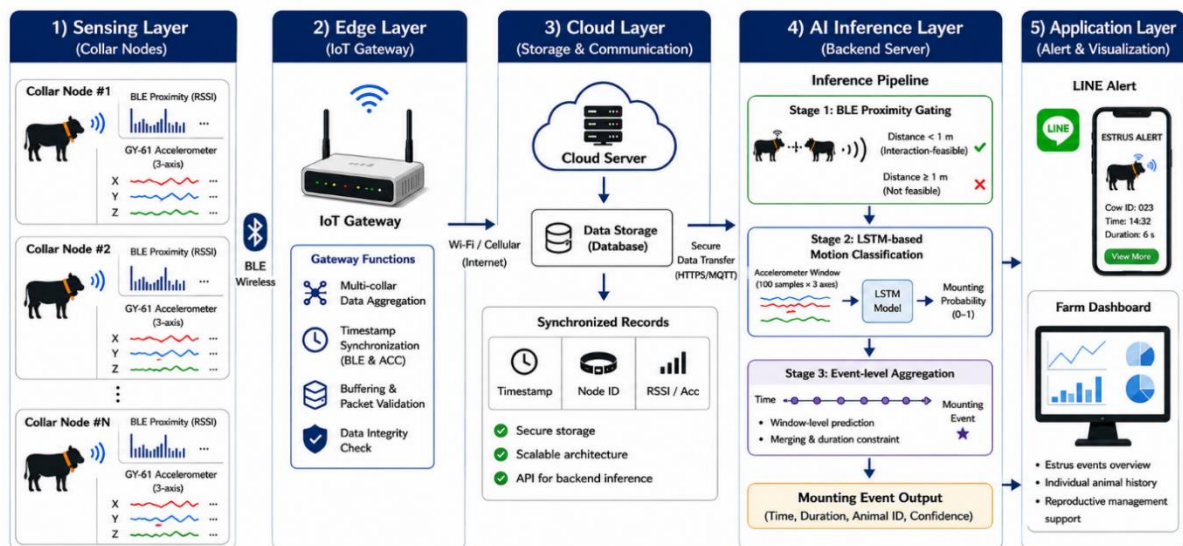


Figure 3 End-to-end data transmission and AI inference pipeline for proximity-gated mounting detection. BLE proximity and tri-axial accelerometer data from multiple collar nodes are synchronized by an IoT gateway, transmitted to cloud storage, and processed through proximity gating, LSTM-based motion classification, and event-level aggregation before generating LINE alerts and farm-dashboard reports.

Figure 3 summarizes the end-to-end data flow of the proposed estrus-detection system across five layers: sensing, edge gateway, cloud storage, AI inference, and application. Each collar-mounted node collects BLE RSSI data for inter-animal proximity estimation and tri-axial accelerometer signals for motion analysis. The IoT gateway aggregates data from multiple collars, synchronizes BLE and accelerometer timestamps, validates and buffers packets, and transfers the synchronized records to cloud storage via Wi-Fi or cellular connectivity.

Backend inference is performed in three stages. First, BLE proximity gating retains only close-contact intervals in which mounting is physically feasible. Second, the LSTM classifier analyzes 4-s accelerometer windows to estimate mounting probability. Third, adjacent positive windows are aggregated into complete mounting events using temporal voting and duration constraints. The resulting events are delivered to farmers through LINE alerts and summarized on the farm dashboard. This sequential design improves biological interpretability by requiring both interaction feasibility and mounting-related motion before an estrus alert is generated.

2.3 Data collection and ground-truth construction

Mounting behavior data were collected under controlled conditions to establish a calibrated development dataset for proximity-gated model training and systematic algorithm comparison prior to real-farm deployment. This controlled phase serves as a model development and calibration stage within the unified inference architecture, ensuring high-quality ground-truth labels for supervised learning and fair model comparison while isolating mounting-specific motion dynamics from uncontrolled social interactions.

2.3.1 Controlled dummy-mounting protocol

Data collection was conducted at a bull stud farm during routine semen collection sessions, where breeding bulls are trained to mount a dummy cow. This setting provides a controlled yet biologically relevant environment in which mounting behavior occurs repeatedly and can be observed unambiguously. The controlled data collection setup is shown in Figure 4.



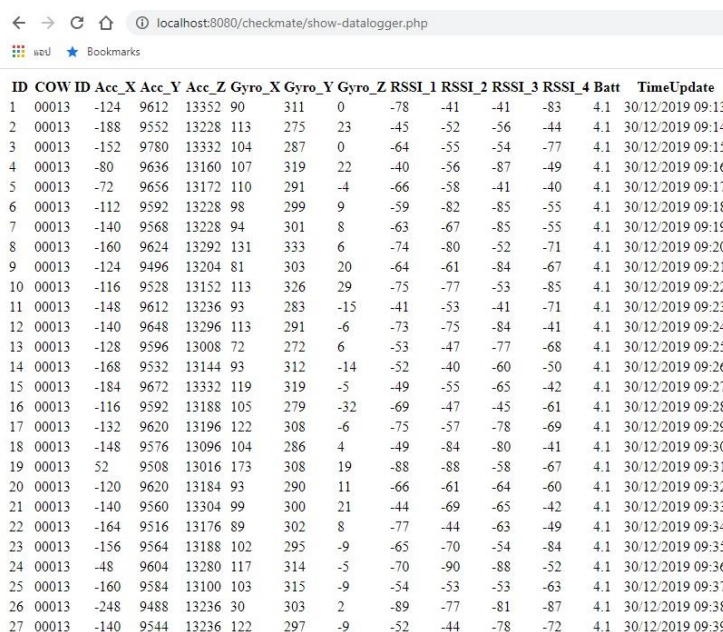
Figure 4 Controlled data collection of mounting behavior.

