IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 11 December 2025, accepted 24 December 2025, date of publication 29 December 2025, date of current version 2 January 2026.

Digital Object Identifier 10.1109/ACCESS.2025.3648965

RESEARCH ARTICLE

A Machine Learning Framework for Large-Scale Behavioral Monitoring to Assess Thermotolerance in Dairy Cows

NSEABASI S. UDONDIAN[1], DANIEL OGBUIGWE[1], MOSTAFA MAHMOUD SELIM[1], AHMED ABDELMOAMEN AHMED[1], (Member, IEEE), MADISON E. BACON[2], AND COURTNEY L. DAIGLE[2]

[1]Department of Computer Science, Prairie View A&M University, Prairie View, TX 77446, USA
[2]Department of Animal Science, Texas A&M University, College Station, TX 77840, USA

Corresponding author: Ahmed Abdelmoamen Ahmed (amahmed@pvamu.edu)

This work was supported in part by the National Science Foundation (NSF) under Grant 2011330, Grant 2200377, and Grant 2302469; and in part by the Agriculture and Food Research Initiative–Foundational and Applied Science Program, through U.S. Department of Agriculture's National Institute of Food and Agriculture, under Award 2022-08297.

ABSTRACT Based on projected climatic trends, the U.S. dairy industry faces escalating heat stress challenges, with milk production losses projected to reach 1.4 kg/day per cow by the 2050s and 1.9 kg/day per cow by the 2080s, resulting in economic losses of $1.7 billion and $2.2 billion, respectively. While extensive cooling mechanisms are in place, they cannot fully mitigate heat stress effects. Thermotolerance, influenced by physiological and behavioral traits, varies significantly among individual cows and is crucial for sustaining dairy productivity. This paper presents an approach for automatically monitoring and recognizing behavior for dairy cows potentially relevant to thermotolerance, which utilizes various Machine Learning (ML) (e.g., CNN and YOLOv8) and Computer Vision (CV) (e.g., Video Processing and Annotation) techniques aimed at quantifying key behavioral indicators, specifically drinking frequency, brush use-behaviors, and social interactions such as headbutting. We evaluated the developed system using two different video datasets. The first dataset was captured using cameras installed at a side angle, while the second dataset was obtained using top-down camera views. Evaluation results showed that the YOLOv8 and CNN models achieved classification accuracies of 93% and 96%, respectively, on the first dataset, while the second dataset yielded 94% for YOLOv8 and 96% for CNN. Beyond experimental validation, the model is currently deployed for large-scale analysis of over 38,000 video recordings from a commercial ranch in Texas. This deployment enables automated extraction of behavioral event counts, specifically drinking and brushing, for each video, providing high-resolution behavioral summaries at scale. These findings show the robustness and operational viability of the system. The insights gained from this work hold significant promise for advancing genetic selection and management practices, equipping the dairy industry to proactively address heat stress and enhance overall herd resilience in the face of evolving climatic challenges.

INDEX TERMS Heat stress, dairy cows, computer vision (CV), machine learning (ML), behavior monitoring, thermotolerance.

I. INTRODUCTION

The dairy industry continues to face growing challenges due to climate change, with heat stress emerging as a critical factor affecting dairy cattle worldwide [1], [2]. In the United States, projected climatic trends suggest significant declines in milk production due to heat stress, with losses expected to reach 1.4 kg/day per cow by the 2050s and 1.9 kg/day per cow by the 2080s. These reductions translate into economic losses of approximately $1.7 billion and $2.2 billion annually.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Soufiane Jouini.

Despite widespread implementation of cooling systems, such interventions remain insufficient to fully mitigate the adverse effects of heat stress on dairy herds [3]. Thermotolerance—cattle's ability to cope with heat stress–is shaped by both physiological and behavioral factors, which are heritable yet vary significantly between individuals [4].

Traditional approaches to monitoring heat stress rely heavily on manual observation, a labor-intensive, subjective, and often reactive method. In the face of rising global demand for dairy products, there is a pressing need for automated, scalable, and precise behavior monitoring tools [5]. Artificial intelligence (AI) [6], notably Computer Vision (CV) [7] and Machine Learning (ML) [8], has emerged as a transformative solution for this challenge [9]. This paper proposes an automated, large-scale behavioral monitoring approach for dairy cows, potentially relevant to thermotolerance [10], [11], utilizing CV and ML techniques. The developed system tracks and quantifies key behavioral indicators, including drinking frequency and brush use, as well as detects social interactions among cows (e.g., headbutting). The system demonstrated high accuracy in detecting these behaviors, validating the feasibility of automated monitoring for dairy cows.

This paper introduces a robust ML-based cow identification model trained using coat pattern analysis [12] utilizing a video dataset collected at the T&K Dairy, a commercial dairy partner in Snyder, Texas. In particular, we collected two different video datasets with more than 38,000 videos. The first dataset was captured using CCTV cameras oriented at a side angle, while the second was captured using top-down cameras. These datasets were compiled using various disparate data collection strategies involving 288 focal cows, specifically those between 45 and 90 Days in Milk. A focused subset of cows (n = 96; 16 cows per pen across 6 pens) was selected for intensive monitoring.

We developed an end-to-end video processing pipeline that allows system users to upload the captured videos to an interactive Graphical User Interface (GUI) developed using ReactJS, HTML, and CSS. The GUI enables users to interact with the system and visualize cow behavior trends conveniently. The developed system processes the videos, identifies individual cows using coat pattern analysis, detects drinking, brushing, and headbutting social behaviors, and visualizes the results in a web-based dashboard that displays behavior statistics such as frequency and duration. We utilized YOLOv8 [13] for target object detection and DeepSORT for behavior tracking [14], providing real-time and actionable insights for farm management.

We conducted a set of experiments to evaluate the developed system regarding inference accuracy and performance. The experimental evaluation demonstrated that the YOLOv8 and CNN models attained classification accuracies of 93% and 96%, respectively, on the side-angle dataset. On the top-down angle dataset, the models achieved accuracies of 94% for YOLOv8 and 96% for CNN. To rigorously assess the behavioral detection capability of the system, a cohort of human observers manually reviewed 457 video samples and compared their annotations with the system's outputs. The analysis indicated that the system demonstrated a high level of accuracy in detecting the target cow behaviors.

The remainder of this paper is organized as follows. Section II reviews relevant prior research in automated livestock monitoring. Section III outlines the system design, covering data collection protocols and core algorithmic frameworks. Section IV details the implementation, including the hardware and software environments. Section V presents a comprehensive evaluation of the system's performance, and finally, Section VI provides concluding remarks and discusses future work.

II. BACKGROUND AND RELATED WORK

Recent advances in animal science have increasingly underscored the influence of climate change on livestock systems, with heat stress recognized as one of the most critical environmental stressors affecting dairy cattle. A substantial body of evidence demonstrates that heat stress not only diminishes milk yield but also compromises reproductive performance and elevates disease susceptibility, thereby contributing to considerable economic losses [1], [3]. Conventional monitoring methods, which typically rely on manual observation, are labor-intensive and subjective, rendering them impractical for large-scale herd management. This limitation has stimulated growing interest in developing automated technologies capable of delivering timely and objective animal welfare assessments. Initial research in this domain employed approaches such as infrared thermography to identify physiological manifestations of heat stress [15].

More recent investigations have integrated ML and CV techniques to automate the detection and quantification of thermoregulatory behaviors. For example, automated tracking systems have been implemented in swine to monitor feeding and social interactions, demonstrating the capacity of computer vision to detect subtle indicators of welfare [16]. Integrating Deep Learning (DL) methodologies has markedly advanced the analysis of livestock behavior, with cattle serving as a primary focus of investigation. Architectures such as Convolutional Neural Networks (CNNs), YOLO-based object detectors, and Squeeze-and-Excitation Networks (SENets) have demonstrated high accuracy in both identifying individual animals within large herds and classifying behavioral patterns in real time [17], [18].

Compared to conventional feature-engineered approaches, these ML-based systems offer superior performance and have facilitated the development of scalable, continuous monitoring systems for precision livestock farming. However, most existing implementations have been validated primarily under controlled experimental conditions characterized by favorable lighting and limited dataset complexity, raising concerns regarding their robustness in heterogeneous, real-world farm environments. Moreover, current systems predominantly prioritize object detection tasks rather than end-to-end behavioral analysis pipelines, constraining their capacity to quantify behavior duration, frequency, and transitions. Overcoming these limitations necessitates the development of models that jointly capture spatial and temporal dynamics while maintaining computational efficiency suitable for on-farm deployment.

Substantial progress has been achieved in the automation of dairy cattle identification, with particular emphasis on utilizing biometric features such as coat pattern recognition. Deep learning frameworks, including Region-based Convolutional Neural Networks (R-CNNs), have been successfully applied to static imagery and aerial footage captured by unmanned aerial vehicles, yielding high identification accuracy in Holstein Friesian cattle [12]. More sophisticated pipelines integrating YOLOv5 for object detection with VGG16 for feature extraction, followed by classification via Support Vector Machines (SVMs), have reported near-perfect accuracy [19]. Despite these advancements, most existing systems remain dependent on high-resolution close-range imagery or constrained acquisition conditions, which are often impractical in commercial farm settings where crowding, occlusion, and variable lighting present significant challenges. Consequently, the large-scale deployment of such methods in operational dairy barns has been limited. These constraints highlight the need for robust and scalable identification approaches to maintain high accuracy under heterogeneous, real-world environmental conditions.

Inadagbo et al. [2] developed a CV-based system for detecting drinking and brushing behaviors, proposed as key indicators of thermotolerance in dairy cattle. Although this system achieved high levels of accuracy, its generalizability across heterogeneous farm environments was limited, and its scope did not extend to the recognition of a broader spectrum of social behaviors. This paper seeks to address these limitations by advancing analytical capabilities beyond simple resource-use monitoring to encompass agonistic social interactions –such as headbutting—which offer more comprehensive insights into herd dynamics and animal welfare [20], [21].

Beyond object-level identification (e.g., individual cow recognition), a growing body of research has examined behavior recognition tasks such as lying, standing, and drinking. For example, Unmanned Aerial Vehicle (UAV)-based monitoring systems have been developed to capture cattle postures and enable real-time detection of standing behavior [22]. Similarly, the Siamese Group Chunking Capsule Network (SGCCN) has been proposed as a computationally efficient architecture that reduces parameter requirements, thereby improving model performance in limited-data settings [23]. Despite these contributions, current approaches face practical limitations: UAV platforms are unsuitable for indoor barn environments, and many systems have not been systematically validated on large-scale, continuous datasets. As a result, their capacity to generalize to the scale, variability, and environmental complexity of commercial dairy operations remains insufficiently demonstrated.

