

Advances in precision dairy and beef farming technologies

Edited by Professor Daniel Berckmans and Professor Tomás Norton,
Katholieke University of Leuven, Belgium

E-CHAPTER FROM THIS BOOK



The use of computer vision technologies for monitoring the behavior of dairy cattle

Oleksiy Guzhva, Swedish University of Agricultural sciences, Sweden; Marwa Mahmoud and Ozgur Civan Dogan, University of Glasgow, UK; David Berthet, Sony Nordic (Sweden), Sweden; and Niclas Högberg, Swedish University of Agricultural Sciences, Sweden

- 1 Introduction
- 2 Bridging the gap: where do computer vision and animal science meet?
- 3 Importance of model interpretability in cattle behavior monitoring
- 4 From theory to practice: how do we apply advanced computer vision algorithms in real-world scenarios?
- 5 Case study 1: the use of scalable computer vision algorithms for calving event monitoring
- 6 Case study 2: dairy cattle behaviour tracking with pose estimation models
- 7 Case study 3: implementing a 3D pose estimation system for cow behavior tracking
- 8 Conclusion
- 9 Where to look for further information
- 10 Future trends in research
- 11 References

1 Introduction

The dairy industry faces increasing demands to enhance productivity, animal welfare, and environmental sustainability. Precision livestock farming (PLF) has emerged as a vital approach to meet these challenges by integrating advanced technologies for real-time monitoring and management of livestock (Wathes et al., 2005; Berckmans, 2017; Norton et al., 2019). Within PLF, computer vision (CV) has gained significant attention as a non-invasive and efficient method for studying dairy cattle behavior and health (García et al., 2020; Li et al.,

2021). Over the past decade, advancements in computer/machine vision and artificial intelligence (AI) have enabled automated analysis of visual data and provided better opportunities for it to be collected continuously, offering new possibilities for monitoring individual animals and entire herds.

Traditionally, monitoring dairy cattle behavior has relied on visual/manual observation and the use of wearable sensors (Stygar et al., 2021). While these methods have provided valuable insights, they come with limitations such as labor intensiveness, potential stress to animals, and limited data granularity (Cockburn, 2020). This labor intensiveness could be related to the fact that each sensor must be individually purchased, attached, calibrated, and regularly maintained – potentially driving up both financial costs per unit and time spent by personnel. Frequent (depending on data resolution and richness) battery charging or replacement also adds to the workload, requiring additional oversight of energy levels and device performance. The advent of CV technology offers a transformative approach to potentially overcome these challenges, at least by minimizing the number of sensors needed for animal monitoring. By leveraging advanced algorithms and machine/deep learning techniques (ML/DL), CV enables non-intrusive, continuous, and automated monitoring of cattle behavior, opening new avenues for research and application in PLF (Guzhva et al., 2016; García et al., 2020; Fuentes et al., 2020; Li et al., 2021).

Previous and recent studies have demonstrated the potential of CV systems in various applications, such as lameness detection (Kang et al., 2021), body condition scoring (Alvarez et al., 2018; Nir et al., 2018), feeding behavior analysis (Saar et al., 2022), and estrus detection (Lodkaew et al., 2023). For instance, the use of DL algorithms, particularly convolutional neural networks (CNNs), has improved the accuracy of identifying and classifying cattle behaviors from video footage (Kuncheva et al., 2022; Mar et al., 2023). Three-dimensional imaging technologies have also been utilized to assess body conformation and detect subtle changes in posture associated with health issues (Gong et al., 2022; Kroese et al., 2024).

An emerging area in the application of CV techniques within the dairy cattle domain is the use of pose estimation and facial expression analysis to monitor behavior and well-being. Pose estimation involves tracking key points on an animal's body to understand its posture and movement dynamics. This technique allows for detailed analysis of behaviors such as lying, standing, walking, and social interactions (e.g. Gao et al., 2023; Wang et al., 2023). Tools like DeepLabCut have been adapted for use in livestock, enabling precise tracking of animal poses with higher accuracy.

Despite these advancements, several gaps remain in the application of CV/DL in dairy cattle studies. One major challenge is the lack of open, standardized datasets and benchmarks, which hinders the development and comparison of algorithms across different studies (Collins et al., 2008). The amount of time

it takes to collect, select, annotate/label, and prepare a sufficient number of images for building a high-quality dataset is often overlooked due to limited resources and/or initial experiment planning (Wei et al., 2016; Kuncheva et al., 2022). Another layer of complexity is linked to different perceptions of what a CV model is and how it functions, depending on the background of the user. In many cases, CV models utilized by animal scientists are treated as black boxes, lacking interpretability, which makes their widespread use more limited (Buhrmester et al., 2021; Rudin 2019; Rudin and Radin, 2019; Feng et al., 2019). Computer scientists and engineers, on the contrary, might treat the same CV models as glass boxes, understanding their complexity in terms of time and deployment costs. Variability in farm environments, such as lighting, lack of standardized area layout, and different ways to integrate farm equipment (e.g. robots, cubicles, feeding area), also affects the deployment effort and cost of CV systems in real-world scenarios. Ensuring data privacy and addressing ethical considerations in data collection and analysis are also important factors that need attention and add yet another layer of complexity to planning experiments.

In our experience, one of the major challenges related to integrating CV and ML/DL algorithms into cattle research is a distinct separation between technology itself and its translation into animal science vocabulary (Guzhva and Siegford, 2022). In a research project aiming to study cattle behavior using CV and DL, merging the animal-centered and technical tracks is crucial for meaningful outcomes. The first track focuses on developing ethograms – detailed catalogs of cattle behaviors – that are not only scientifically comprehensive but also tailored to what CV models can realistically detect and interpret. The second track deals with the technical and algorithmic challenges of adapting CV models to the nuances of animal behavior in a real-world farm setting. By merging these tracks, researchers ensure that the behavioral analyses are both biologically significant and technically feasible. This synergy enhances the reliability of behavior detection, facilitates the development of more robust algorithms, and ultimately contributes to advancements in both animal science and CV fields.