Controlled dummy-mounting enables precise identification of event onset and offset while minimizing confounding from overlapping social interactions [10, 24]. This phase ensures that mounting-specific motion signatures can be characterized prior to deployment under heterogeneous farm conditions. However, this controlled setting differs from natural cow-cow mounting in real herds, and this difference is explicitly considered as a limitation when interpreting generalizability.

2.3.2 Definition of mounting events and annotation

Direct field observation confirmed that individual mounting events typically last 3–5 seconds. Accordingly, a mounting event was defined as a continuous interval characterized by sustained physical contact with the dummy cow, a distinct postural change consistent with mounting, and elevated dynamic acceleration captured by the tri-axial accelerometer for a minimum duration of 3 seconds.

Ground-truth labels were constructed by synchronizing sensor timestamps with manual observation logs. Each interval was annotated at the event level as either mounting (positive class) or non-mounting (negative class). Event-level labels were subsequently propagated to window-level segments during preprocessing. An example of the resulting raw collar data records is shown in Figure 5.



ID	COW ID	Acc_X	Acc_Y	Acc_Z	Gyro_X	Gyro_Y	Gyro_Z	RSSI_1	RSSI_2	RSSI_3	RSSI_4	Batt	TimeUpdate
1	00013	-124	9612	13352	90	311	0	-78	-41	-41	-83	4.1	30/12/2019 09:13
2	00013	-188	9552	13228	113	275	23	-45	-52	-56	-44	4.1	30/12/2019 09:14
3	00013	-152	9780	13332	104	287	0	-64	-55	-54	-77	4.1	30/12/2019 09:15
4	00013	-80	9636	13160	107	319	22	-40	-56	-87	-49	4.1	30/12/2019 09:16
5	00013	-72	9656	13172	110	291	-4	-66	-58	-41	-40	4.1	30/12/2019 09:17
6	00013	-112	9592	13228	98	299	9	-59	-82	-85	-55	4.1	30/12/2019 09:18
7	00013	-140	9568	13228	94	301	8	-63	-67	-85	-55	4.1	30/12/2019 09:19
8	00013	-160	9624	13292	131	333	6	-74	-80	-52	-71	4.1	30/12/2019 09:20
9	00013	-124	9496	13204	81	303	20	-64	-61	-84	-67	4.1	30/12/2019 09:21
10	00013	-116	9528	13152	113	326	29	-75	-77	-53	-85	4.1	30/12/2019 09:22
11	00013	-148	9612	13236	93	283	-15	-41	-53	-41	-71	4.1	30/12/2019 09:23
12	00013	-140	9648	13296	113	291	-6	-73	-75	-84	-41	4.1	30/12/2019 09:24
13	00013	-128	9596	13008	72	272	6	-53	-47	-77	-68	4.1	30/12/2019 09:25
14	00013	-168	9532	13144	93	312	-14	-52	-40	-60	-50	4.1	30/12/2019 09:26
15	00013	-184	9672	13332	119	319	-5	-49	-55	-65	-42	4.1	30/12/2019 09:27
16	00013	-116	9592	13188	105	279	-32	-69	-47	-45	-61	4.1	30/12/2019 09:28
17	00013	-132	9620	13196	122	308	-6	-75	-57	-78	-69	4.1	30/12/2019 09:29
18	00013	-148	9576	13096	104	286	4	-49	-84	-80	-41	4.1	30/12/2019 09:30
19	00013	52	9508	13016	173	308	19	-88	-88	-58	-67	4.1	30/12/2019 09:31
20	00013	-120	9620	13184	93	290	11	-66	-61	-64	-60	4.1	30/12/2019 09:32
21	00013	-140	9560	13304	99	300	21	-44	-69	-65	-42	4.1	30/12/2019 09:33
22	00013	-164	9516	13176	89	302	8	-77	-44	-63	-49	4.1	30/12/2019 09:34
23	00013	-156	9564	13188	102	295	-9	-65	-70	-54	-84	4.1	30/12/2019 09:35
24	00013	-48	9604	13280	117	314	-5	-70	-90	-88	-52	4.1	30/12/2019 09:36
25	00013	-160	9584	13100	103	315	-9	-54	-53	-53	-63	4.1	30/12/2019 09:37
26	00013	-248	9488	13236	30	303	2	-89	-77	-81	-87	4.1	30/12/2019 09:38
27	00013	-140	9544	13236	122	297	-9	-52	-44	-78	-72	4.1	30/12/2019 09:39

Figure 5 Example of raw data records collected from collar-mounted sensing devices.

2.4 Temporal signal characteristics and dataset composition

Example accelerometer signal patterns during confirmed mounting events are illustrated in Figure 6. Sustained high-amplitude dynamic acceleration over approximately 3-5 seconds distinguishes mounting behavior from transient non-reproductive activities, motivating the formulation of mounting detection as a temporal pattern recognition problem.