Recent studies have increasingly explored livestock monitoring frameworks, underscoring their potential and inherent limitations [24]. In [25], the authors introduced a recurrent learning scheme for temporal modeling of behaviors, but the framework neglected spatial interactions between cattle and barn resources. Tsai et al. [26] designed an embedded vision system to monitor drinking as an indicator of heat stress; however, the system did not encompass additional behaviors such as brushing or social interactions. Similarly, Qiao et al. [27] employed YOLOv5-ASFF for cattle body detection, achieving high localization accuracy but without integration of behavior classification. These contributions reveal a persistent gap: while effective for specific and isolated tasks, existing pipelines rarely integrate individual identification, multi-behavior classification, and real-time scalability within a unified monitoring system.

Several recent approaches aimed to integrate spatial and temporal reasoning in livestock monitoring. For instance, Fuentes et al. [28] introduced a hierarchical deep learning framework for cattle behavior classification; however, the system lacked real-time applicability and did not preserve individual identity. Wang et al. [29] investigated open-set recognition to ensure identity consistency under variable conditions, yet this contribution was confined to identity verification and did not extend to behavior recognition. Yu et al. [30] applied YOLO-based edge inference for detecting feeding behavior, demonstrating feasibility for embedded deployment, but the scope was restricted to feeding alone. Islam et al. [31] developed a vision-based system to monitor drinking behavior in beef cattle, though its narrow behavioral coverage limits broader applicability. Fuentes et al. [32] proposed a multi-view learning framework enabling three-dimensional action recognition, but the reliance on complex multi-camera infrastructure poses challenges for scalability. More recently, Jia et al. [33] adapted YOLOv8n for behavior recognition in naturalistic barn conditions, yet without integration of tracking or long-term behavior logging. These studies highlight that despite methodological advances, no existing system provides a comprehensive and practically deployable solution for multi-behavior monitoring in commercial herds.

More closely aligned with the present study, Porto et al. [20] employed pose estimation in combination with long short-term memory (LSTM) sequence models to detect social behaviors such as headbutting and grooming. While methodologically sophisticated, pose-estimation models are computationally demanding and prone to reduced reliability when applied to low-quality farm video footage. Nasirahmadi et al. [21] investigated herd-level movement patterns and clustering dynamics, yet their system lacked the capacity to identify or monitor individual animals. Cheng et al. [34] integrated YOLOv5 with DeepSORT for behavior detection based on fixed proximity thresholds; however, this approach relied on rigid heuristics that overlooked the temporal variability inherent in behavioral sequences.

Collectively, these studies offer valuable contributions, but their respective limitations –ranging from narrow behavioral scope to computational inefficiency and the absence of individual-level tracking– underscore the need for more integrative and scalable solutions.

In summary, existing research underscores the potential of DL for cattle monitoring but also highlights enduring challenges related to generalizability, behavioral coverage, and real-time applicability. This paper seeks to address these gaps through the development of an integrated framework that combines three key advances: (i) multi-behavior detection encompassing drinking, brushing, and headbutting; (ii) reliable individual identification via coat-pattern recognition; and (iii) scalable real-time video processing suitable for deployment in commercial barn environments. By consolidating these components within a unified system, this work advances precision livestock farming by providing a practical and extensible platform for monitoring animal welfare under climate stress conditions.

III. SYSTEM DESIGN

This section outlines the methodological framework adopted for developing and evaluating the proposed system designed to characterize dairy cow phenotypes potentially associated with thermotolerance through the application of CV and ML techniques. It presents a comprehensive account of the experimental design, data acquisition procedures, and the architecture and implementation of the underlying artificial intelligence models and algorithms.

As shown in Figure 1, the developed system is designed as a comprehensive full-stack pipeline to enable seamless interaction between users and the AI-powered system via an interactive GUI. This includes an application layer for backend services, integrated AI models for core analysis, and a database for persistent data storage. The system architecture of the proposed approach is divided into five layers.

Layer 1 represents the data collection stage, during which video recordings were obtained using both top-down and side-angle camera perspectives. Layer 2 corresponds to the video preprocessing stage, wherein the recorded datasets were segmented into individual frame images at predefined intervals using a Python 3.13 script in conjunction with the FFmpeg framework [35]. Subsequently, the target objects were extracted to facilitate the training of the cow clustering algorithm and Convolutional Neural Network (CNN) model. Object annotation for cows, brushing tools, and waterers was performed using the Roboflow platform [36], which supported the development of both detection and segmentation models.

Layer 3 encompasses the core analytical modules, including the cow detection and segmentation component implemented with the YOLOv8 model, the cow clustering component utilizing the K-means algorithm, and the cow identification component employing a CNN and a Squeeze-and-Excitation Network (SENet) architectures [13], [37]. Layer 4 pertains to the application layer, developed using the Python Flask framework [38] in conjunction with the SQLite 3.50.4 database engine [39], which supports the construction of a web-based GUI application. Finally, Layer 5 represents the end-user interaction layer, enabling system users to access and operate the system with enhanced usability and convenience.

A. DATA COLLECTION AND PRE-PROCESSING

Cows were housed in a tunnel-ventilated barn (118 $m\times$ 115 m) designed to accommodate 1100 Holstein, Jersey, and crossbred dairy cows in six free-stall pens (approximately 180 cows/pen). Each pen provided access to three mechanical rotating cattle brushes and four water troughs. A commercially available fogging system and an additional sprinkler system were activated for heat stress mitigation when ambient temperatures reached 26.7^{o}C. An automatic robotic milking system with 18 stalls was also present in the barn. The videos were recorded using 8MP PoE bullet cameras, GW Security Inc., with four cameras installed per pen above the free stalls, angled to capture the water troughs and rotating brushes –see Figures 2 and 3.

The cameras were connected to PoE (Power over Ethernet) boxes and a network video recorder, with video saved to 8TB internal hard disk drives that were replaced every two months. The side-angle cameras were mounted at a height of approximately 2.2 meters from the ground, positioned laterally at an angle of about 45 degrees relative to the cows' movement path. The top-down cameras were installed at a height of approximately 3.5 meters, oriented perpendicular to the ground (90 degrees) to capture overhead views of the animals. These configurations were selected to ensure optimal visibility of behavioral indicators such as drinking, brushing, and social interactions, while minimizing occlusion and maximizing coverage of the activity zones.

We collected two different datasets, containing CCTV videos, from side and top-down angles. The side-angle dataset, shown in Figure 4(a), comprised of 154,368 videos captured from a side-angle perspective, of which 3,584 were processed for analysis. The top-down dataset, shown in Figure 4(b), consisted of 150,000 videos recorded from overhead camera views, with 38,152 processed. This selection was driven primarily by the need for manual annotation, which is highly time- and labor-intensive. To mitigate potential selection bias, videos were randomly sampled across different days, times, and animal groups to ensure diversity in behavioral expression and environmental conditions. Figure 5 shows representative samples from both video datasets.

1) DATA PRE-PROCESSING AND ANNOTATION

Figure 6 illustrates the dataset preparation workflow, which transforms raw video data into a format suitable for model training through sequential processes of video conversion, image extraction, and annotation. The side-angle dataset, initially in.DAT format, and the top-down dataset, initially in.AVI format, were standardized to.MP4 format using

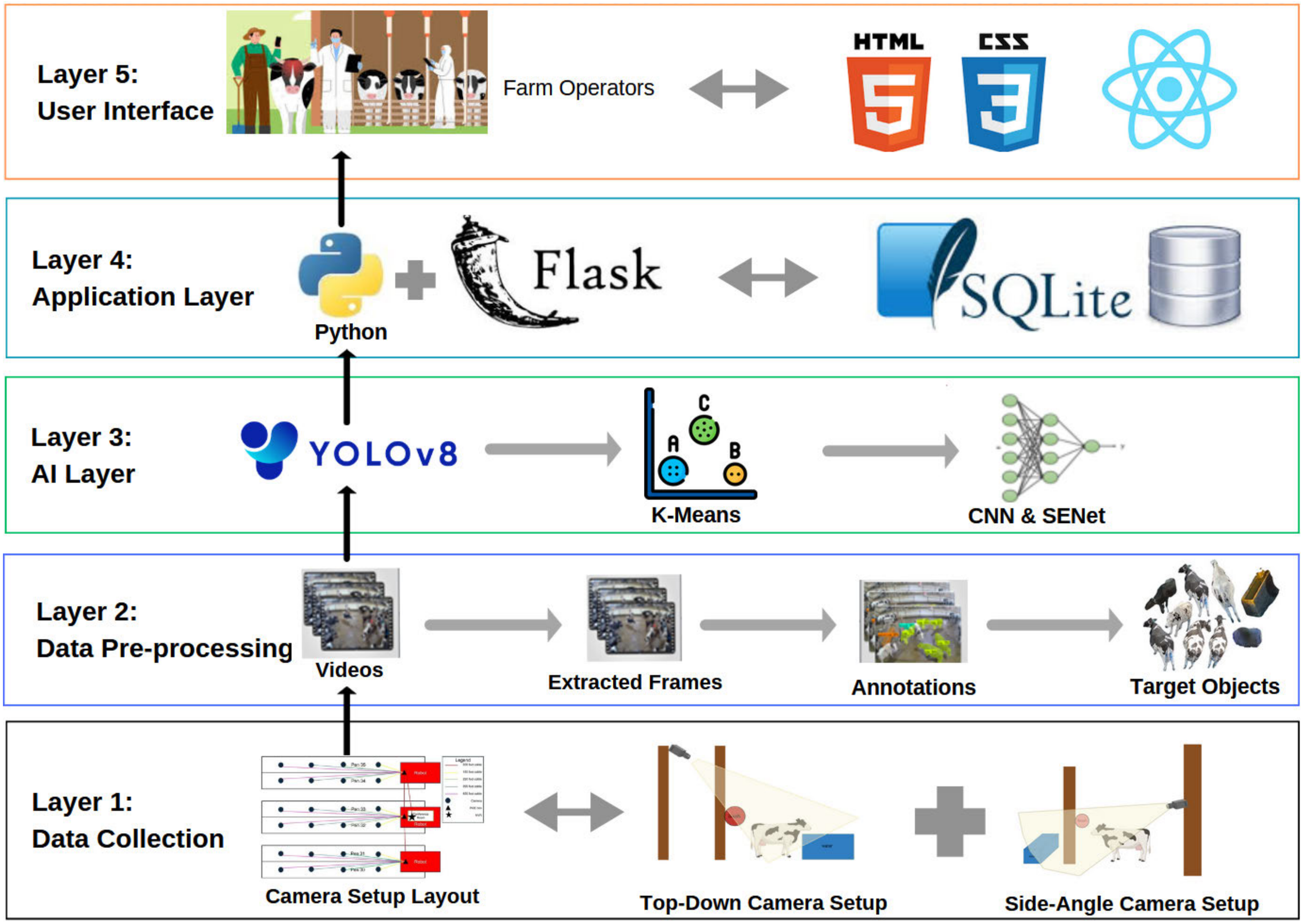


FIGURE 1. System architecture.