This chapter will delve into these topics, discussing the current technologies and methodologies employed, evaluating their effectiveness, and highlighting the challenges that need to be addressed to fully realize the potential of CV in dairy cattle behavior analysis. Recognizing this, our chapter addresses the most common challenges researchers encounter and provides guidance through the critical stages of experiment planning. By highlighting key considerations and offering practical solutions, we aim to equip researchers with the tools and knowledge necessary to design effective CV setups tailored to their unique research objectives.

2 Bridging the gap: where do computer vision and animal science meet?

2.1 Challenges with traditional ethograms

Traditional ethograms are often crafted based on human observations, capturing nuanced behaviors that may be difficult for computer algorithms to recognize. These behaviors can be:

- 1 Complex and multi-faceted: Behaviors that involve subtle body language or context-dependent actions;
- 2 Subjectively defined: Variations in definitions and interpretations among researchers;
- 3 Non-standardized: Lack of uniformity in behavior categories across different studies or species;
- 4 Although human interpretation can be extremely rich, any assessment is individual, varies over time, thus degrading consistency.

These characteristics pose significant challenges for CV systems, which require clear, objective, and quantifiable data to accurately detect and classify behaviors. To harness the full potential of CV in animal behavior studies, scientists must reconsider how they develop ethograms. The goal is to create ethograms that are:

- Machine-readable: Behaviors defined in terms that can be detected by algorithms;
- Standardized: Consistent definitions to allow for generalization and scalability;
- Quantifiable: Behaviors broken down into measurable visual features;
- Repeatable: Constant output opening for valuable comparison.

2.2 Strategies for developing computer vision-friendly ethograms

2.2.1 Simplify and standardize behaviors

- Action primitives: Break down complex behaviors into simpler, discrete actions (e.g. 'lifting leg,' 'head turn') that are easier for algorithms to detect;
- Clear definitions: Use precise, unambiguous terminology to describe behaviors, minimizing subjective interpretation.

2.2.2 Focus on observable features

- Visual cues: Emphasize behaviors with distinct visual characteristics, such as specific postures, movements, or interactions;

- Pose estimation: Utilize key anatomical landmarks (e.g. joints, limbs) to define keypoints that both the CV algorithms and the human observer can identify, when creating ground truth; Target a minimized keypoint set, since any additional keypoint in the annotation workflow increases the effort tremendously when creating datasets.

2.2.3 Incorporate multimodal data

- Sensor fusion: Combine video data with other sensors (e.g. accelerometers, RFID, UWB, microphones) to capture behaviors or behavioral proxies that are not easily observable visually;
- Data fusion: Merge different data types to enhance the detection and interpretation of complex behaviors.

2.2.4 Develop annotated, high-quality datasets

- Appropriate pre-experiment planning: CV models require a lot of data for training, and the initial data acquisition and preprocessing take a lot of time;
- High-quality labels: Create extensive datasets with accurately annotated behaviors to train and validate CV models;
- Context-focused training material with high diversity: The richness and the relevance of the training material have a major impact on the performance of the CV algorithms;
- Highly relevant validation material: To understand and improve the performance of a CV algorithm, use distinct training and validation sets, with validation sets focused on the monitored behaviors;
- Open access: Share datasets within the research community to promote collaboration and standardization.

2.2.5 Collaboration between engineers, data scientists and animal scientists

- Interdisciplinary teams: Work alongside computer scientists and engineers to understand the capabilities and limitations of current technologies and to develop a common/shared vocabulary to allow concept development and cross-disciplinary hypothesis testing;
- Algorithm development: Participate in the creation of algorithms tailored to specific behavioral analyses by transferring animal-based features into the computational domain.

2.2.6 Iterative refinement

- Feedback loops: Continuously refine ethograms based on the performance of CV models;

- Validation studies: Conduct experiments to validate that the revised ethograms and algorithms accurately reflect the intended behaviors.

2.2.7 Benefits of revised ethograms

- Enhanced accuracy: Improved detection and classification of behaviors by CV systems since the behaviors will be described with model specifics in mind;
- Scalability: Ability to monitor larger populations over extended periods without the need for constant human observation;
- Real-time monitoring: Immediate detection of critical behaviors (e.g. signs of distress or illness), enabling prompt interventions;
- Data consistency: Standardized behavior definitions lead to more consistent data across studies and species.

2.3 Technical challenges and setting up the computer vision infrastructure

2.3.1 Execution plan

- Plan thoroughly every step of the execution to maximize the outcome. A detailed assessment will minimize drawbacks, efforts, and maximize results.
- Invest time in preparation: Investigate existing methods and achievements, define clear ethograms and targets, and identify key elements in the realization.

2.3.2 No 'off-the-shelf' vision system(s)

- While different cameras can be found in abundance, marker-free CV systems are not off-the-shelf products waiting to be used in experimental setups;
- Vendor-specific camera Application Programming Interface (API) mandates tailor-made software implementations for frame synchronization;
- Setting up a marker-free vision system requires a wide range of engineering expertise (camera management and calibration, network, model training and dataset creation, data management, processing units);
- Complex toolchain of open-source code requiring numerous adjustments for animal-centered scenarios.

2.3.3 3D vs 2D cameras and/or single vs multi-camera setup

- 2D cameras are cost-effective and simple to deploy but lack depth perception, leading to challenges in pose estimation and occlusion

handling. They are suitable for basic behavioral analysis that does not require in-depth information;

- 3D cameras offer enhanced spatial data and improved accuracy in detecting subtle behaviors and interactions. However, they come with higher costs, complexity in setup, limited field of view (FOV) range (<5 m), and greater data management needs, which might make it difficult to be used in commercial farm settings;
- Single-camera (2D) setups are easy to install with lower initial costs but suffer from limited FOV, occlusion challenges, and a lack of redundancy. They may miss important behaviors that require multiple perspectives;
- Multi-camera setups (3D) provide comprehensive coverage and improved accuracy through multiple viewpoints, aiding in better occlusion handling and behavioral analysis. The trade-offs include increased costs, a complex installation process, and higher data management overhead.