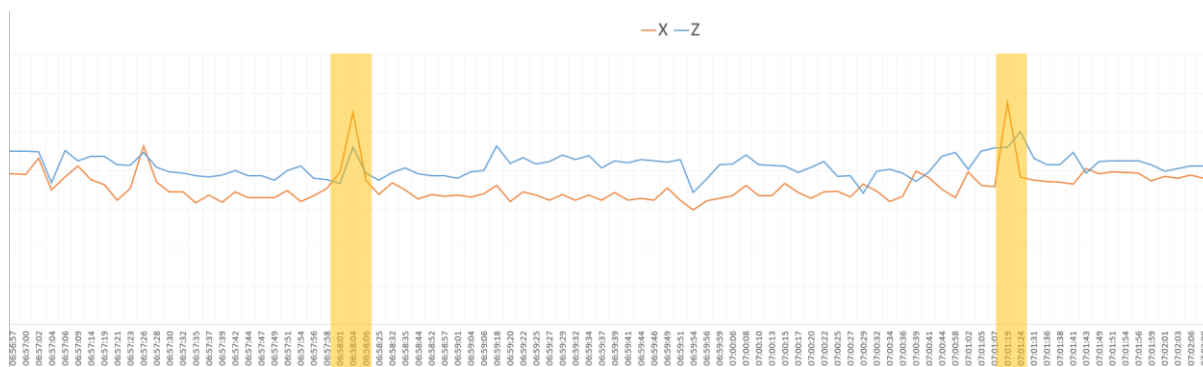


Figure 6 Example of tri-axial accelerometer (GY-61) signal patterns during a confirmed mounting event.

Across all sessions, the controlled dataset comprised 120 confirmed mounting events. The tri-axial accelerometer was sampled at approximately 25 Hz, while BLE RSSI measurements were recorded at approximately 1-s intervals. Controlled data collection was conducted across 10 semen-collection sessions over 5 days, yielding approximately 18.4 h of synchronized raw recordings prior to preprocessing. Using sliding-window segmentation, the annotated events yielded approximately 2,000 mounting windows and 6,000 non-mounting windows. Given the inherent class imbalance, subsequent model evaluation emphasizes event-level performance metrics in addition to window-level accuracy [25].

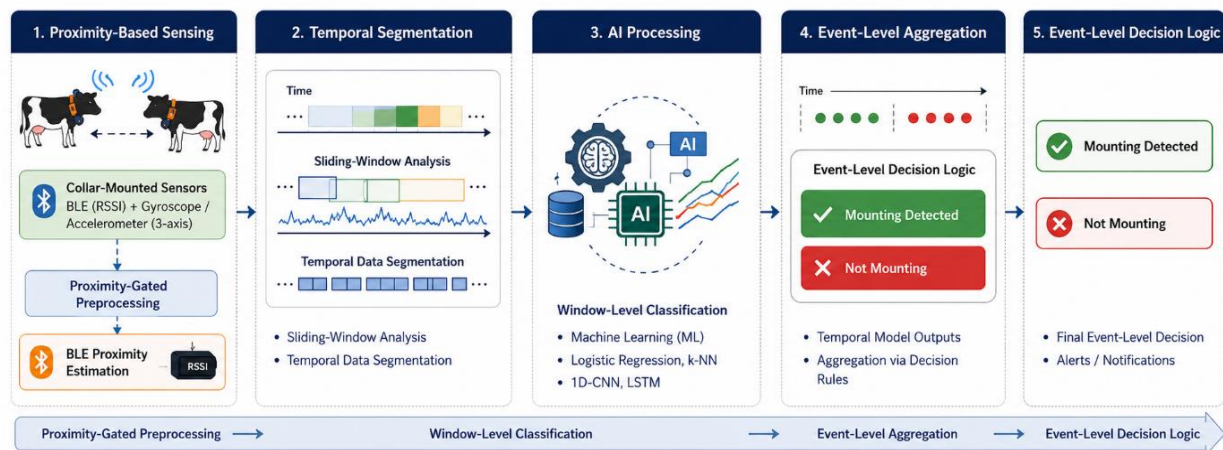


Figure 7 Overview of the AI development pipeline for proximity-gated mounting detection.

2.5 AI development pipeline overview

Figure 7 summarizes the AI development pipeline used in this study. The workflow comprised proximity-gated interaction filtering, temporal segmentation, model training using ML and DL approaches, and event-level aggregation aligned with the biological duration of mounting behavior. Detailed experimental settings are presented in Sections 2.7-2.12.

2.6 Field deployment protocol for estrus monitoring

In addition to controlled mounting data collection at the bull stud farm, field deployment was conducted to evaluate practical estrus-monitoring performance under real farm conditions. Following controlled calibration and comparative model evaluation, the best-performing model was selected for field deployment. The same proximity-gated AI framework described in Sections 2.1-2.5 was then deployed without structural modification to assess robustness and operational feasibility in heterogeneous farm environments.

A total of 71 collars were installed on breeding cows across three smallholder dairy farms in Sisaket and Buriram provinces, Thailand: Sanfun Farm (45 collars), Thong-In Farm (5 collars), and Homkhao Farm (21 collars). The system transmitted collar data through IoT gateways using Wi-Fi or cellular connectivity and delivered estrus alerts to farmers through the LINE application, with summary reports available through a farm dashboard. During deployment, estrus alerts were verified by farm owners and livestock officers, and subsequent return-to-estrus observations during the first expected post-insemination cycle were recorded for reproductive performance assessment.

2.7 Experimental design overview

All experiments were conducted using the same dataset, preprocessing pipeline, proximity-gated filtering mechanism, and data partitioning strategy described in Section 2. The objective was to isolate the effect of the learning algorithm while controlling for variability introduced by sensing, interaction validation, segmentation, and preprocessing procedures. Model performance was evaluated at two complementary levels: window-level classification and event-level detection.