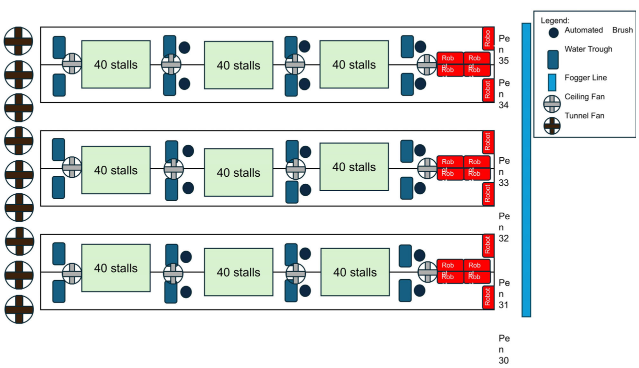


FIGURE 2. Schematic of the experimental barn layout, showing the placement of stalls, automated brushes, water troughs, and robotic milking stations in a representative pair of pens.

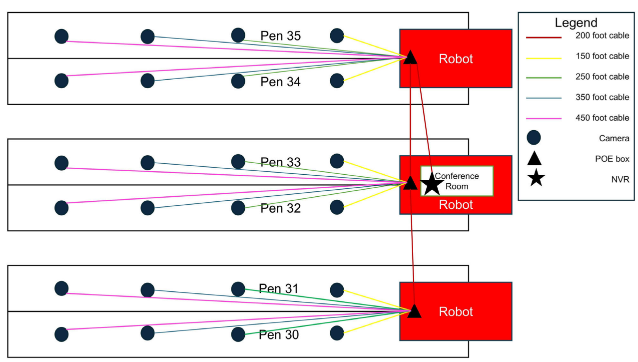
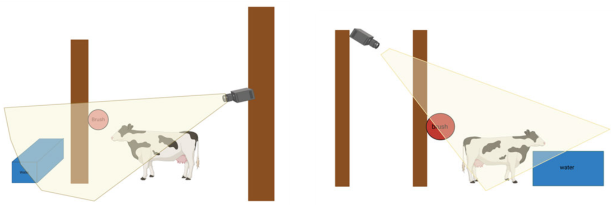


FIGURE 3. The cameras and data recording setup shows that each camera is connected via varying lengths of PoE cable to a central Network Video Recorder (NVR).

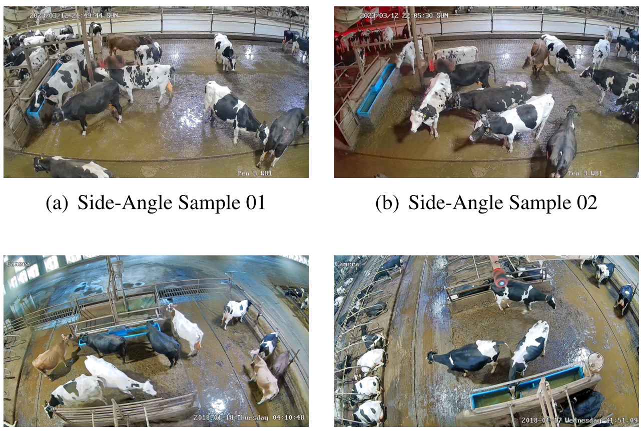
FFmpeg and VLC Media Player, respectively, to enhance compatibility and processing efficiency. From these converted videos, image frames were systematically extracted at a rate of one frame per second.

Dataset annotation for the object detection models was conducted using the Roboflow platform, which incorporates the Segment Anything Model (SAM) to accelerate instance segmentation tasks. While SAM substantially improved annotation efficiency, its performance was limited in cases where object boundaries were indistinct or blended into the background. Both datasets were subjected to a structured preprocessing and annotation pipeline to ensure consistency



(a) Side-Angle Camera Setup (b) Top-Down Camera Setup

FIGURE 4. **Camera placement in the barn showing side-angle and top-down viewpoints.**



(a) Side-Angle Sample 01 (b) Side-Angle Sample 02

(c) Top-Down Sample 01 (d) Top-Down Sample 02

FIGURE 5. **Representative dataset frames used for ML model annotation and training. Samples (a) and (b) show side-angle views; Samples (c) and (d) show top-down views. Photos are cropped for clarity and displayed at consistent aspect ratios.**

and quality for model training. The side-angle dataset produced 564 annotated images, partitioned into 500 for training, 40 for validation, and 24 for testing. The top-down dataset yielded 2,184 annotated images, comprising 1,528 for training, 437 for validation, and 219 for testing. This two-stage preparation process established a robust foundation for supervised learning across heterogeneous camera perspectives and variable barn conditions.

B. COMPUTATIONAL FRAMEWORK AND ALGORITHMS

This section outlines the central framework of our system, which employs AI models and algorithmic techniques to detect cow objects, distinguish individual cows, monitor their movements, and categorize behavioral patterns. The object detection and identification pipeline constitutes the foundational architecture of the behavioral characterization system. It is organized into four sequential stages: target object detection, unsupervised clustering ML model, supervised ML-based cow identification, and multi-object behavior tracking.

1) TARGET OBJECT DETECTION AND COW IDENTIFICATION MODULE

The target object detection process, using YOLOv8 model, begins with object detection using the YOLOv8 architecture, which is selected for its balance of speed and accuracy in real-time video analysis. YOLOv8 detects and localizes key objects—namely, full cow bodies, cow heads, water tubs, and mechanical brushes—within individual video frames. For each object, the model returns bounding boxes and associated class probabilities. These detections serve as the input for all downstream modules.

We employed an unsupervised clustering model, specifically the K-means algorithm, to implement the individual cow identification module. Each detected cow instance was encoded as a feature vector derived from its visual attributes, serving as the clustering input representation. The K-means algorithm subsequently grouped visually similar instances by minimizing within-cluster variance, as formalized by the following objective function:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{1}$$

where $S = \{S_1, S_2, \ldots, S_k\}$ is the set of k clusters, $\mathbf{x}$ is a feature vector, and $\boldsymbol{\mu}_i$ is the centroid of cluster S_i. The resulting cluster labels are treated as provisional cow identities and used to supervise a CNN in the next step.

We used a supervised ML-based identification model, namely Convolutional Neural Network (CNN) enhanced with Squeeze-and-Excitation CNN-SENet, to improve the accuracy and generalizability of the cow identification module, which was trained using the clustered dataset. The CNN is designed to capture both local texture features and global channel-wise dependencies, which are critical for distinguishing between visually similar cows. Table 1 summarizes the architecture used for this task.

TABLE 1. **CNN architecture for cow identification module.**

Layer	Output Size	Parameters
Conv1	$64 \times 64 \times 16$	3×3 Conv, ReLU, SE
Conv2	$32 \times 32 \times 32$	3×3 Conv, MaxPool, ReLU
Conv3	$16 \times 16 \times 64$	3×3 Conv, MaxPool, SE
FC1	128	Fully Connected, ReLU
FC2	k	Softmax Output Layer

The CNN network is optimized using the cross-entropy loss function, described as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{k} y_i \log(\hat{y}_i) \tag{2}$$

where y_i is the ground-truth class label and $\hat{y}_i$ is the predicted probability for class i.

2) MULTI-OBJECT BEHAVIOR TRACKING USING DEEPSORT MODEL

To maintain cow identity consistency across frames, we employed the DeepSORT model, which couples a Kalman filter for motion prediction with deep appearance embeddings for data association. Let $\mathbf{x}_t$ denote the object state (position

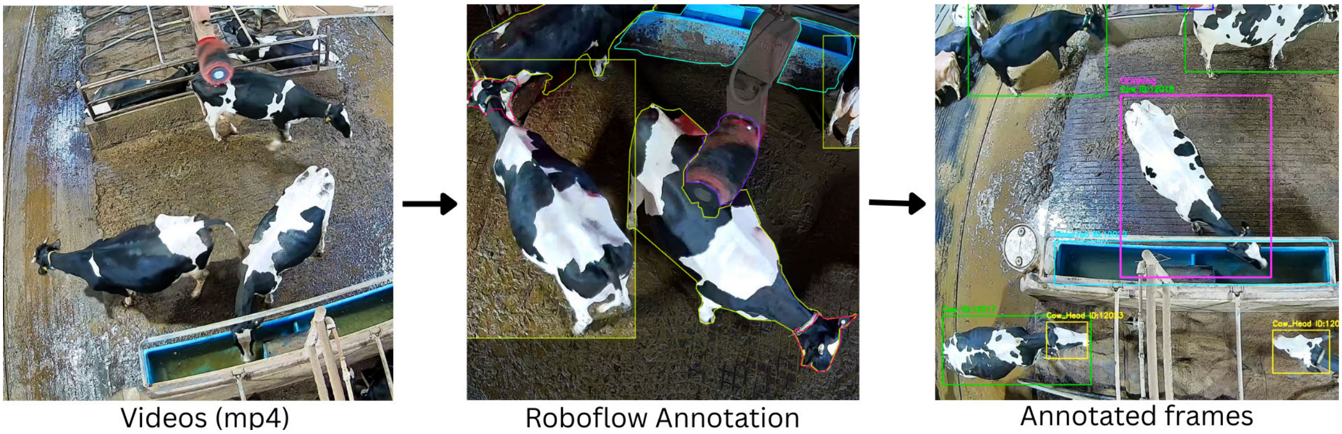


FIGURE 6. Dataset preprocessing phase: Slicing videos into individual frames, annotating images. This process illustrates the key frame extraction, conversion, and annotation steps.

and velocity) at time t and $\mathbf{z}_t$ the measurement derived from detector outputs. The motion and observation models can be expressed as:

$$\mathbf{x}_t = \mathbf{F}\,\mathbf{x}_{t-1} + \mathbf{w}_t \tag{3}$$

$$\mathbf{z}_t = \mathbf{H}\,\mathbf{x}_t + \mathbf{v}_t \tag{4}$$

where $\mathbf{w}_t$ and $\mathbf{v}_t$ are process and measurement noise, respectively.

For association, object detections are matched to predicted tracks by minimizing a cost that blends motion consistency and appearance similarity, which is represented formally as follows:

$$\psi_{ij} = \lambda\, d_{\mathrm{mahal}}(\hat{\mathbf{y}}_i, \mathbf{y}_j) + (1-\lambda)\left(1 - \cos(\mathbf{f}_i, \mathbf{f}_j)\right) \tag{5}$$

where d_{mahal} is the Mahalanobis distance between the predicted state $\hat{\mathbf{y}}_i$ and detection $\mathbf{y}_j$, and $\cos(\mathbf{f}_i, \mathbf{f}_j)$ is cosine similarity between deep appearance embeddings $\mathbf{f}_i$ and $\mathbf{f}_j$.

The behavior tracker predicts the cow states using Equations (3) and (4) and associates detections by solving a bipartite matching problem on Equation (5). This enables robust identity preservation under occlusion and crowded scenes, providing temporally coherent cow IDs for downstream behavior analysis.