2.3.4 Complexity of farm environments

- The complex background of dairy farms and crowded areas with many individuals reduces the efficiency of automated behavior annotation/recognition;
- Achieving an optimal camera specification and setup can be complex due to the physical constraints of the environment (fixation, dirt, distances to target, humidity, light, cable lengths, electrical noises, dust, aggressive gases like NH₃, rodents, high-pressure cleaning systems, insects) and often setup adjustments;
- An aggressive environment sets higher requirements on hardware, regular maintenance, and cleaning, which will further increase the deployment and running costs at the farm level;
- Long cable lengths along power supply lines in wide indoor areas increase frame loss and synchronization issues.

3 Importance of model interpretability in cattle behavior monitoring

Interpretability in machine learning refers to the extent to which a human can understand the cause of a decision made by a model. In cattle behavior monitoring, interpretability is essential for several reasons:

- 1 Trust and adoption: Interpretability bridges the gap between complex algorithms and practical livestock/PLF applications, facilitating wider acceptance;

- 2 Error diagnosis and model improvement: Explainable models allow users to identify and correct errors or biases in the system. Understanding why a model misclassifies a behavior or health condition enables developers to refine algorithms and improve performance;
- 3 Regulatory compliance and ethical considerations: Animal welfare regulations may require explanations for decisions affecting livestock. Interpretable models ensure compliance by providing evidence and rationale behind specific interventions or lack thereof;
- 4 Customized interventions: By understanding the factors influencing a model's prediction, farmers can tailor interventions to individual animals, improving welfare and productivity outcomes.

3.1 Issues related to the use of computer vision/deep learning models in animal science domain

In the development of CV models for animal behavior studies, researchers often grapple with a critical trade-off among model size, performance, and the time required for annotation and training. Larger models, such as deep CNNs with numerous layers and parameters, typically achieve higher accuracy and better performance in tasks like object detection, pose estimation, and behavior classification. However, the complex architecture of CNNs makes them inherently opaque due to:

- Complexity: The multi-layered nonlinear transformations in DL models are difficult to parse and interpret, making it challenging to trace input features to output decisions;
- Data dependency: These models require large amounts of data, and their decision boundaries are shaped by the training data, which may contain biases or be unrepresentative of all scenarios in cattle behavior;
- Overfitting and generalization: Without interpretability, it is hard to detect overfitting, where a model performs well on training data but poorly on unseen data. This is critical in diverse farm environments.

When model complexity, data dependency, and problems with overfitting and generalization circle back to ethograms and how they are created, the need for a consistent annotation strategy becomes even more prominent.

Annotation challenges are particularly pronounced in animal behavior studies due to the complexity and subtlety of behaviors, as well as the necessity for expert knowledge to accurately label data. For instance, distinguishing between similar behaviors like feeding and rumination in cattle requires detailed observation and expertise. The time investment for annotating large datasets can slow down research progress and limit the scalability of CV applications in

this field. To address these issues, several strategies have been proposed and implemented:

Accurate, reliable, and handcrafted dataset:

- A diverse, high-quality dataset ensures models can accurately recognize and classify a wide range of behaviors. This, in turn, boosts the reliability of any downstream analyses or decisions that depend on the CV outputs.
- Including varied conditions (e.g. lighting, camera angles, different species or subspecies) helps models generalize well across real-world scenarios, reducing biases and improving model performance in unseen conditions.
- By providing a comprehensive representation of the animals' behaviors, robust datasets enable deeper biological insights, ensuring findings are backed by strong, quantitative evidence.

3.1.1 Data augmentation and synthetic data

- Data augmentation techniques such as rotation, scaling, flipping, and adding noise can artificially expand the dataset without additional annotation efforts.
- Synthetic data generation using simulations or generative models can provide additional training examples. These methods enhance model robustness and generalization by exposing it to a wider variety of scenarios.

3.1.2 Automated annotation tools

- Utilizing semi-automated annotation software (e.g. Labelbox, Roboflow, CVAT) that incorporates techniques like object tracking can speed up the labeling process.
- Tools that suggest annotations based on model predictions allow experts to validate and correct labels rather than annotate from scratch. This reduces manual effort and accelerates dataset preparation.

3.1.3 Collaborative data sharing and standardization

- Establishing open-access repositories and standardized datasets facilitates data sharing among researchers.
- Collaborative efforts reduce duplication of annotation work and promote the development of benchmark datasets for the community.
- Shared resources accelerate progress and improve model comparability across studies.

3.1.4 Domain-specific pretraining

- Pretraining models on related animal datasets before fine-tuning on the target species can improve performance.
- For example, models trained on common livestock behaviors can be adapted to specific breeds or environments with minimal additional data.

3.1.5 Transfer learning and fine-tuning

- Utilizing pre-trained models on large datasets like ImageNet allows researchers to fine-tune these models on smaller, domain-specific datasets. This approach reduces the need for extensive annotation while maintaining high performance.
- For example, pre-trained human pose estimation models can be adapted for cattle by retraining the final layers with a smaller set of annotated cattle images.

3.1.6 Semi-supervised and unsupervised learning

- Leveraging unlabeled data through semi-supervised learning can improve model performance with fewer labeled examples.
- Self-supervised learning methods enable models to learn useful features from unlabeled data, which is abundant in animal monitoring systems. This reduces the annotation burden while still benefiting from large datasets.