Window-level classification evaluates the ability of models to discriminate mounting-related motion patterns from non-mounting activity within short temporal windows derived from proximity-gated accelerometer signals. Event-level detection evaluates biologically meaningful mounting events by aggregating window-level predictions over the natural duration of mounting behavior, which typically spans 3–5 seconds. This dual-level evaluation ensures that reported performance reflects both algorithmic discrimination capability and practical relevance for automated estrus detection, where decisions must correspond to complete behavioral events rather than isolated signal fragments [10, 23].

2.8 Data partitioning strategy

To prevent information leakage and overly optimistic performance estimates, data were partitioned at the session level, consistent with the event definitions described in Section 2.3.2. Window-level samples originating from the same mounting event were not allowed to appear across different subsets. The dataset was divided into 70% training data for model fitting, 15% validation data for hyperparameter tuning and early stopping, and 15% testing data reserved exclusively for final performance evaluation.

This partitioning strategy ensures that model performance reflects generalization to unseen mounting events rather than memorization of specific motion patterns. Event-level separation is particularly important in proximity-gated behavioral detection systems intended for deployment under heterogeneous farm conditions [26].

2.9 Training and validation protocol

For classical ML models, feature normalization parameters were computed using only the training set and subsequently applied to the validation and testing sets. Model hyperparameters were optimized using grid search or randomized search based on validation performance.

For DL models, training was performed using mini-batch gradient descent with early stopping based on validation loss to mitigate overfitting. Dropout regularization was applied where appropriate. Model checkpoints corresponding to the best validation performance were retained for testing.

No model-specific data augmentation, heuristic rule injection, or post hoc threshold tuning was applied. All models were trained and evaluated under identical proximity-gated constraints and preprocessing configurations, ensuring fairness in performance comparison and preserving methodological continuity with subsequent field deployment [17, 27].

To improve methodological transparency, the main implementation settings of the best-performing LSTM model and the computationally simpler GBDT classifier are reported here. For the LSTM model, each input sequence consisted of 100 time steps derived from a 4-s accelerometer window sampled at approximately 25 Hz, with three input features corresponding to the x-, y-, and z-axis acceleration signals. The LSTM architecture consisted of two recurrent layers with 64 and 32 units, respectively. The first LSTM layer returned the full temporal sequence, whereas the second LSTM layer returned the final hidden representation. Dropout regularization with a rate of 0.2 was applied after each LSTM layer. The recurrent representation was passed to a dense hidden layer with 16 units and ReLU activation, followed by a single sigmoid output unit for binary mounting probability estimation. The model was trained using the Adam optimizer with a learning rate of 0.001, binary cross-entropy loss, a batch size of 32, and early stopping based on validation loss with a patience of 10 epochs.

For the GBDT classifier, engineered features were extracted from each sliding accelerometer window. These features included axis-wise statistical descriptors from the x, y, and z-axis acceleration signals, including mean, standard deviation, maximum, minimum, root mean square, and signal-magnitude-related descriptors, resulting in 18 window-level features. Hyperparameters were selected using grid search on the validation set. The final GBDT configuration used 150 estimators, a learning rate of 0.05, maximum tree depth of 5, subsampling rate of 0.8, and log-loss as the optimization criterion.

2.10 Evaluation metrics

Given the inherent class imbalance between mounting and non-mounting samples, multiple evaluation metrics were employed. Overall classification accuracy was reported to provide a general measure of correctness. Precision quantified the proportion of detected mounting instances that were true positives, while recall (sensitivity) quantified the proportion of true mounting events correctly detected. The F1-score, defined as the harmonic mean of precision and recall, was used as a balanced performance metric.

Precision and recall are particularly important for automated estrus detection systems. False positives may lead to unnecessary alerts and interventions, whereas false negatives may result in missed estrus events and reduced reproductive efficiency [5, 25]. Accordingly, event-level metrics are emphasized as the primary indicators of biological and operational reliability.

2.11 Event-Level evaluation criteria

The event-level evaluation was adopted to align AI decisions with the natural duration of mounting events (3–5 seconds). Window-level predictions were aggregated into event-level decisions using a voting strategy within each annotated event interval.

An event-level detection was considered correct only if it occurred within the ground-truth event duration and within proximity-gated interaction windows, thereby enforcing both temporal consistency and spatial feasibility. This biologically constrained aggregation strategy reduces sensitivity to transient motion artifacts and sensor noise while preserving meaningful detection outcomes for automated estrus monitoring [25].

2.12 Consistency and fairness across models

To ensure statistical consistency and unbiased performance comparison, identical data partitions were used for all evaluated models. Preprocessing parameters, including proximity-gated filtering, window length, step size, and event aggregation rules, were held constant across experiments. No model-specific heuristics or post-processing techniques were applied.

Performance was reported using the same evaluation metrics at both window and event levels, with particular emphasis on event-level precision, recall, and F1-score, as these metrics directly reflect the operational reliability of automated estrus detection systems. For field-level proportions derived from deployment outcomes, exact 95% confidence intervals were additionally computed to support cautious interpretation. Formal statistical significance testing between models was not performed because the controlled evaluation was based on a single event-partitioned holdout design and the field deployment was intended primarily as system validation rather than hypothesis testing. This controlled and proximity-consistent experimental design nevertheless enables meaningful interpretation of accuracy-complexity trade-offs between ML and DL models and supports conclusions regarding their suitability for subsequent real-farm deployment.

2.13 Ablation analysis of proximity-gated inference

To quantify the independent contribution of BLE-based proximity gating, a retrospective ablation analysis was performed using the same controlled event-level test partition, LSTM classifier, preprocessing pipeline, model decision threshold, and event-level aggregation rule. The only difference between the two inference configurations was whether BLE proximity gating was applied before event-level decision making.