3) BOUNDING BOX GEOMETRIC AND OVERLAP ALGORITHM

Object bounding boxes are represented as $B = (x_c, y_c, w, h)$ with center (x_c, y_c) and size (w, h). For overlap computation, we used edge coordinates (l, t, r, b) by the conversion formula shown in Algorithm 1.

Algorithm 1 Conversion of Bounding Box to Edge Coordinates

Require: $B = (x_c, y_c, w, h)$
Ensure: (l, t, r, b)

1: $l \leftarrow x_c - \frac{w}{2}, \quad t \leftarrow y_c - \frac{h}{2}$
2: $r \leftarrow l + w, \quad b \leftarrow t + h$

Algorithm 2 shows the process of checking the coordinate overlap between two bounding boxes. Once an overlap area is found, we use Algorithm 3 to obtain the intersection area. The Intersection-Over-Union (IoU) between B_1 and B_2 can be calculated using this formula:

$$\mathrm{IoU}(B_1, B_2) = \frac{A(B_1 \cap B_2)}{A(B_1) + A(B_2) - A(B_1 \cap B_2)} \tag{6}$$

Algorithm 2 Check Coordinate Overlap Between Two Bounding Boxes

Require: $B_1 = (l_1, t_1, r_1, b_1);\ B_2 = (l_2, t_2, r_2, b_2)$
Ensure: Overlap flag and overlap area A

1: **if** $l_1 \geq r_2$ or $r_1 \leq l_2$ or $t_1 \geq b_2$ or $b_1 \leq t_2$ **then**
2: Overlap $\leftarrow$ False; $A \leftarrow 0$
3: **else**
4: Overlap $\leftarrow$ True; $A \leftarrow \mathrm{OverlapArea}(B_1, B_2)$

Algorithm 3 Computation of Overlap Area Between Two Bounding Boxes

Require: $B_1 = (l_1, t_1, r_1, b_1);\ B_2 = (l_2, t_2, r_2, b_2)$
Ensure: Overlap area A

1: $x_o \leftarrow \max\big(0,\ \min(r_1, r_2) - \max(l_1, l_2)\big)$
2: $y_o \leftarrow \max\big(0,\ \min(b_1, b_2) - \max(t_1, t_2)\big)$
3: $A \leftarrow x_o \times y_o$

4) INTEGRATED ML INFERENCE PIPELINE

Algorithm 4 integrates the detection, tracking, and re-identification stages to generate temporally consistent cow identities and associated behavioral events. Identity predictions are cached across time to enhance label stability, while per-cow buffers preserve short-term histories required for reasoning based on motion dynamics and spatial overlap.

5) PER-FRAME BEHAVIORAL ANALYSIS

Algorithm 5 describes the cow behavioral detection module on a per-frame basis that incrementally updates event timelines. Temporal persistence is enforced by stipulating that each predicate must be satisfied for at least κ consecutive

Algorithm 4 Integrated Cow Behavior and Identity Inference

Input: Video sequence $F = \{f_1, \dots, f_T\}$
Output: Annotated video and event log
Initialize YOLO detector, DeepSORT tracker, CNN–SENet identifier
for each frame $f_t \in F$ **do**
  $D_t \leftarrow$ YOLO.detect(f_t)
  $T_t \leftarrow$ DeepSORT.update(D_t)
  for each cow track $c \in T_t$ **do**
    Extract image patch of c and predict identity with CNN–SENet
    Update per-cow state buffers (head, body, velocities)
    Evaluate brushing and drinking logic; update event timelines

frames before event initiation and must remain unsatisfied for at least κ frames prior to termination, thereby mitigating spurious fluctuations arising from detector jitter.

Algorithm 5 Per-Frame Behavioral Analysis

Input: Frame f_t; detections D_t; tracks T_t
Output: Updated brushing/drinking logs
for each cow $c \in T_t$ **do**
  if BrushingDetected$c$, D_t **then**
    Update brushing start/end for c
  if DrinkingDetected$c$, D_t **then**
    Update drinking start/end for c

6) BRUSHING BEHAVIOR DETECTION

Brushing behavior is defined as sustained contact between a cow and the rotating brush. A cow i is considered in proximity of a brush if $\mathrm{IoU}(B_i^{\mathrm{body}}, B_b^{\mathrm{brush}}) > \theta_p$. Brush motion is quantified as the mean frame-to-frame centroid displacement over a window of N frames, which is defined as:

$$\bar{D}_b(t) = \frac{1}{N-1} \sum_{u=t-N+2}^{t} \|\mathbf{C}_b(u) - \mathbf{C}_b(u-1)\|_2 \quad (7)$$

where $\mathbf{C}_b(u)$ is the brush centroid at frame u. A brushing event is declared when proximity holds and $\bar{D}_b(t) > \tau_{\mathrm{mot}}$ for at least κ consecutive frames.

Distinct brushing events are separated by a minimum inter-event gap $\Delta_{\min}$ (5s) to avoid fragmentation. Equation (7) suppresses false positives from cows resting near an idle brush by requiring measurable brush motion, while the proximity condition ensures spatial plausibility of interaction.

7) DRINKING BEHAVIOR DETECTION

Drinking behavior prioritizes head-trough alignment when a cow's head detection is available and falls back to a relaxed body-trough overlap otherwise. Let B_i^{head} and B_i^{body} denote the head and body boxes for cow i, and $\mathcal{T}$ be the set of trough boxes, we define the bounding box overlap (IoU) area as follows:

$$IoU(B, \mathcal{T}) = \max_{T \in \mathcal{T}} \mathrm{IoU}(B, T) \quad (8)$$

As described in Algorithm 6, we impose a more stringent threshold, denoted θ_h, to detect head-trough contact. In contrast, a comparatively lower fallback threshold, θ_b, is employed to identify body-trough overlap in order to mitigate the effects of intermittent head occlusions. Consistent with the brushing detection procedure, temporal persistence over κ consecutive frames and a minimum inter-event interval $\Delta_{\min}$ are enforced to ensure the robustness and stability of event duration estimates.

Algorithm 6 Drinking Behavior Detection

Require: Cow body B_i^{body}; head set $\mathcal{H}$; troughs $\mathcal{T}$
Ensure: Boolean indicator *is_drinking*
$H^* \leftarrow \textsc{MatchHeadToBody}(B_i^{\mathrm{body}}, \mathcal{H})$
if $H^* \neq \varnothing$ **then**
  $is\_drinking \leftarrow \mathrm{IoU}(H^*, \mathcal{T}) > \theta_h$
else
  $is\_drinking \leftarrow \mathrm{IoU}(B_i^{\mathrm{body}}, \mathcal{T}) > \theta_b$

8) HEADBUTT DETECTION

Agonistic interactions, with particular emphasis on headbutts, are detected through a multi-stage rule-based framework that integrates kinematic and spatial constraints. The rationale for this dual approach lies in the behavioral characteristics of headbutting: the action is defined not only by rapid, forceful head movements (kinematics) but also by the close physical proximity between the interacting cows (i.e., spatial positioning). Relying exclusively on kinematic signatures risks misclassifying non-agonistic rapid head movements, while spatial cues alone cannot disambiguate passive body contact from active aggression. To address these limitations, the detection procedure requires simultaneous satisfaction of both domains of evidence. Specifically, head acceleration and velocity profiles are analyzed to identify sudden, directed movements consistent with a striking motion, while spatial overlap and relative orientation constraints ensure that the movement occurs in close proximity to another cow's anterior region. Only when these conditions are jointly met is a headbutt event logged, thereby reducing false positives and improving the robustness of agonistic interaction detection.

A headbutt event is detected only when all the following criteria are satisfied.

a: VELOCITY CALCULATION

Let $\vec{h}_i(t)$ be the center position of the head of cow i at time t, and $\vec{v}_i(t)$ be the velocity vector of cow i's head at time t, then the velocity can be defined as follows:

$$\vec{v}_i(t) = \begin{cases} \vec{h}_i(t) - \vec{h}_i(t-1), & \text{if } \vec{h}_i(t-1) \text{ exists} \\ \vec{0}, & \text{otherwise} \end{cases} \quad (9)$$

b: PROXIMITY CHECK

Let $\vec{c}_j(t)$ be the center position of the body of cow j at time t, the proximity is defined using a threshold distance ϵ. That check confirms that the head of cow i is sufficiently close to the body of cow j:

$$||\vec{h}_i(t) - \vec{c}_j(t)|| < \epsilon \tag{10}$$

c: DIRECTIONAL IMPACT (HEADBUTT LIKELIHOOD)

Let $\vec{d}_{ij}(t) = \vec{c}_j(t) - \vec{h}_i(t)$ be the direction vector from cow i's head to cow j's body, and $||\vec{v}_i(t)||$ be the speed of the head movement, a headbutt event is detected if the movement speed exceeds a minimum threshold and is directed toward the target:

$$||\vec{v}_i(t)|| \geq v_{\min} \quad \text{and} \quad \vec{v}_i(t) \cdot \vec{d}_{ij}(t) \geq \tau \tag{11}$$

where $v_{\min}$ is the minimum speed threshold and τ is the directional threshold.

d: HEADBUTT LOGGING

If all of the conditions above are satisfied, a headbutt event, HeadbuttEvent(i, j, t) is logged, which signifies that cow i is inferred to have headbutted cow j at time t.

Algorithm 7 provides a concise computational summary of these rules. The algorithm iterates through each pair of cows (i, j). It evaluates a single conditional statement (line 4) that directly combines the mathematical definitions for proximity check ($||\vec{d}_{ij}|| < \epsilon$) and directional impact ($||\vec{v}_i(t)|| \geq v_{\min}$ and the dot product check). If the conditions are met, it performs the Headbutt Logging step (line 5). The Velocity Calculation is treated as a pre-computed input required by the algorithm.

Algorithm 7 Headbutt Behavior Detection

Require: Head positions $\{\vec{h}_i(t)\}$, body positions $\{\vec{c}_j(t)\}$, velocities $\{\vec{v}_i(t)\}$. Thresholds $v_{\min}, \tau, \epsilon$.
Ensure: Set of headbutt events E_{headbutt}.

1: $E_{\text{headbutt}} \leftarrow \varnothing$
2: **for** each pair of distinct cows (i, j) **do**
3: $\quad \vec{d}_{ij} \leftarrow \vec{c}_j(t) - \vec{h}_i(t)$
4: $\quad$ **if** $||\vec{d}_{ij}|| < \epsilon$ **and** $||\vec{v}_i(t)|| \geq v_{\min}$ **and** $\vec{v}_i(t) \cdot \vec{d}_{ij} \geq \tau$ **then**
5: $\quad\quad E_{\text{headbutt}} \leftarrow E_{\text{headbutt}} \cup \{(i, j, t)\}$
6: **return** E_{headbutt}

IV. IMPLEMENTATION

This section details the implementation of the automated behavior characterization system, which necessitated the coordinated integration of high-performance computing hardware, optimized deep learning frameworks, and an intuitive user interface. Dual objectives guided the design: enabling large-scale video analysis with computational efficiency, and providing accessible visualization tools to facilitate interpretation of results. Accordingly, the following subsections describe the hardware and software environments, as well as the architecture of the web-based application that supports user interaction with system outputs.