3.1.7 Active learning

- Active learning algorithms identify the most informative samples for annotation, optimizing the use of annotation resources.
- By focusing on uncertain or misclassified instances, researchers can improve model performance with fewer annotated examples. This iterative process accelerates training while maintaining or enhancing accuracy.

3.1.8 Efficient model architectures

- Developing or adopting lightweight models like MobileNets or SqueezeNet can achieve a balance between performance and computational efficiency.
- These models require less computational power and can be trained faster, making them suitable for real-time applications and deployment in farm environments.

Despite the seemingly complex set of requirements for successful CV system implementation for cattle behavioral studies, the generalized workflow is

rather straightforward (Fig. 1). It all distills down to careful conceptualization/ visualization of the research question and a detailed step-by-step description of how it translates into programming and what the CV model can achieve under the right circumstances. Another crucial step influencing Data Preprocessing and Augmentation steps is clarity and quality of instructions for annotators who will create ground truth from images and videos to allow further model development.

Selecting the appropriate CV model for animal behavioral studies is a complex task that significantly influences the success of the research project and its potential outcomes. The model architecture plays a pivotal role in determining not only the accuracy and robustness of the system but also its applicability to specific behavioral analyses and further transfer toward more commercial applications. One of the primary difficulties in choosing the right CV model stems from the need to balance model complexity with practical

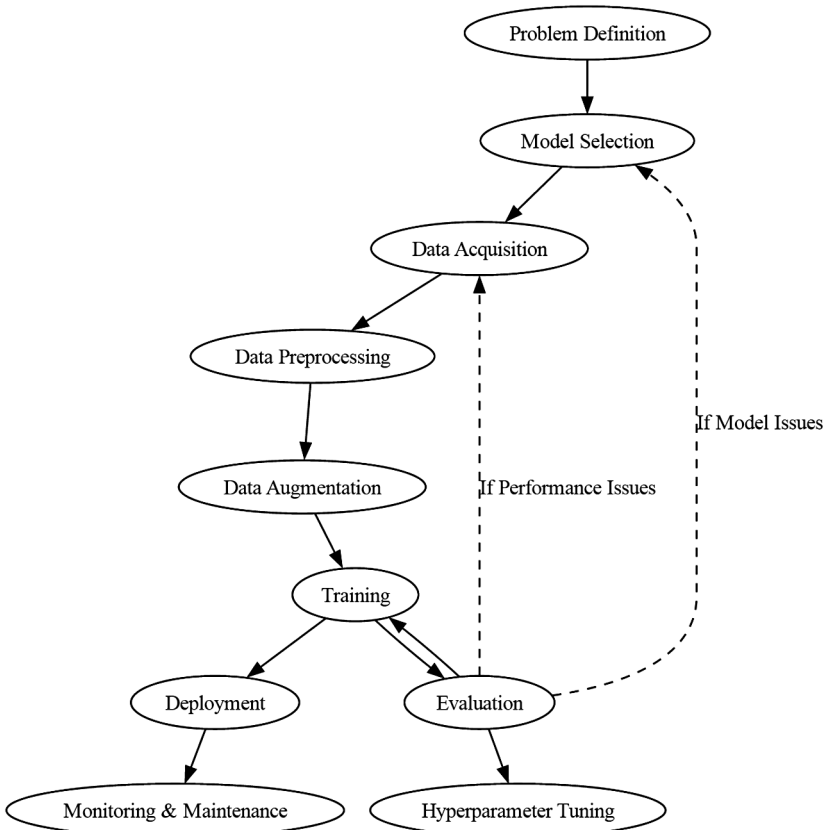


Figure 1 Generalized workflow for computer vision (CV) model development and deployment.

considerations such as computational resources, data availability, and the specific objectives of the study.

Figure 2 shows the simplified overview of the CV model architecture, with further details being influenced by both study requirements and the model family itself.

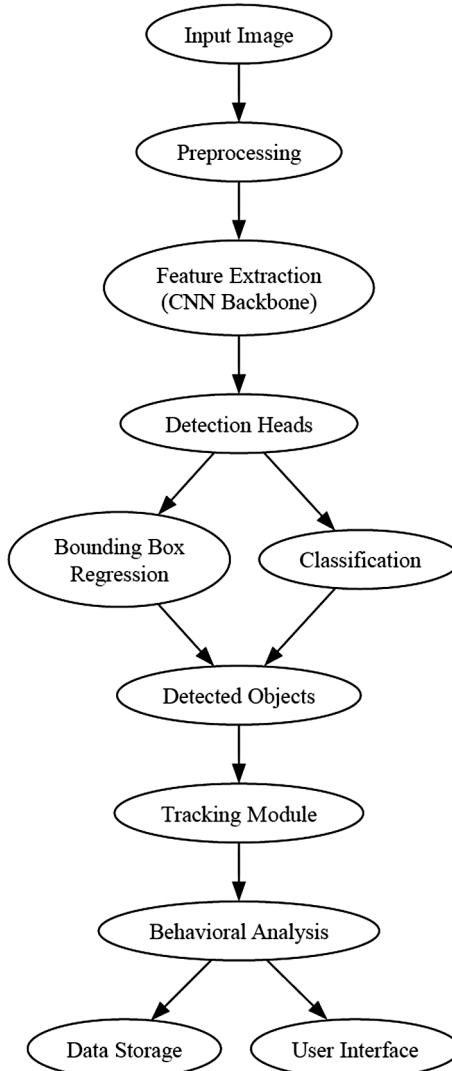


Figure 2 Simplified computer vision (CV) model architecture with main structural components affecting the outcome of behavioral studies and final model performance in real-world scenarios.

Different behavioral studies may require varying levels of detail and specificity from the CV models. For instance, pose estimation models like DeepLabCut or OpenPose are designed to capture fine-grained movements and are well-suited for studies focusing on detailed posture analysis or subtle behavioral cues in cattle. These models leverage deep neural network architectures with multiple layers that can learn complex patterns from the data. However, they often require large amounts of annotated data and significant computational power for training and real-time processing.