Tri-axial accelerometer signals sampled at approximately 25 Hz were segmented into 4-s windows with a 2-s step size, corresponding to 50% overlap. Each window therefore contained 100 accelerometer samples. BLE RSSI measurements were recorded at approximately 1-s intervals and synchronized with accelerometer windows. In the ungated motion-only configuration, all accelerometer windows classified as mounting by the LSTM model were retained as mounting candidates regardless of BLE proximity status. In the proximity-gated configuration, a window was retained as a mounting candidate only when it was classified as mounting by the LSTM model and satisfied the predefined BLE close-contact criterion used by the system.

Window-level predictions were converted into event-level decisions using majority voting over the biologically plausible mounting-duration interval. Adjacent positive event candidates separated by 2 s or less were merged into a single predicted event. Event-level true positives, false positives, and false negatives were determined by matching predicted events with manually annotated ground-truth mounting events. Each ground-truth event was matched to at most one predicted event. Predicted events without overlap with any ground-truth event were counted as false positives, whereas ground-truth events without a matched predicted event were counted as false negatives. Precision, recall, F1-score, and false-positive event reduction were used to compare gated and ungated inference.

3. Results

This section presents the results obtained during the controlled calibration phase and subsequent field deployment. All models were evaluated within the unified proximity-gated inference framework prior to and during real-farm implementation. Performance is reported at both window and event levels to ensure alignment between statistical classification accuracy and biologically meaningful mounting-event detection.

3.1 Controlled calibration results and model evaluation

3.1.1 Window-Level classification performance

Window-level classification evaluates the ability of each model to distinguish mounting-related motion patterns from non-mounting activity within short temporal windows derived from proximity-gated tri-axial accelerometer signals, as defined in Section 2.

Table 1 Window-level classification performance of machine learning and deep learning models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Logistic Regression (LR)	88.3	81.2	74.6	77.8
Decision Tree (DT)	86.9	79.5	72.1	75.6
k-Nearest Neighbors (kNN)	89.4	83.1	76.8	79.8
Support Vector Machine (RBF)	91.6	86.7	79.9	83.1
Random Forest (RF)	93.2	89.5	83.7	86.5
Gradient Boosting Decision Trees (GBDT)	94.1	91.3	85.6	88.4
1D Convolutional Neural Network (1D-CNN)	93.8	90.7	84.9	87.7
Long Short-Term Memory (LSTM)	94.0	91.0	86.2	88.5

As presented in Table 1, ensemble-based machine learning (ML) models and deep learning (DL) models consistently outperformed linear and shallow classifiers across all evaluation metrics. Logistic regression and single decision tree models exhibited comparatively lower recall values, indicating limited capacity to model non-linear and temporally distributed acceleration patterns associated with mounting behavior.

Among the classical ML approaches summarized in Table 1, Gradient Boosting Decision Trees (GBDT) achieved the highest window-level F1-score (88.4%), demonstrating strong discriminative power when operating on engineered accelerometer features. Deep learning models, including 1D Convolutional Neural Network (1D-CNN) and Long Short-Term Memory (LSTM), achieved comparable performance (87.7% and 88.5% F1-score, respectively) by learning temporal representations directly from raw accelerometer sequences [26, 27].

However, consistent with the evaluation rationale described in Section 2, window-level performance alone does not fully determine suitability for automated estrus detection, as isolated window predictions may not necessarily correspond to complete biologically meaningful mounting events. This limitation motivates the subsequent event-level aggregation analysis presented in Section 3.1.2.

3.1.2 Event-Level mounting detection performance

To align algorithmic outputs with biologically defined mounting behavior, window-level predictions were aggregated over 3–5 second intervals, consistent with the event-level criteria described in Section 2.

Event-level aggregation substantially improved detection performance across all evaluated models compared with window-level evaluation. This improvement confirms the importance of embedding biological duration constraints into AI decision logic rather than relying solely on fragment-level classification [10, 23].

Table 2 Event-level mounting detection performance of ML and DL models.

Model	Precision (%)	Recall (%)	F1-score (%)
Logistic Regression (LR)	86.1	78.3	82.0
Decision Tree (DT)	84.5	77.5	80.9
k-Nearest Neighbors (kNN)	88.2	81.7	84.8
Support Vector Machine (RBF)	91.6	85.8	88.6
Random Forest (RF)	94.1	89.2	91.6
Gradient Boosting Decision Trees (GBDT)	95.3	90.0	92.6
1D Convolutional Neural Network (1D-CNN)	94.6	90.8	92.7
Long Short-Term Memory (LSTM)	94.8	91.7	93.2

Among all evaluated models reported in Table 2, the LSTM achieved the highest event-level F1-score (93.2%), indicating the strongest overall performance in detecting complete mounting events under proximity-constrained evaluation. However, the performance margin relative to Gradient Boosting Decision Trees (GBDT) was modest (92.6%). Notably, GBDT achieved near-optimal event-level performance with substantially lower computational complexity, highlighting its practical suitability for real-time deployment within the proposed proximity-gated estrus detection framework.

3.1.3 Impact of BLE proximity-gated processing

A formal event-level ablation analysis was conducted to directly quantify the contribution of BLE proximity gating. The analysis compared the proposed proximity-gated inference configuration with an ungated motion-only configuration using the same LSTM

classifier, controlled test partition, decision threshold, preprocessing pipeline, and event-level aggregation rule. Therefore, the comparison isolates the effect of the BLE proximity gate on mounting-event detection.