A. PHYSICAL COMPUTING CAPABILITIES

All computationally demanding operations –including model training and processing more than 38,000 surveillance videos– were executed on a dedicated cluster comprising four Alienware Aurora R16 desktop systems [40]. The two primary workstations are equipped with Intel Core i9-14900KF processors, 64 GB DDR5 RAM, 2 TB NVMe solid-state drives, and NVIDIA GeForce RTX 5090 graphics processing units (GPUs) with 32 GB of dedicated VRAM. The two secondary workstations were configured with identical CPU and memory specifications but utilized NVIDIA GeForce RTX 4090 GPUs with 24 GB of VRAM. This heterogeneous multi-GPU architecture facilitated efficient task distribution and parallelization, thereby accelerating both model inference and large-scale data processing while markedly reducing overall analysis time.

B. SOFTWARE PIPELINE

The software infrastructure was developed primarily in Python, with deep learning components implemented using the PyTorch framework [41]. Experimental prototyping and visualization were conducted within Jupyter Notebooks [42], while image annotation for the custom object detection dataset was performed using the Roboflow platform [36]. The core processing pipelines integrated video decoding, frame extraction, and YOLOv8-based inference, yielding structured event-level data amenable to subsequent higher-order behavioral analyses.

The system backend was implemented using the Python Flask framework [38], which managed video uploads, orchestrated AI inference APIs, and recorded metadata within a lightweight SQLite database. To ensure reproducibility, consistency across environments, and portability, the application was fully containerized using Docker. While this configuration was well-suited for the present study, the architecture was designed with scalability in mind. Specifically, the lightweight database layer can be migrated to distributed systems such as PostgreSQL [43] or cloud-native solutions (e.g., Amazon RDS [44], Google Cloud SQL [45]) to accommodate higher data throughput and concurrent access. Likewise, containerized services can be deployed in orchestrated environments (e.g., Kubernetes [46] or Docker Swarm [47]), enabling elastic resource allocation, fault tolerance, and support for large-scale, multi-user deployments. This forward-compatible design ensures that the system can evolve from a laboratory-scale prototype to a production-ready platform suitable for broader research and field applications.

C. GRAPHICAL USER INTERFACE AND VISUALIZATION

To enhance user usability, the system incorporates a web-based graphical interface developed with React, HTML, and CSS. The GUI provides interactive dashboards for monitoring processed videos, visualizing detected behavioral events, and examining aggregated statistics across large datasets. Figure 7 presents a screenshot of the primary dashboard, while Figure 8 depicts the video inspection module. Beyond functionality, the interface design was guided by established human-computer interaction principles. Particular attention was given to intuitiveness, ensuring that complex analytical outputs can be navigated without extensive technical training; accessibility, with layouts and color schemes optimized for clarity and inclusivity; and responsiveness, enabling smooth performance across different devices and screen sizes. These design considerations allow users to seamlessly transition between raw inference outputs and higher-level analytical summaries, thereby strengthening the interpretability and practical utility of the system.

V. EVALUATION

This section evaluates the end-to-end pipeline system with respect to two complementary dimensions: (i) concordance with human annotations at both the video and event levels, and (ii) the characterization of large-scale behavioral patterns derived from the video datasets.

Accurate detection and classification of key behaviors represent a critical requirement for applying the system to the monitoring of thermotolerance. Figure 9 illustrates qualitative outputs from the inference pipeline, demonstrating successful identification of both brushing and drinking events under typical barn conditions. These examples underscore the model's robustness in accommodating variability in cow posture, minor occlusions, and interactions with environmental objects.

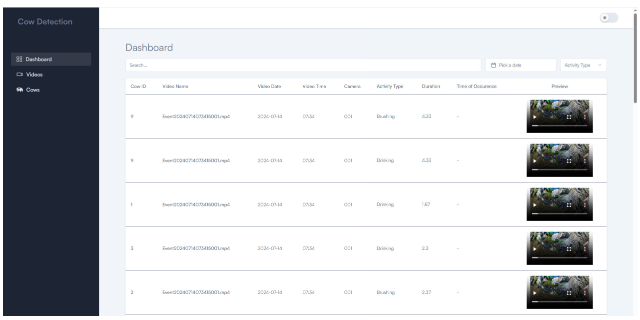


FIGURE 7. **Interactive GUI dashboard for visualizing behavioral analytics and aggregated event outputs.**

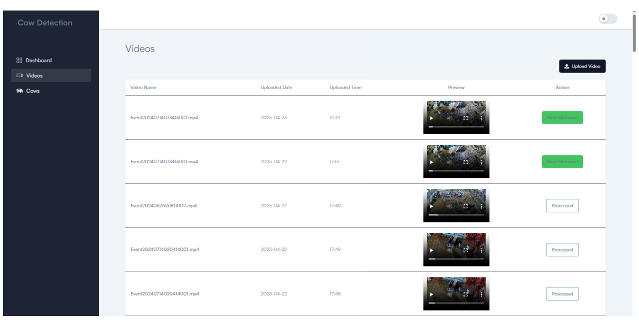
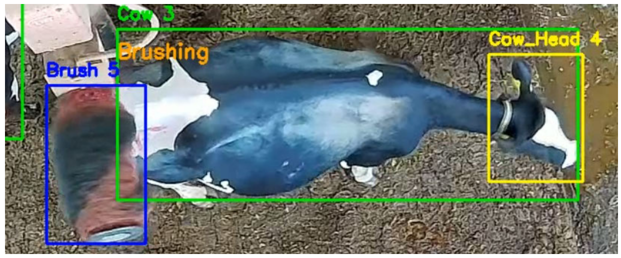
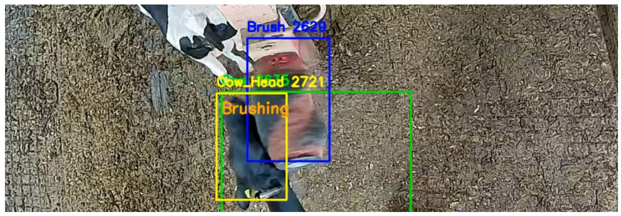


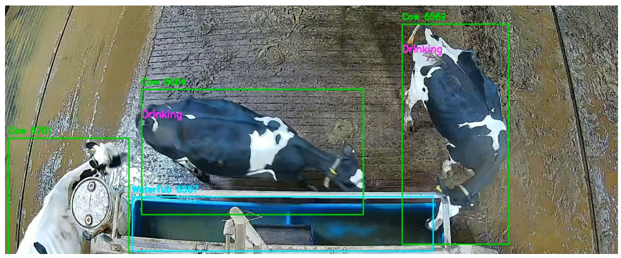
FIGURE 8. **Video dashboard page that allows farm operators to launch inference kernels to analyze video recordings.**



(a) Brushing event from a top-down perspective.



(b) Brushing event showing head-brush interaction.



(c) Drinking events with multiple cows at the waterer.

FIGURE 9. **Qualitative examples of the system's output. The ML models successfully detect and label brushing (Figures 9(a), 9(b)) and drinking (Figure 9(c)) behaviors by identifying the target objects (i.e., cows, brush, and waterer), along with their spatial interactions. These examples illustrate the system's performance in typical, often crowded, barn conditions.**

A. HUMAN-OBSERVATION VALIDATION

To rigorously evaluate the system's behavioral detection performance, human observers (four undergraduate student research workers) from the Texas A&M University (TAMU), Animal Science team manually annotated a validation dataset. Students were employed after completing structured training sessions led by animal behavior specialists. The training included guided labeling exercises and iterative feedback to ensure consistency with expert standards. This approach enabled efficient large-scale annotation while maintaining an acceptable level of accuracy. The dataset comprised 457 videos, each approximately two minutes in duration, which were annotated using the Behavioral Observation Research Interactive Software (BORIS, v8.0.6) [48].

We used the ethogram catalog that operationalizes brushing and drinking events using a threshold of approximately five seconds to delineate the termination of one event and the onset of a subsequent event. Every human observer reviewed each video clip. Only videos overlapping with the large-scale deployment dataset were retained for validation. The final curated validation subset comprised videos after excluding those outside the deployment set.

1) BEHAVIOR EVENT FRAGMENTATION

Behavioral events are represented as time intervals $I = (s, e)$ corresponding to the start and end times of a given cow's behavior. Model-generated timelines were compared against human observations, revealing that events detected by the system were often fragmented into multiple short intervals occurring in rapid succession. This fragmentation arises from sensitivity to transient positional changes, such as brief head movements away from the water trough. To address this issue, consecutive intervals associated with the same cow and behavior are merged whenever the inter-interval gap does not exceed Δ_{merge}. Formally, if $I_1 = (s_1, e_1)$ and $I_2 = (s_2, e_2)$ with $e_1 \leq s_2$ and $s_2 - e_1 \leq \Delta_{\text{merge}}$, the consolidated interval, $I^\star$, is represented as follows:

$$I^\star = (s_1, e_2). \tag{12}$$

Consolidation was applied post-detection and before metric computation to ensure that evaluation metrics capture continuous behavioral patterns rather than artifacts introduced by detector jitter. Model performance was then assessed on the human-validated video set using the YOLOv8n detector, with two validation metrics: (i) event duration accuracy, measuring agreement between the total time assigned to behavioral events by the system and human annotations, and (ii) event frequency accuracy, quantifying correspondence in the number of events detected.

2) EVENT DURATION AGREEMENT

At the video level, we aggregate the total duration per behavior and compute the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson correlation, r, as follows:

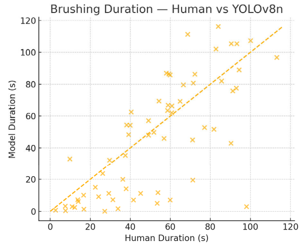
$$\text{MAE} = \frac{1}{N}\sum_{v=1}^{N}\left|D_v^{\text{model}} - D_v^{\text{human}}\right| \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{v=1}^{N}\left(D_v^{\text{model}} - D_v^{\text{human}}\right)^2} \tag{14}$$

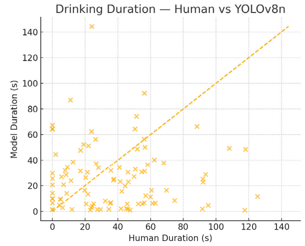
As illustrated in the scatter plots in Figure 10, brushing duration estimates show strong concordance with human annotations (MAE = 11.07s, RMSE = 19.88s, Pearson $r = 0.867$). In contrast, performance for drinking events is notably weaker (MAE = 36.75s, RMSE = 47.53s, $r = 0.012$), primarily attributable to occlusions and animal crowding around the water tub.