On the other hand, if the study's goal is to monitor general activity patterns or detect the presence or absence of animals in a given area, object detection models such as YOLO (you only look once) or single shot multibox detector might be more appropriate. These models are optimized for speed and can operate effectively with fewer computational resources, but they may not capture intricate details of animal behavior.

The environmental conditions of the study also impact the choice of model architecture. Models deployed in uncontrolled farm environments must contend with varying lighting conditions, occlusions, and background clutter. Architectures that incorporate attention mechanisms or are trained with robust data augmentation techniques can improve performance under these challenging conditions, but may add to the model's complexity.

The trade-off between model size and performance is also crucial. Larger models with more parameters can capture complex patterns but are prone to higher computational costs and longer inference times, which may not be feasible for real-time monitoring systems. Contrariwise, smaller models may offer faster processing but at the expense of reduced accuracy.

The final choice of CV model architecture profoundly affects the outcomes and applicability of behavioral studies in cattle (and other species). Researchers must carefully consider the specific requirements of their study, including the types of behaviors to be analyzed, environmental factors, available resources, and the feasibility of data collection and annotation. By aligning the model architecture with these considerations, it is possible to develop effective CV solutions that enhance our understanding of animal behavior while remaining practical and scalable within the constraints of the animal science domain (Table 1).

3.2 Approaches to achieving computer vision model interpretability

To address these challenges, researchers have explored various methods to make computer vision models more interpretable:

- Explainable AI (XAI) techniques: Methods such as saliency maps, attention mechanisms, and layer-wise relevance propagation help visualize which

Table 1 The most used CV models and the prerequisites for their implementation alongside the expected performance and initial data/time investments

Architecture	Use case	Resource cost	Dataset size	Extension possibilities	F1 score (approx.)
YOLOX	Static & video object detection	High (GPU intensive)	Large	Good (custom heads, plugins)	High
Mask R-CNN	Static detection & segmentation	Very high	Very large	Good (modular, flexible API)	Very high
Detectron2	Static & video object detection	Very high	Large	Excellent (highly extensible)	High
ResNet	Backbone for various tasks	Moderate	Large	Good (widely used as backbone)	Depends on use case
U-Net	Image segmentation	Moderate	Medium	Limited (mostly segmentation)	High in segmentation tasks
DINOv2	Self-supervised learning	High	Large	Limited to specific tasks	N/A
DeeplabV3	Efficient object detection	Moderate	Medium	Moderate	High
EfficientNet	Varied (scalable architecture)	High	Large	Good (scalable, versatile)	High

parts of an image contribute most to the model's prediction. For instance, in lameness detection, heatmaps can highlight specific limb movements influencing the decision.

- **Model simplification:** Using simpler models or combining complex models with interpretable ones (e.g. hybrid models) can balance performance and transparency. Decision trees or rule-based systems, though less powerful, offer greater interpretability. Aside from improving the computational performance and final model accuracy, the simpler models are easier to adapt to varying on-farm conditions and different production scenarios, bringing more real-world value to the farmer.
- **Post-hoc explanations:** Tools like Local Interpretable Model-agnostic Explanations provide explanations for individual predictions by approximating the black-box model locally with an interpretable model.
- **Domain-specific visualizations:** Developing visualization tools tailored to cattle behavior can help users intuitively understand model outputs. For example, overlaying detected gait patterns on video frames can illustrate how the model assesses locomotion.

Another crucial component interlinked with CV model development and interpretability is the lack of standardized datasets for training as well as difficulties related to estimating the optimal performance-to-resource-investment sample size. The overview of the problem is presented in Table 2.

4 From theory to practice: how do we apply advanced computer vision algorithms in real-world scenarios?

There are several rather important hurdles in cattle behavioral research relying on CV solutions – general algorithm scalability, flexibility, and adaptability, adding to pose estimation analysis, and finally, multi-camera setups and CV-related infrastructure, which align synergistically when a comprehensive and effective model implementation is needed. Algorithm scalability ensures that CV systems can handle the vast and growing datasets generated by continuous monitoring of large herds, adapting to increasing computational demands without loss of performance. This scalability is essential when integrating multi-camera setups, which expand coverage areas and provide multiple viewpoints to mitigate occlusions and capture more comprehensive behavioral data. The enriched data from these multi-camera systems enable more accurate and robust pose estimation and facial expression analysis, allowing for detailed assessments of cattle postures and movements that are indicative of health and well-being. By building scalable algorithms that can process data from multi-camera networks, researchers can effectively implement advanced CV techniques for precise pose estimation, ultimately leading to more accurate

Table 2 Challenges related to lack of standardized datasets and estimating optimal sample size in dairy cattle behavior monitoring