Table 3 Event-level ablation analysis of LSTM inference with and without BLE proximity gating.

Config.	TP	FP	FN	Precision	Recall	F1	FP reduction
Motion-only (ungated)	112	48	8	70.0	93.3	80.0	Ref.
BLE proximity-gated	110	6	10	94.8	91.7	93.2	87.5

As shown in Table 3, the ungated motion-only configuration achieved high recall but produced many false-positive event candidates. Specifically, ungated LSTM inference detected 112 true mounting events but generated 48 false-positive events, resulting in a precision of 70.0%, recall of 93.3%, and F1-score of 80.0%. After applying BLE proximity gating, the number of false-positive events decreased from 48 to 6, corresponding to an 87.5% false-positive event reduction. Precision increased substantially from 70.0% to 94.8%, while recall decreased only slightly from 93.3% to 91.7%. Consequently, the event-level F1-score improved from 80.0% to 93.2%. These results demonstrate that BLE proximity gating contributed directly to false-positive suppression rather than serving only as a conceptual component of the system architecture. The slight reduction in recall suggests that a small number of true mounting events may have been excluded by proximity filtering, possibly due to short-term BLE signal fluctuation or collar orientation effects. However, the large reduction in false-positive events indicates that requiring both motion evidence and close-contact feasibility substantially improves the biological specificity of mounting-based estrus detection.

3.1.4 Accuracy–Complexity trade-off between ML and DL models

While DL models, particularly LSTM, achieved the highest event-level detection performance, their advantage over advanced ML models remained limited under the controlled proximity-gated setting. Given the short-duration and structured nature of mounting behavior, engineered-feature-based models such as Gradient Boosting Decision Trees effectively captured discriminative acceleration patterns without requiring deep temporal representation learning.

Considering training stability, inference latency, and deployment simplicity within IoT-based livestock monitoring systems, GBDT represents a favorable trade-off between accuracy and computational complexity. In contrast, DL models may be preferred in scenarios involving larger datasets, higher behavioral variability, or richer multimodal integration [17, 28].

Based on this comparative evaluation, the LSTM model, which achieved the highest event-level F1-score (93.2%) under the proximity-gated configuration, was selected for subsequent field deployment.

3.2 Field deployment results

To evaluate real-world applicability, the LSTM model selected from the controlled phase was deployed without farm-specific retraining across three smallholder dairy farms in Sisaket and Buriram provinces, Thailand. A total of 71 estrus-monitoring collars were installed: 45 at Sanfun Farm, 5 at Thong-In Farm, and 21 at Homkhao Farm.

3.2.1 Field detection performance

During the monitoring period, the platform generated estrus alerts for 23 dairy cows. All 23 cases were subsequently verified by farm owners or livestock officers and confirmed as true estrus events. No false-positive estrus alerts were recorded during the observation period; all 23 alerts corresponded to confirmed estrus events within this dataset. Given the limited field sample, this finding is reported descriptively in counts rather than emphasized as a percentage estimate; the exact 95% lower confidence bound for precision was 85.2%.

In parallel, farm observation identified 8 additional cows that were in estrus but were not detected by the system. According to farmer reports, most of these cows exhibited silent estrus, characterized by reduced or absent mounting behavior despite underlying ovarian cyclicity [3, 29]. Silent estrus is widely reported in dairy cattle, particularly under metabolic stress, environmental challenges, or health disorders [3, 30].

Because the proposed framework relies on proximity-confirmed mounting behavior as its primary detection criterion, estrus cases without overt behavioral expression were not identifiable by design. Considering both detected and undetected estrus cases, the total number of confirmed estrus events during the study period was 31. Based on this value, the field-level sensitivity (recall) of the system was 74.2% (23/31; exact 95% CI: 55.4–88.1%).

The reduction in sensitivity was therefore more plausibly attributable to biological expression variability than to observed instability in the classification model and should be interpreted as a limitation of behavior-dependent estrus monitoring rather than as evidence of unstable inference under the present deployment conditions.

3.2.2 Reproductive management outcomes

Among the 23 cows detected by the system, 13 underwent artificial insemination (AI), while 10 were not inseminated. The primary reason for withholding insemination was lumpy skin disease infection, which required veterinary management and posed potential reproductive risks. Farmers therefore postponed insemination for affected animals. Post-insemination follow-up was conducted during the first expected return-to-estrus cycle.

Of the 13 inseminated cows, 11 cows (C001, C003, C004, C005, C007, C008, C009, C010, C011, C012, and C013) did not exhibit a subsequent estrus event during the first expected post-insemination cycle follow-up. Two cows (C002 and C006) returned to estrus on days 22 and 24 after AI, respectively, indicating return to the next estrous cycle rather than an immediate false alert. Accordingly, the first-cycle non-return proportion was calculated as $11/13 \times 100 = 84.6\%$ (exact 95% CI: 54.6–98.1%). This metric was used as a practical field proxy for reproductive outcome and should not be interpreted as a confirmed pregnancy rate.

Although non-return rate does not replace confirmatory pregnancy diagnosis through ultrasonography or hormonal assays, it remains an established and operationally feasible fertility metric in routine dairy practice [31]. The observed non-return rate suggests

that system-generated estrus alerts were biologically meaningful and appropriately timed for insemination decision-making, but it should not be interpreted as a confirmed pregnancy rate [32].

These results indicate that the proposed platform can function not only as a behavioral estrus detection system but also as a practical reproductive decision-support tool under smallholder farm conditions, while acknowledging that the reproductive evidence remains preliminary.