3) EVENT COUNTS AND EVENT-LEVEL ALIGNMENT

Event counts diverged more than durations due to fragmentation –the model often splits one continuous bout into multiple



(a) Brushing Behavior Durations



(b) Drinking Behavior Durations

FIGURE 10. Scatter plots show the correlation between system output and duration of human-observed events for brushing and drinking behaviors.

short detections. We quantify the over-counting per video with the fragmentation index, ϕ, as follows:

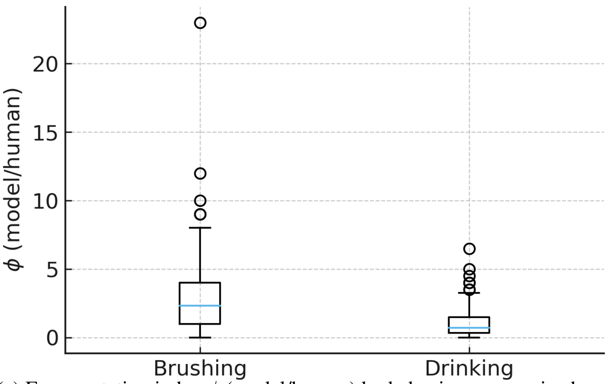
$$\phi = \frac{\#\text{events}^{\text{model}}}{\#\text{events}^{\text{human}}}. \tag{15}$$

Within the matched video cohort, brushing events exhibit systematic over-counting relative to human annotations (median $\phi = \mathbf{2.33}$, $N = 82$), whereas drinking events align more closely with human logs (median $\phi = \mathbf{0.71}$, $N = 128$). The distributions of these ratios are summarized in Figure 11(a). We applied a 5-second consolidation window to merge adjacent detections to mitigate the over-counting observed for brushing. The choice of a 5s threshold reflects the ethogram's definition of inter-event separation, ensuring that brief positional changes –such as transient head movements away from the brush– are not misclassified as distinct behavioral bouts. The effect of this adjustment is illustrated in Figure 11(b), which presents a representative clip with three comparative bars: Human annotations (top), Model (pre-consolidation) outputs (middle), and Model (5,s consolidated) outputs (bottom).

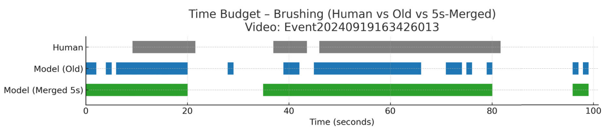
For event-level agreement, we align intervals with report precision (fraction of model events that overlap a human event), recall (fraction of human events covered by at least one model event), and temporal Intersection-over-Union (IoU), which is defined as follows:

$$\text{IoU}_t = \frac{|[s_1, e_1] \cap [s_2, e_2]|}{|[s_1, e_1] \cup [s_2, e_2]|} \geq \tau_t, \quad \tau_t = 0.30. \tag{16}$$

Brushing events demonstrated high recall but modest precision (per-video medians $\approx$ 0.67 recall and $\approx$ 0.25 precision; micro-averaged 0.65/0.22), indicating that while the model reliably captured the majority of human-annotated bouts, it frequently fragmented them into multiple detections. This imbalance reflects the detector's heightened sensitivity to short positional deviations –such as transient head movements away from the brush– that are not treated as distinct bouts under the ethogram definition. Drinking events exhibited substantially weaker alignment (medians$\approx$ 0.12 recall and $\approx$ 0.17 precision; micro-averaged 0.18/0.24), a result attributable primarily to occlusions and the geometric complexity of head–trough interactions, which reduce the system's ability to maintain consistent bout boundaries. The precision-recall trade-offs observed here highlight the methodological tension between model sensitivity and ethogram-based behavioral definitions: while higher sensitivity ensures that true bouts are rarely missed, it also increases susceptibility to over-fragmentation. These error modes and the improvement gained through temporal consolidation are visually illustrated in the three-bar time plot of Figure 11(b).



(a) Fragmentation index ϕ (model/human) by behavior, summarized per video.



(b) Time budget for brushing in `Event20240919163426013`, showing *Human* (top), *Model (old)* before consolidation (middle), and *Model (5 s merged)* (bottom).

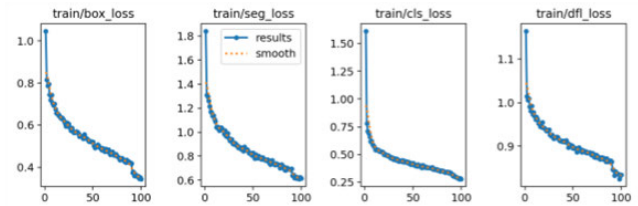
FIGURE 11. Event count agreement and temporal alignment. (a) Brushing behavior exhibits higher over-counting (higher ϕ) than drinking behavior. (b) The three-row time plot highlights how consolidation reduces fragmentation and better aligns model detections with human bouts.

B. MODEL TRAINING AND DIAGNOSTICS

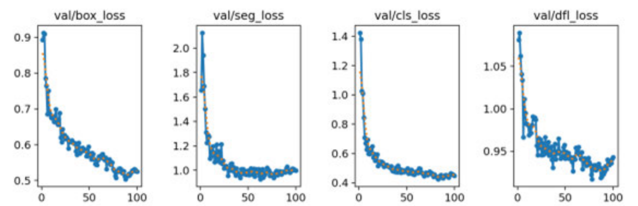
The developed detection pipeline utilizes the YOLOv8n model, which is trained on both side-angle and top-down images to enhance robustness across diverse viewing geometries. This dual-perspective training strategy ensures the model captures discriminative features regardless of camera placement, thereby improving generalization to varied barn layouts. Standard geometric and photometric augmentations were further applied to expand the effective training distribution, mitigating overfitting and increasing resilience to lighting variability and minor pose differences. As shown in Figure 12, training and validation losses decrease monotonically, indicating stable convergence under the selected optimization schedule. Confusion matrices (Figures 3 and 4) demonstrate high separability across cow, cow_head, brush, and tub categories. Residual misclassifications are primarily attributable to structural ambiguities, such as head-body adjacency and partial occlusions near troughs, which present intrinsic challenges to visual discrimination. Table 2 shows the ML detection performance metrics for the validation data subset.

TABLE 2. ML detection performance metrics for the validation data subset.

Class	Precision (P)	Recall (R)	mAP@.5	mAP@.5:.95
brush	0.993	0.975	0.995	0.903
cow	0.924	0.802	0.956	0.784
cow head	0.918	0.805	0.942	0.785
watertub	0.956	0.899	0.980	0.849
All Classes	**0.977**	**0.871**	**0.955**	**0.806**



(a) Training loss



(b) Validation loss

FIGURE 12. Training and validation loss trends for the YOLOv8 segmentation model over 100 epochs. The first row shows the training losses, including bounding box regression loss (train/box_loss), segmentation loss (train/seg_loss), classification loss (train/cls_loss), and distribution focal loss (train/dfl_loss)—while the second row shows the corresponding validation losses (val/box_loss, val/seg_loss, val/cls_loss, and val/dfl_loss).

TABLE 3. Training confusion matrix (row-normalized). Values are fractions; entries <0.01 shown as 0.00.

Pred.\True	brush	cow	cow_head	watertub	background
brush	1.00	0.00	0.00	0.00	0.00
cow	0.00	0.90	0.00	0.00	0.05
cow_head	0.00	0.00	0.88	0.00	0.07
watertub	0.00	0.00	0.00	0.98	0.00
background	0.00	0.10	0.02	0.00	0.85

TABLE 4. Validation confusion matrix (row-normalized). Values are fractions; entries <0.01 shown as 0.00.

Pred.\True	brush	cow	cow_head	watertub	background
brush	1.00	0.00	0.00	0.00	0.00
cow	0.00	0.90	0.00	0.00	0.05
cow_head	0.00	0.00	0.85	0.00	0.10
watertub	0.00	0.00	0.00	0.98	0.01
background	0.00	0.10	0.02	0.00	0.80

(a) Input Frame (b) Headbutting Motion (c) After Headbutting

(d) Input Frame – Annotated (e) Headbutting Motion – Annotated (f) After Headbutting – Annotated

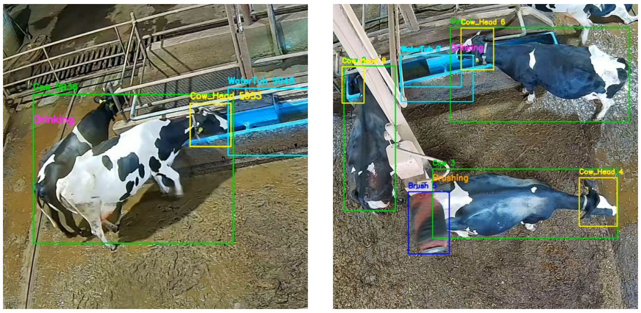
FIGURE 13. **Headbutt Detection Sequence (top: input frames; bottom: model overlays). The aggressor's head velocity is directed toward the recipient while proximity holds, satisfying our detection criteria (velocity, direction, proximity, and temporal persistence).**

Figure 13 presents a representative instance of a headbutt event, shown in raw and annotated formats. Per the detection criteria outlined in Section III, the aggressor's head velocity exceeds the defined threshold, with the movement vector directed toward the recipient while spatial proximity is maintained. From an ethological perspective, this interaction aligns with established definitions of agonistic behavior, where forceful head contact serves as a mechanism of dominance assertion. In this case, the headbutt results in the recipient being displaced from the brush, reflecting the role of such interactions in mediating competition over limited resources.

C. ERROR ANALYSIS AND FAILURE MODES

System performance analysis revealed several recurrent failure modes, predominantly arising from challenges in object detection and tracking under conditions of heavy occlusion. These errors frequently occurred when multiple cows overlapped or when key anatomical features, such as the head or muzzle, were partially obscured by environmental structures. From a methodological perspective, such limitations underscore the sensitivity of detection pipelines to dataset composition and annotation practices. While the training set incorporated diverse postures and lighting conditions, occlusion-heavy scenarios were comparatively underrepresented, reducing model robustness in crowded contexts. Moreover, reliance on bounding-box-based detectors such as YOLOv8n introduces inherent difficulties in segmenting fine-grained features when individuals are in close proximity, suggesting that alternative architectures (e.g., instance segmentation or keypoint-based models) may offer improved resilience. These observations not only guide potential avenues for model refinement but also contribute to the interpretability of failure cases, situating them within the broader methodological trade-offs of dataset design and architectural choice.

The most pronounced detection failures arose when cows were closely aligned or when one individual was positioned directly behind another. Under these conditions of severe occlusion, the model occasionally merged two distinct animals into a single bounding box, as illustrated in Figure 14. Such errors are particularly consequential, as they result in undercounting of animals and the misattribution of behaviors (e.g., drinking) to a non-existent, merged entity. These observations underscore the inherent limitations of bounding-box representations in crowded or overlapping scenes. Potential strategies to address these challenges include adopting instance segmentation architectures capable of delineating overlapping individuals at the pixel level, keypoint-based tracking models that exploit anatomical landmarks for more precise identification of partially occluded animals, and multiview fusion approaches that integrate complementary perspectives to resolve ambiguities arising from single-camera occlusions. Implementing such solutions could enhance the robustness of behavioral detection in high-density settings and improve the fidelity of subsequent ethological analyses.