Challenge	Previously solved by	Currently solved by	Future research
Annotation inconsistencies and data quality	<p>Manual annotation: Individual labeling with varied criteria and standards.</p> <p>Inconsistent labels: Different researchers using varied definitions for behaviors.</p>	<p>Inter-lab protocols: Developing shared annotation guidelines within research groups.</p> <p>Crowdsourcing annotations: Utilizing online platforms, though quality varies.</p> <p>Semi-automated tools: Limited use of tools to assist annotation.</p>	<p>Standardized annotation protocols: Establishing universal guidelines for labeling behaviors and classification of material using detailed description (AI-assisted), i.e. scene description for generating focused datasets.</p> <p>Quality assurance mechanisms: Implementing validation steps to maintain high data quality.</p>
Estimating optimal sample size	<p>Heuristic methods: Relying on rules of thumb without empirical backing.</p> <p>Small sample studies: Limited data leading to overfitting or underpowered models.</p>	<p>Statistical techniques: Applying power analysis based on preliminary studies.</p> <p>Cross-validation: Using available data to estimate model performance.</p> <p>Iterative data collection: Collecting data incrementally based on ongoing analysis.</p>	<p>Adaptive sampling strategies: Developing algorithms that determine sample size dynamically.</p> <p>Resource-efficient data collection: Minimizing unnecessary data collection to conserve resources.</p> <p>Collaborative data gathering: Pooling data across institutions to achieve optimal sample sizes sustainably.</p>
Generalization across diverse environments	<p>Single-environment training: Models trained on data from one farm, limiting applicability.</p> <p>Ignoring environmental variability: Overlooking differences in farm setups, breeds, and management practices.</p>	<p>Multi-environment data collection: Gathering data from various sources to improve robustness.</p> <p>Domain adaptation techniques: Adjusting models for new environments using limited additional data.</p>	<p>Inclusive data collection: Ensuring datasets represent a wide range of environments and conditions.</p> <p>Sustainable data sharing: Facilitating data exchange without excessive resource use.</p> <p>Ethical considerations: Avoiding biases that disadvantage certain farms or regions.</p>
Lack of benchmarking and evaluation standards	<p>Independent evaluations: Researchers using their own metrics, making comparisons difficult.</p> <p>Limited reproducibility: Difficulty in replicating studies due to lack of shared benchmarks.</p>	<p>Community workshops: Initial efforts to standardize evaluation methods.</p> <p>Shared metrics: Adoption of common performance indicators in some subfields.</p>	<p>Standardized benchmarks: Developing widely accepted datasets and metrics for evaluation.</p> <p>Open challenges and competitions: Encouraging innovation through community engagement.</p>

and insightful interpretations of cattle behavior. This integrated approach addresses the complexities of real-world farm environments and is pivotal for the successful deployment of CV technologies in animal behavioral studies as well as their refinement within the PLF domain.

We would like to share our experiences related to these three issues – algorithm complexity and how it affects the final product value, pose estimation and multi-class behavioral analysis for more individual behavioral assessment and lastly, multi-camera setups and how to make those work. The overview of the problem and how the solution was implemented will be presented in the form of three different case scenarios linked to three different research projects.

5 Case study 1: the use of scalable computer vision algorithms for calving event monitoring

5.1 To calve or not to calve

Implementing rigorous management routines for monitoring dairy cattle during the pre- and post-calving periods is vital for ensuring cow and calf welfare and health. Traditional wearable sensor-based methods (accelerometers) aimed at predicting calving onset often demand continuous tracking of multiple behavioral and physiological parameters (Chang et al., 2022). This complexity renders them less feasible for practical, on-farm application, as final model performance and prediction window still make it difficult to be used as a guidance for farm personnel. To address these challenges, CV algorithms present a flexible and non-invasive solution for observing calving boxes with a known number of individuals, offering valuable data for situational analysis and decision support.

5.2 Methodology and dataset

In a pilot study, we proposed a two-step approach for calving monitoring. Two custom-designed CNN detectors were developed for detecting and tracking cows and calves based on continuous data from 16 calving events recorded in a conventional farm environment, where cows were group-housed in deep-straw bedding with free access to a feeding area. See Fig. 3 for the general overview of how the calving box looked.

Each binary detector was trained on 15 000 manually annotated frames sourced from two different farms, classifying images into two distinct classes, such as cow and calf. Additionally, the cow detector's functionality was extended by adding a 7-point shape model (Guzhva et al., 2016) based on anatomically relevant key points to include monitoring of some of the pose changes (standing or lying down). The annotation process applied to each frame consisted of the following

steps and was performed in a custom tool written in the Python programming language: class assignment (cow or calf), selection of anatomical landmarks, and pose confirmation (standing or lying down). However, after the initial testing and for the sake of lowering the computational costs, the 7-point model was trimmed to only include three points – head, root of the neck, and middle back to allow extraction of features such as body orientation and head-to-body angle.

5.3 Results and challenges

The final detectors achieved F1 detection scores of 95.04% for calves, 98.07% for cows, and 98.62% for cow pose estimation (standing/lying down), with a root mean square pixel error of 6.52. For the prediction of the calving events, we utilized video segments spanning 5 h before calving and 4 h after the calving event as input for a custom Kalman Filter-based tracker (Guzhva et al., 2018). This enabled the extraction of continuous animal-based features like general activity (assessed through individual variations in speed and acceleration), head-to-body angle, and posture shifts between standing and lying. The spatiotemporal preference of each cow was investigated for each calving event through individual detection-based heatmaps (Fig. 4).

All the anatomically relevant features, similar or corresponding to those used in classification attempts of calving events based on accelerometer data (Borchers et al., 2017), were extracted for each individual cow and each calving case (70 cows and 16 distinct calving events). These features were visualized to provide a continuous stream of behavioral and activity data to capture the so-called ‘signature patterns’ correlated with the onset of calving (Fig. 5).



Figure 3 Example frame from an overview top-down camera installed in a calving box, covering the main region of interest – the area where calving events occurred.

Dairy cattle exhibit several distinct/signature behavioral patterns correlating with the onset of calving, which can be critical indicators for farmers to provide timely assistance and ensure animal welfare. Key behaviors include:

- Restlessness and increased activity: As calving approaches, cows often become more restless. They may increase their walking time, shift positions frequently, and show signs of discomfort (Lidfors et al., 1994);
- Isolation seeking: Pregnant cows may separate themselves from the herd to find a quiet, secluded area for giving birth. This instinctual behavior helps protect the newborn calf from potential disturbances or threats (Lidfors et al., 1994);
- Frequent lying down and standing up: Cows may lie down and get up more frequently due to discomfort from uterine contractions. This increased movement is associated with the progression of labor (Fadul et al., 2017);
- Frequent head swinging: Cows may frequently turn their head toward their belly due to discomfort from uterine contractions (Barrier et al., 2012).