4. Discussion

4.1 Controlled-versus-field performance and precision-oriented design

The controlled experimental phase demonstrated strong mounting classification capability, achieving an event-level F1-score of 93.2% under the proximity-gated LSTM configuration. During field deployment, sensitivity decreased to 74.2% (95% CI: 55.4-88.1%), reflecting the expected gap between structured validation and heterogeneous farm operation in precision livestock systems. No false-positive alerts were observed during the study period in this dataset, so field precision is interpreted descriptively rather than as a definitive population parameter.

A central design objective of the proposed framework was to prioritize high precision, because false estrus alerts can increase insemination costs and reduce user trust. The ablation analysis provides quantitative evidence for this design rationale. Compared with ungated motion-only inference, BLE proximity-gated inference reduced false-positive event candidates by 87.5% and improved precision from 70.0% to 94.8%. This result indicates that many abrupt acceleration patterns that resembled mounting at the motion-signal level were not biologically plausible when inter-animal proximity was considered. Therefore, the proximity gate improved the biological specificity of the detection pipeline by requiring both motion evidence and interaction feasibility before generating event-level mounting decisions.

4.2 Biological limitation of behavior-dependent estrus detection

The eight undetected estrus events were categorized as silent estrus cases. Silent estrus is widely documented in dairy cattle and is characterized by minimal or absent overt mounting behavior despite ovulation [3, 29]. Because the proposed framework relies on observable mounting interactions, detection capability is inherently constrained by behavioral expressivity.

These missed detections should therefore be interpreted as biological limitations of behavior-based estrus monitoring rather than algorithmic instability. Additional limitations also warrant caution: field validation covered three farms within a specific geographic region and a relatively small number of confirmed estrus events, reproductive outcomes were assessed using a first-cycle non-return proportion as a practical proxy rather than confirmed pregnancy diagnosis, and the study did not include longitudinal multi-cycle monitoring. Broader multi-site, multi-season, and longitudinal validation is required to assess generalizability across herd sizes, housing systems, environmental conditions, and management practices.

4.3 Practical implications and study limitations

Scalable smart farming systems require contextual validation beyond signal classification alone. By formalizing estrus detection as a socially constrained behavioral event rather than a single-animal activity spike, the proposed proximity-gated framework supports biologically interpretable and farmer-oriented decision support in resource-constrained dairy environments.

Practical deployment of the proposed framework requires attention to device maintenance, data connectivity, and user training. Collar-mounted devices should be checked routinely for fit, physical damage, battery condition, and sensor orientation because loose collars or rotated sensor modules may affect accelerometer signal quality. Gateway placement is also important because BLE RSSI values are sensitive to animal body obstruction, pen structure, multipath propagation, and temporary signal fluctuation. In this study, collar data were transmitted through an IoT gateway using Wi-Fi or cellular connectivity before backend inference and alert generation. Therefore, reliable network coverage and local buffering are important for minimizing data loss during temporary connectivity interruptions.

From the user perspective, farmers need simple training on LINE alert interpretation, dashboard review, visible estrus confirmation, insemination decision-making, and reporting of health conditions such as lumpy skin disease. These operational factors are particularly important for smallholder farms, where ease of maintenance, trust in alerts, and low management burden strongly influence adoption.

5. Conclusions

This study tested the hypothesis that mounting-based estrus detection can be improved by constraining accelerometer-based motion classification to biologically feasible inter-animal proximity intervals. The results support this hypothesis. In the controlled evaluation, the proximity-gated LSTM configuration achieved an event-level F1-score of 93.2%. More importantly, the ablation analysis showed that BLE proximity gating reduced false-positive event candidates from 48 to 6 compared with ungated motion-only inference, increasing precision from 70.0% to 94.8% and improving the F1-score from 80.0% to 93.2%. These findings indicate that the main contribution of the proposed framework lies not only in AI-based motion classification but also in interaction-aware validation of biological plausibility.

Field deployment across three smallholder dairy farms demonstrated operational feasibility using 71 collar-mounted devices, with 23 of 31 confirmed estrus events detected and no false-positive alerts observed within the limited field dataset. Post-insemination follow-up showed that 11 of 13 inseminated cows did not return to estrus during the first expected post-insemination cycle follow-up, corresponding to a first-cycle non-return proportion of 84.6%. This reproductive outcome should be interpreted as a field proxy rather than confirmed pregnancy.

Several limitations should be acknowledged. First, the controlled training dataset was collected from breeding bulls mounting a stationary dummy cow, which enabled accurate annotation but may not fully represent the movement dynamics, acceleration forces, and behavioral variability of natural cow-cow mounting in real farm environments. Second, field validation was limited to three farms in two Thai provinces and involved a relatively small number of confirmed estrus events. Third, the system depends on overt mounting behavior and therefore may miss silent estrus cases. Future work should include larger multi-site and multi-season validation, natural

cow-cow mounting datasets, longitudinal reproductive follow-up with confirmatory pregnancy diagnosis, and multimodal integration with additional physiological or behavioral indicators.

6. Acknowledgements

The authors gratefully acknowledge the Dairy Farming Promotion Organization of Thailand (DPO) for their valuable support in facilitating the collection of mounting behavior data and coordinating field activities. The authors also sincerely thank Sanfun Farm, Thong-In Farm, and Homkhao Farm for serving as pilot sites for system deployment and field validation. The cooperation of the farm owners and staff in device installation, animal monitoring, data collection, and reproductive record verification was essential to the successful completion of this study.