(a) Occlusion Case (b) Failed Object Localization

FIGURE 14. Detection localization inaccuracy. Figure 14(a) shows a case where a partial occlusion occurred, causing two cows to be incorrectly identified as one entity. Figure 14(b) shows a case where the bounding box of the waterer is partially invisible due to occlusion. Hence, the drinking event for cow 2 (left of the pillar) is not recognized and logged.

Localization inaccuracy and re-identification errors represent key limitations in the current detection pipeline. For static objects such as water troughs, improving localization could be achieved through instance segmentation models that delineate objects at the pixel level rather than relying on bounding boxes, thereby increasing the likelihood of sufficient overlap with interacting animals. Additionally, multiview camera setups could reduce occlusion effects by providing complementary perspectives, ensuring more complete object coverage.

Re-identification errors following prolonged occlusions could be mitigated using appearance-based embedding models, which encode individual-specific visual features to maintain identity continuity, even when animals are temporarily obscured. Temporal consistency constraints and motion trajectory modeling could further reinforce identity assignments by leveraging predictable movement patterns, reducing the probability of assigning new IDs to previously tracked individuals. Together, these methodological refinements would enhance both object localization accuracy and the fidelity of longitudinal behavioral records, supporting more robust and interpretable analyses of individual animal behavior.

Expanding the training dataset to include scenarios with dense crowds, severe occlusions, and environmental clutter is expected to reduce the frequency of detection and tracking errors; however, increasing data scale alone cannot overcome the intrinsic limitations of bounding-box representations. Bounding boxes inherently conflate body parts and contact geometry, which are critical for accurately capturing social, contact-intensive behaviors such as headbutts. To better resolve these interactions, moving beyond box overlap and adopting pixel- or keypoint-level reasoning through instance segmentation or pose estimation is necessary. Off-the-shelf tools, including the SAM2 segmentor and markerless pose estimation pipelines such as DeepLabCut, can provide richer spatial information—capturing orientation, physical contact, and displacement—while remaining robust even under conditions of partial occlusion.

TABLE 5. Training configuration and data summary (YOLOv8n-seg). Ultralytics defaults are used where not specified.

Setting	Value
Model	YOLOv8n-seg (Ultralytics v8.0.106)
Framework	PyTorch 2.1.0 + cu118
Input size	640 × 640
Epochs	100
Batch size	default (Ultralytics)
Optimizer / LR	default (Ultralytics)
Augmentation	flips, HSV, mosaic, random crop
Classes	cow, cow_head, brush, tub (`watertub`)
Train/Val/Test	1183 / 651 / 350 images (54/30/16%)
Device	Tesla T4 (15 GB)
Params / GFLOPs	11.79M / 42.4

Re-identification failures following prolonged occlusions highlight the need for more robust appearance embeddings and temporally informed association strategies. To address this, we plan to evaluate self-supervised visual feature representations such as DINOv3 [49], alongside end-to-end tracking architectures like TrackFormer [50], which integrate both motion and appearance cues to preserve identity continuity through extended periods of occlusion.

We archived the cohort manifest, training configuration (Table 5), and figure-generation scripts, together with detailed run metadata, including commit hashes, random seeds, and applied thresholds. This comprehensive archive supports full reproducibility of the reported analyses and figures, aligning with best practices in open science by enabling independent verification, methodological transparency, and the reuse of computational workflows in future studies.

D. SYSTEM LIMITATIONS

The system was primarily trained on daylight video captured at a single site, and its generalization to red-light or nighttime conditions remains limited. A single rater annotated most video clips; although quality-control checks and a double-coded subset demonstrated strong agreement, comprehensive inter-rater reliability studies would further strengthen external validity. Finally, behavioral predicates are currently defined based on bounding-box overlaps and are therefore sensitive to occlusions. Methodological enhancements, such as instance segmentation and multi-view data fusion, are expected to improve robustness under these challenging conditions.

VI. CONCLUSION AND FUTURE WORKS

This paper presented an ML-driven framework for the automated detection of cows, brushes, and water tanks in video streams, alongside recognizing context-specific behaviors such as drinking and brushing. The system leverages state-of-the-art neural network architectures –most notably YOLOv8 for real-time object detection and a CNN enhanced for individual cow identification and clustering. Empirical evaluations demonstrate strong performance in both object localization and classification of cow interactions with brushes and water troughs, even within the complexities of a dynamic farm environment. However, the system's ability to capture more complex social interactions, such as headbutting, requires further refinement. The reliance on bounding-box representations limits precision, and validated percent agreement for these behaviors has not yet been established, making this an important area for continued validation. Overall, these findings highlight the promise of integrating computer vision and machine learning into livestock monitoring systems, showing potential to enhance resource utilization, support early detection of behavioral anomalies, and ultimately improve both operational efficiency and animal welfare.

Several directions remain open for expanding and enhancing the system's analytical depth and technical capabilities. A primary goal is to refine the characterization of agonistic behaviors. For instance, the system can be extended to detect headbutts and localize them within a specific spatial context, such as competition at a water trough or brush. Additionally, by analyzing the subsequent motion vectors, the model could be enhanced to differentiate between headbutts with and without displacement, enabling a quantitative classification of interaction severity. Achieving this level of nuanced understanding requires moving beyond current methods toward more advanced vision models.

A pivotal direction involves transitioning from bounding-box-based detection to high-fidelity instance segmentation. Leveraging foundation models like Meta's Segment Anything (SAM) [51] and its next-generation versions would allow the system to generate precise pixel-level masks for each cow. This transition from coarse bounding boxes to fine-grained masks would dramatically improve the accuracy of spatial reasoning for all contact-based behaviors, significantly reducing ambiguity and false positives—a known failure mode of the current approach.

We plan to explore two synergistic technologies to address the challenge of identity preservation and tracking drift. First, features from the latest generation of self-supervised models, such as Meta's recently released DINOv3 [49], can provide even more powerful and generalizable visual representations for re-identification. Second, we can move beyond the classic tracking-by-detection paradigm by implementing end-to-end Transformer-based trackers like TrackFormer [50]. Such architectures unify object detection and data association into a single network, allowing them to model long-range temporal dependencies for more robust and context-aware tracking, directly mitigating identity-switch errors.

Another promising avenue involves creating a more holistic welfare model through sensor fusion. Integrating data from IoT-based environmental sensors (e.g., temperature, humidity) and physiological monitors would enable a more comprehensive assessment of heat stress by correlating observed behaviors with empirical data. These advanced models must be optimized for real-time performance on edge devices such as NVIDIA Jetson platforms for practical on-farm implementation. The deployment of a resulting real-time alert system would then empower farm managers to receive notifications of critical events, such as early signs of heat stress or problematic aggressive interactions, enabling prompt and targeted interventions. Collectively, these enhancements will advance the system toward a fully autonomous, multi-modal, intelligent livestock monitoring platform capable of supporting precision agriculture and improving animal health outcomes. We are also working on incorporating environmental (e.g., ambient temperature, humidity, THI) and physiological measures (e.g., respiration rate, rectal temperature) that would enable a more direct linkage between visual behavioral indicators and thermotolerance.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect NSF's views.

DATA AVAILABILITY STATEMENT

The data and source code that support the findings of this study are available online. **Source Code**: https://github.com/nseudondian/dairy_cow_behavior_detection.
Video Dataset: https://zenodo.org/records/17918160.

REFERENCES

[1] P. J. Hansen, "Effects of heat stress on mammalian reproduction," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 364, no. 1534, pp. 3341–3350, Nov. 2009.

[2] O. Inadagbo, G. Makowski, A. A. Ahmed, and C. Daigle, "On developing a machine learning-based approach for the automatic characterization of behavioral phenotypes for dairy cows relevant to thermotolerance," *AgriEngineering*, vol. 6, no. 3, pp. 2656–2677, Aug. 2024.

[3] N. R. St-Pierre, B. Cobanov, and G. Schnitkey, "Economic losses from heat stress by U.S. livestock industries," *J. Dairy Sci.*, vol. 86, pp. E52–E77, Jun. 2003.

[4] M. Gauly et al., "Future consequences and challenges for dairy cow production systems arising from climate change in central Europe—A review," *Animals*, vol. 7, no. 5, pp. 843–859, 2012.

[5] P. K. Thornton, "Livestock production: Recent trends, future prospects," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 365, no. 1554, pp. 2853–2867, Sep. 2010.

[6] A. A. Ahmed and B. Nyarko, "Smart-watcher: An AI-powered IoT monitoring system for small-medium scale premises," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2024, pp. 139–143.

[7] A. A. Ahmed and G. H. Reddy, "A mobile-based system for detecting plant leaf diseases using deep learning," *AgriEngineering*, vol. 3, no. 3, pp. 478–493, Jul. 2021. [Online]. Available: https://www.mdpi.com/2624-7402/3/3/32

[8] A. A. Ahmed and J. Ahmed, "An image-based plant weed detector using machine learning," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2024, pp. 193–197.

[9] W. Grzesiak, D. Zaborski, M. Pluciński, M. Jędrzejczak-Silicka, R. Pilarczyk, and P. Sablik, "The use of selected machine learning methods in dairy cattle farming: A review," *Animals*, vol. 15, no. 14, p. 2033, Jul. 2025.

[10] G. Ceja, S. Paudyal, J. Spencer, J. M. Piñeiro, and C. L. Daigle, "149 case study: The impact of a fogging system on dairy cow comfort in cows housed in a barn with tunnel ventilation and an automatic milking system," *J. Animal Sci.*, vol. 101, no. 1, pp. 94–95, May 2023.

[11] M. J. Carabaño, C. Díaz, and M. Ramón, "Breeding for thermotolerance in dairy cattle: Production versus fertility traits," *J. Dairy Sci.*, vol. 108, no. 9, pp. 9915–9929, Sep. 2025.

[12] W. Andrew, J. Gao, S. Mullan, N. Campbell, A. W. Dowsey, and T. Burghardt, "Visual identification of individual holstein-friesian cattle via deep metric learning," *Comput. Electron. Agricult.*, vol. 185, Jun. 2021, Art. no. 106133, doi: 10.1016/j.compag.2021.106133.

[13] (2025). *Ultralytics YOLOv8*. Accessed: Dec. 10, 2025. [Online]. Available: https://docs.ultralytics.com/models/yolov8/

[14] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," 2017, *arXiv:1703.07402*.