While the extraction of relevant features based on the CV approach has worked well, predicting the precise onset of calving proved impractical in real-world scenarios due to significant individual variability among cows. In our pilot study, we could only confirm 5 out of 16 calving events based on the output from support vector machine classification algorithm, with a confidence score above 80%, thus rendering the remaining cases inconclusive. Considering that animal caretakers often lack the time and opportunity to intervene during the calving

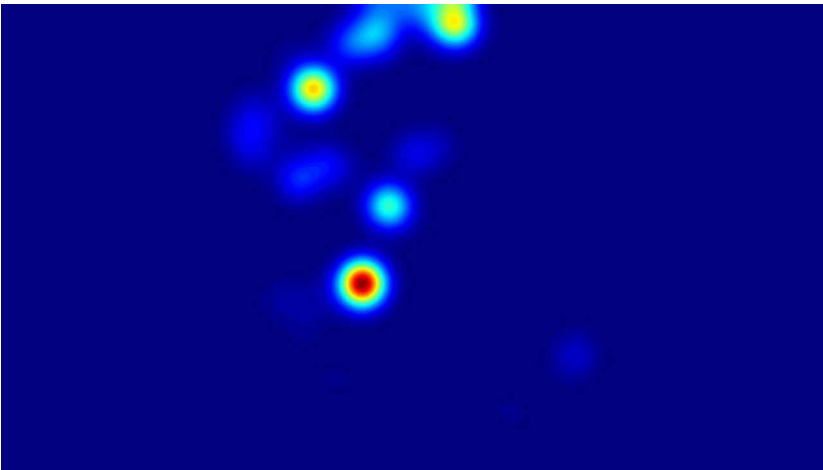


Figure 4 The example of the individual heatmaps produced for each cow highlighting the time spent and preferred location within the intended ROI in a calving box.

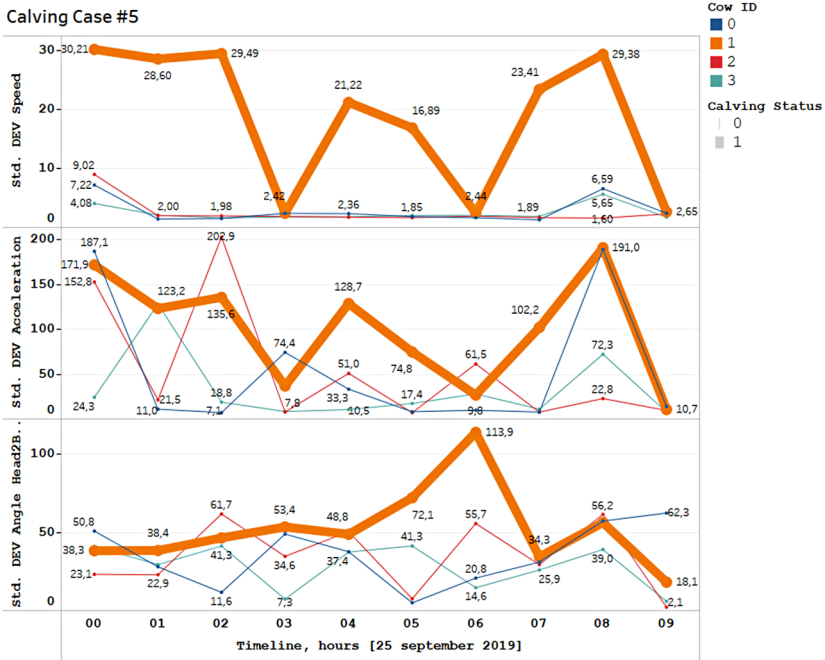


Figure 5 Visualized individual activity streams of cows prior to the calving event, with a bold line corresponding to the cow that calved vs the remaining cows in the calving box.

process, anyway, focusing instead on assisting the newborn calf or mother cow during delivery might be considered a better option.

5.4 Going from complex to basic algorithms

To aid in timely neonatal care and potentially create real added value for farm personnel, we tested a simplified approach based on frame-to-frame object detection. Instead of trying to predict the time window for the onset of calving, the focus was shifted toward the confirmation of calving as a finished event, and calf vitality confirmation. The basic (stripped of all the additional features) cow/calf detector continuously processed frames from a camera installed in the calving box, tallying detections of cow and calf classes. When calf detections surpassed a certain threshold – set to mitigate false positives – an alarm was triggered to indicate that calving had occurred (Fig. 6), and that manual assistance was necessary.

In the context of CV applications for dairy farming, particularly the detection of calving events within calving boxes, model scalability is paramount for practical deployment and product development. While complex models that incorporate multiple features – such as pose estimation, tracking functionality, and behavioral analysis – can provide comprehensive data, they often come

with increased computational demands and potential for inconclusive results due to their complexity. Scientific studies have indicated that simpler models focused on binary classification can offer extreme precision in detecting specific events like calving (Miller et al., 2020; Benaissa et al., 2020; Liseune et al., 2021).

5.5 Research or real-world value for farmers?

Moreover, the scalability of these simpler models makes them more suitable for deployment in agricultural settings, where computational resources may be limited. Evidence from the literature suggests that models with fewer parameters not only require less processing power but also have faster inference times, which is critical for real-time monitoring applications on farms. This aligns with the immediate needs of farmers by providing clear and actionable insights without the ambiguity that can accompany more complex analyses. Therefore, when targeting product development that requires reliability and efficiency, a simple, high-precision binary detector may be more beneficial than a complex model that offers broader analysis but with less definitive outcomes. Focusing on streamlined models enhances the feasibility of widespread adoption and maintenance in practical farming environments, ultimately contributing to improved animal welfare and farm productivity.