7. References

- [1] Walsh SW, Williams EJ, Evans ACO. A review of the causes of poor fertility in high milk producing dairy cows. *Anim Reprod Sci.* 2011;123(3-4):127-38.
- [2] Aungier SPM, Roche JF, Sheehy M, Crowe MA. Effects of management and health on the use of activity monitoring for estrus detection in dairy cows. *J Dairy Sci.* 2012;95(5):2452-66.
- [3] Roelofs J, López-Gatius F, Hunter RHF, van Eerdenburg FJCM, Hanzen C. When is a cow in estrus? Clinical and practical aspects. *Theriogenology.* 2010;74(3):327-44.
- [4] Firk R, Stamer E, Junge W, Krieter J. Automation of oestrus detection in dairy cows: a review. *Livest Prod Sci.* 2002;75(3):219-32.
- [5] Saint-Dizier M, Chastant-Maillard S. Towards an automated detection of oestrus in dairy cattle. *Reprod Domest Anim.* 2012;47(6):1056-61.
- [6] Rutten CJ, Velthuis AGJ, Steeneveld W, Hogeveen H. Invited review: Sensors to support health management on dairy farms. *J Dairy Sci.* 2013;96(4):1928-52.
- [7] Valenza A, Giordano JO, Lopes G, Vincenti L, Amundson MC, Fricke PM. Assessment of an accelerometer system for detection of estrus and treatment with gonadotropin-releasing hormone at the time of insemination in lactating dairy cows. *J Dairy Sci.* 2012;95(12):7115-27.
- [8] Holman A, Thompson J, Routly JE, Cameron J, Jones DN, Grove-White D, et al. Comparison of oestrus detection methods in dairy cattle. *Vet Rec.* 2011;169(2):47.
- [9] Van Eerdenburg FJCM, Loeffler HSH, van Vliet JH. Detection of oestrus in dairy cows: a new approach to an old problem. *Vet Q.* 1996;18(2):52-4.
- [10] Kerbrat S, Disenhaus C. A proposition for an updated behavioural characterisation of the oestrus period in dairy cows. *Appl Anim Behav Sci.* 2004;87(3-4):223-38.
- [11] Senger PL. Pathways to pregnancy and parturition. 3rd ed. Redmond: Current Conceptions Inc.; 2012.
- [12] Roelofs JB, van Eerdenburg FJCM, Soede NM, Kemp B. Various behavioral signs of estrous and their relationship with time of ovulation in dairy cattle. *Theriogenology.* 2005;63(5):1366-77.
- [13] Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: a review. *Sensors.* 2018;18(8):2674.
- [14] Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: a survey. *Comput Electron Agric.* 2018;147:70-90.
- [15] Riaboff L, Shaloo L, Smeaton AF, Couvreur S, Madouasse A, Keane MT. Predicting livestock behaviour using accelerometers: a systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data. *Comput Electron Agric.* 2022;192:106610.
- [16] Nasirahmadi A, Edwards SA, Sturm B. Implementation of machine vision for detecting behaviour of cattle and pigs. *Livest Sci.* 2017;202:25-38.
- [17] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
- [18] Chen C, Zhu W, Norton T. Behaviour recognition of pigs and cattle: journey from computer vision to deep learning. *Comput Electron Agric.* 2021;187:106255.
- [19] Mahmud MS, Zahid A, Das AK, Muzammil M, Khan MU. A systematic literature review on deep learning applications for precision cattle farming. *Comput Electron Agric.* 2021;187:106313.
- [20] Wolfert S, Ge L, Verdouw C, Bogaardt MJ. Big data in smart farming - a review. *Agric Syst.* 2017;153:69-80.
- [21] Paek J, Kim J, Govindan R. Energy-efficient rate-adaptive GPS-based positioning for smartphones. *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys); 2010 Jun 15-18; San Francisco, USA.* New York: ACM; 2010. p. 299-314.
- [22] Stopczynski A, Stahlhut C, Larsen JE, Petersen MK, Hansen LK. The smartphone brain scanner: a portable real-time neuroimaging system. *PLoS ONE.* 2014;9(2):e86733.
- [23] Martiskainen P, Järvinen M, Skön JP, Tiirikainen J, Kolehmainen M, Mononen J. Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. *Appl Anim Behav Sci.* 2009;119(1-2):32-8.
- [24] Guo Y, Zhang Z, He D, Niu J, Tan Y. Detection of cow mounting behavior using region geometry and optical flow characteristics. *Comput Electron Agric.* 2019;163:104828.
- [25] Provost F, Fawcett T. Robust classification for imprecise environments. *Machine Learning.* 2001;42(3):203-31.
- [26] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2009.
- [27] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell.* 2013;35(8):1798-828.
- [28] Breiman L. Random forests. *Machine Learning.* 2001;45(1):5-32.
- [29] Diskin MG, Sreenan JM. Expression and detection of oestrus in cattle. *Reprod Nutr Dev.* 2000;40(5):481-91.
- [30] López-Gatius F. Is fertility declining in dairy cattle? a retrospective study in northeastern Spain. *Theriogenology.* 2003;60(1):89-99.
- [31] Stevenson JS. Reproductive management of dairy cows in high milk-producing herds. *J Dairy Sci.* 2001;84:E128-43.
- [32] Santos JEP, Thatcher WW, Chebel RC, Cerri RLA, Galvão KN. The effect of embryonic death rates in cattle on the efficacy of estrus synchronization programs. *Anim Reprod Sci.* 2004;82-83:513-35.