[15] A. L. Schaefer, N. J. Cook, J. S. Church, J. Basarab, B. Perry, C. Miller, and A. K. W. Tong, "The use of infrared thermography as an early indicator of bovine respiratory disease complex in calves," *Res. Veterinary Sci.*, vol. 83, no. 3, pp. 376–384, Dec. 2007.

[16] S. G. Matthews, A. L. Miller, T. PlÖtz, and I. Kyriazakis, "Automated tracking to measure behavioural changes in pigs for health and welfare monitoring," *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, Dec. 2017.

[17] E. Fazzari, D. Romano, F. Falchi, and C. Stefanini, "Animal behavior analysis methods using deep learning: A survey," *Expert Syst. Appl.*, vol. 289, Sep. 2025, Art. no. 128330.

[18] S. Mahato and S. Neethirajan, "Integrating artificial intelligence in dairy farm management–biometric facial recognition for cows," *Inf. Process. Agricult.*, vol. 12, no. 3, pp. 312–325, 2024.

[19] D. Shao, Z. He, H. Fan, and K. Sun, "Detection of cattle key parts based on the improved YOLOv5 algorithm," *Agriculture*, vol. 13, no. 6, p. 1110, May 2023. [Online]. Available: https://www.mdpi.com/2077-0472/13/6/1110

[20] S. M. C. Porto, C. Arcidiacono, U. Anguzza, and G. Cascone, "Video-based monitoring system to detect and classify cattle behaviors using machine learning," *Animals*, vol. 10, no. 10, pp. 1–15, 2020.

[21] A. Nasirahmadi, S. A. Edwards, and B. Sturm, "Deep learning and computer vision for automated cattle welfare monitoring: A review," *Comput. Electron. Agricult.*, vol. 174, pp. 1–21, Jan. 2020.

[22] C. Wu, J. Fang, X. Wang, and Y. Zhao, "DMSF-YOLO: Cow behavior recognition algorithm based on dynamic mechanism and multi-scale feature fusion," *Sensors*, vol. 25, no. 11, p. 3479, May 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/11/3479

[23] J. Porto, G. Higa, V. Weber, F. Weber, N. Loebens, P. Claure, L. de Almeida, K. Porto, and H. Pistori, "A new Siamese network loss for cattle facial recognition in a few-shot learning scenario," *AgriEngineering*, vol. 6, no. 3, pp. 2941–2954, Aug. 2024.

[24] A. Rohan, M. S. Rafaq, M. J. Hasan, F. Asghar, A. K. Bashir, and T. Dottorini, "Application of deep learning for livestock behaviour recognition: A systematic literature review," 2023, *arXiv:2310.13483*.

[25] P. M. Shakeel, B. B. M. Aboobaider, and L. B. Salahuddin, "A deep learning-based cow behavior recognition scheme for improving cattle behavior modeling in smart farming," *Internet Things*, vol. 19, Aug. 2022, Art. no. 100539.

[26] Y.-C. Tsai, J.-T. Hsu, S.-T. Ding, D. J. A. Rustia, and T.-T. Lin, "Assessment of dairy cow heat stress by monitoring drinking behaviour using an embedded imaging system," *Biosyst. Eng.*, vol. 199, pp. 97–108, Nov. 2020.

[27] Y. Qiao, Y. Guo, and D. He, "Cattle body detection based on YOLOv5-ASFF for precision livestock farming," *Comput. Electron. Agricult.*, vol. 204, Jan. 2023, Art. no. 107579.

[28] A. Fuentes, S. Yoon, J. Park, and D. S. Park, "Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information," *Comput. Electron. Agricult.*, vol. 177, Oct. 2020, Art. no. 105627.

[29] B. Wang, X. Li, X. An, W. Duan, Y. Wang, D. Wang, and J. Qi, "Open-set recognition of individual cows based on spatial feature transformation and metric learning," *Animals*, vol. 14, no. 8, p. 1175, Apr. 2024.

[30] Z. Yu, Y. Liu, S. Yu, R. Wang, Z. Song, Y. Yan, F. Li, Z. Wang, and F. Tian, "Automatic detection method of dairy cow feeding behaviour based on YOLO improved model and edge computing," *Sensors*, vol. 22, no. 9, p. 3271, Apr. 2022.

[31] M. N. Islam, J. Yoder, A. Nasiri, R. T. Burns, and H. Gan, "Analysis of the drinking behavior of beef cattle using computer vision," *Animals*, vol. 13, no. 18, p. 2984, Sep. 2023.

[32] A. Fuentes, S. Han, M. F. Nasir, J. Park, S. Yoon, and D. S. Park, "Multiview monitoring of individual cattle behavior based on action recognition in closed barns using deep learning," *Animals*, vol. 13, no. 12, p. 2020, Jun. 2023.

[33] Q. Jia, J. Yang, S. Han, Z. Du, and J. Liu, "CAMLLA-YOLOv8n: Cow behavior recognition based on improved YOLOv8n," *Animals*, vol. 14, no. 20, p. 3033, Oct. 2024.

[34] T. Cheng, Z. Zhang, B. Zheng, and L. Fu, "Cattle behavior recognition using yolov5 and deepsort tracking in precision livestock farming," *Sensors*, vol. 22, no. 9, pp. 20–34, 2022.

[35] FFmeg. *Cross-Platform Solution to Record, Convert Stream Audio Video*. Accessed: Dec. 10, 2025. [Online]. Available: https://ffmpeg.org/.

[36] Roboflow. *Computer Vision Tools for Developers and Enterprises*. Accessed: Dec. 10, 2025. [Online]. Available: https://roboflow.com

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[38] Pallets Projects. (2024). *Flask: A Python Microframework*. Accessed: Dec. 2025. [Online]. Available: https://flask.palletsprojects.com/

[39] S. Consortium. (2025). *SQLite: Version 3.50.4*. Accessed: Dec. 2025. [Online]. Available: https://www.sqlite.org/

[40] Dell Technologies. (2025). *Alienware Aurora R16 Gaming Desktop*. Accessed: Dec. 2025. [Online]. Available: https://www.dell.com/alienware-aurora-r16

[41] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 8026–8037.

[42] K. Thomas, R.-K. Benjamin, P. Fernando, G. Brian, B. Matthias, F. Jonathan, K. Kyle, H. Jessica, G. Jason, C. Sylvain, I. Paul, A. Damián, A. Safia, W. Carol, and J. Development Team, "Jupyter notebooks—A publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, Eds., IOS Press, 2016, pp. 87–90, doi: 10.3233/978-1-61499-649-1-87.

[43] PostgreSQL Global Development Group. (2025). *PostgreSQL: The World's Most Advanced Open Source Database*. Accessed: Dec. 2025. [Online]. Available: https://www.postgresql.org/

[44] Amazon Web Services. (2025). *Amazon RDS: Relational Database Service*. Accessed: Dec. 2025. [Online]. Available: https://aws.amazon.com/rds/

[45] Google Cloud. (2025). *Google Cloud SQL: Managed Database Service*. Accessed: Dec. 2025. [Online]. Available: https://cloud.google.com/sql

[46] B. Burns, B. Grant, D. Oppenheimer, E. Brewer, and J. Wilkes, "Borg, omega, and kubernetes," *Commun. ACM*, vol. 59, no. 5, pp. 50–57, Apr. 2016.

[47] Docker Inc. (2025). *Docker Swarm: Native Clustering for Docker*. Accessed: Dec. 2025. [Online]. Available: https://docs.docker.com/engine/swarm/

[48] O. Friard and M. Gamba, "BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations," *Methods Ecol. Evol.*, vol. 7, no. 11, pp. 1325–1330, Nov. 2016.

[49] O. Siméoni et al., "DINOv3," 2025, *arXiv:2508.10104*.

[50] T. Meinhardt, A. Kirillov, L. Leal-Taixé, and C. Feichtenhofer, "Track-Former: End-to-end single object tracking with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jun. 2022, pp. 8022–8032.

[51] A. M. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3992–4003.

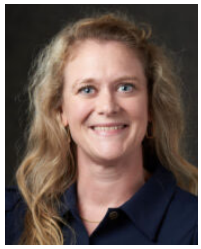
NSEABASI S. UDONDIAN received the master's degrees in computer science and in information technology. She is a Software Engineer and a Researcher with a background in computer vision, artificial intelligence, and sustainability-focused applications. Her research focuses on AI-enabled behavioral analysis in livestock for precision agriculture, with applications in thermotolerance and animal welfare. Her broader interests include building scalable AI/ML infrastructure and developing technologies that contribute to social impact and sustainable development.

DANIEL OGBUIGWE received the B.Eng. degree in mechanical engineering from Baze University, Abuja. He is currently pursuing the M.S. degree in computer science with Prairie View A&M University. He works as a Research Assistant with Prairie View A&M University. His research interests include machine learning, artificial intelligence, and accelerated computing on GPUs. He has previously conducted research on speech recognition for under-resourced languages in Nigeria.

MOSTAFA MAHMOUD SELIM is currently pursuing the M.Sc. degree with the Computer Science Department, Prairie View A&M University, under the Computer Information System (CIS) Program. His research interests lie at the intersection of data science and artificial intelligence. His research focuses on AI-enabled applications in the domains of animal science and transportation, with applications in the IoT, and sensing-based edge computing systems.

AHMED ABDELMOAMEN AHMED (Member, IEEE) received the B.Sc. and M.S. degrees in computer science from Cairo University and the Ph.D. degree in computer science from the University of Saskatchewan, in August 2017. He joined the Department of Computer Science, PVAMU, in Fall 2017. He is currently an Associate Professor with the Computer Science Department, Prairie View A&M University, Prairie View, TX, USA. His research interests are in the fields of Parallel and Distributed Systems with applications in various domains, such as the IoT and AI. Since joining PVAMU, he has been actively involved in several research projects, including the IoT and machine learning applications.

MADISON E. BACON received the B.A. degree in psychology from the University of Wisconsin-Eau Claire, the M.Sc. degree in applied animal behavior and animal welfare from The University of Edinburgh, and the Ph.D. degree in animal science from the University of Minnesota in 2025. She is currently a Postdoctoral Research Associate with the Animal Science Department, Texas A&M University, College Station, TX, USA. Her research interests include dairy cattle behavior and welfare.



COURTNEY L. DAIGLE received the B.S. degree in zoology from Oklahoma State University, and the M.S. degree zoo and aquarium management and the Ph.D. degree in animal science from Michigan State University. She has been an Associate Professor with Texas A&M University, College Station, since 2016. Prior to this, she completed the Postdoctoral and Research Fellowships with Purdue University and Wageningen University. She has secured over $2.5 million in research funding and authored 48 peer-reviewed publications, six book chapters, and more than 80 conference abstracts. Her teaching portfolio spans over 800 students across diverse courses in animal behavior and welfare, including leadership in the International Beef Cattle Academy and the Animal Welfare Judging Team, which has earned multiple national and international accolades under her guidance. She actively contributes to professional organizations such as the International Society for Applied reflecting her commitment to advancing animal welfare through research, education, and industry collaboration.