6 Case study 2: dairy cattle behaviour tracking with pose estimation models

Automatic cattle behavior classification can open the door for many research areas regarding automating cattle welfare estimation. A lot of practical considerations, including some mentioned before, make this a challenging problem. Lack of large, standardized, labeled datasets is one of the main challenges. Fine-tuning is one of the possible solutions for this problem. In this case study, we investigate the use of pre-trained pose estimation models to extract context-aware animal body pose features as an intermediate step for cattle behavior classification. Using handcrafted features also means that our models are explainable, compared to black-box models. We use a publicly available dataset to test our models and prove that we could achieve an accuracy of 94% using pose data alone, validating our hypothesis that vision-based pose estimation can effectively classify cattle behavior.

6.1 Dataset

For this work, we used the publicly available Video Dataset of Cattle Visual Behaviors dataset CVB (Zia et al., 2023). The dataset contains raw video footage that was captured using GoPro5 Black cameras. These cameras were positioned at four corners of a 25 × 25 meter square area containing 8 Angus

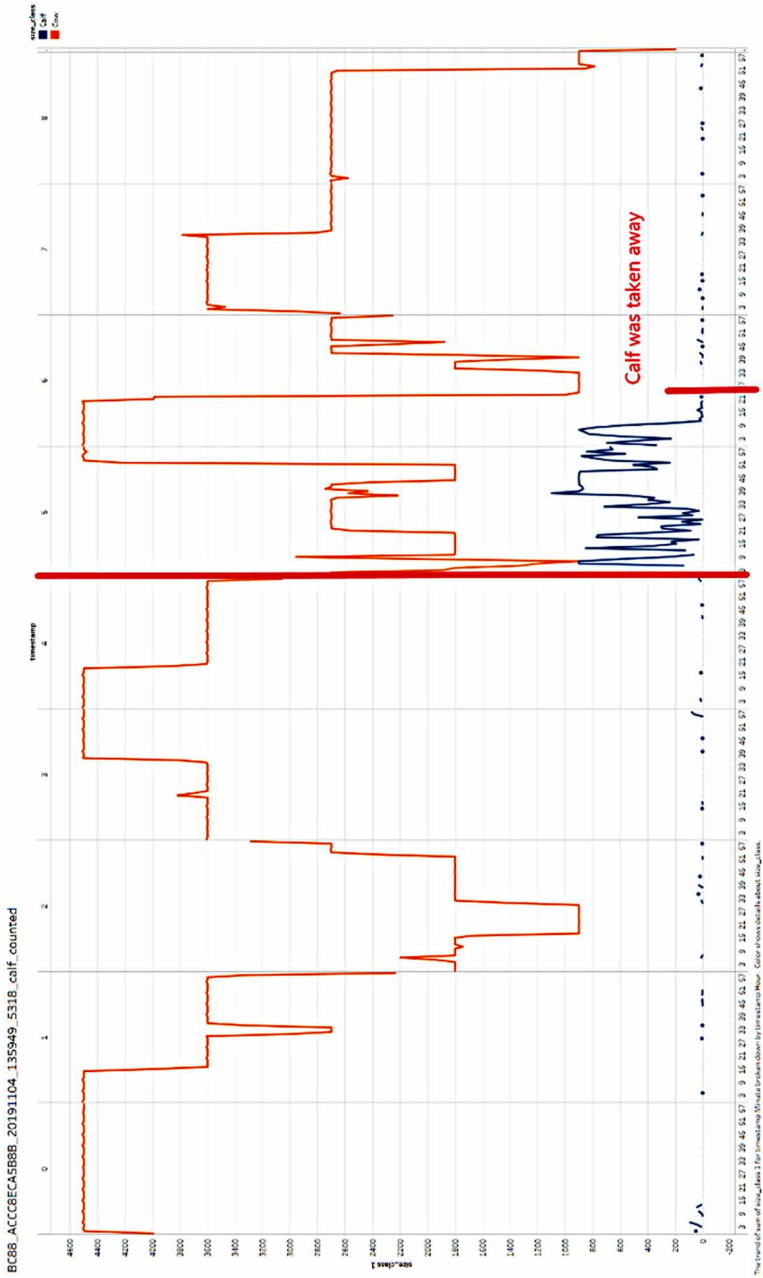


Figure 6 Output from two binary detectors running in parallel, processing the live video data. The orange line indicates the cow class detections, and the blue one is for the calf class. First red line indicates the positive outcome of a calving event (calf is born and detected with 'ok' vitality status due to confirmed activity), while the second one indicates calf being taken away by caretakers, thus explaining the decrease/stop in detection.

beef cattle. This setup was chosen to maximize the coverage, ensuring each behavior is captured. The video data contains various illumination/lightning conditions and four different viewing angles, which can be seen in Fig. 7. The video resolution is 1920×1080 pixels, offering a high quality that is beneficial for identifying behaviors of cattle. The videos are captured at a frame rate of 30 frames per second, sufficient to capture the movements of cattle. The dataset contains 502 sections, each precisely 15 s long, in total approximately 2 h long of video data. The video frames were labeled with 12 behaviors. The video frames were labeled with 12 cattle behaviors, namely: walking, running, ruminating-standing, ruminating-lying, resting-standing, resting-lying, hidden, grooming, grazing, drinking, other, none.

The CVB dataset was annotated by domain experts, ensuring accurate labeling of cattle behaviors. These experts manually assigned a behavior to each cow in every frame. Our analysis of the dataset revealed that nearly 8.6% of cattle change their behavior within 15-s clips. This highlights the challenge of identifying transitions between behaviors in a dynamic setting. More detailly, of these transitions, 7.27% were between two behaviors, 1% between three, and a maximum of five behaviors were observed within a single clip.

It is worth mentioning that the distribution of the labels/behaviors in the dataset was unbalanced, with Grazing and Resting classes appearing in more than 50% of the data (see Fig. 8 for details of all classes' distributions in the dataset), which is common in these kinds of datasets collected in a natural,

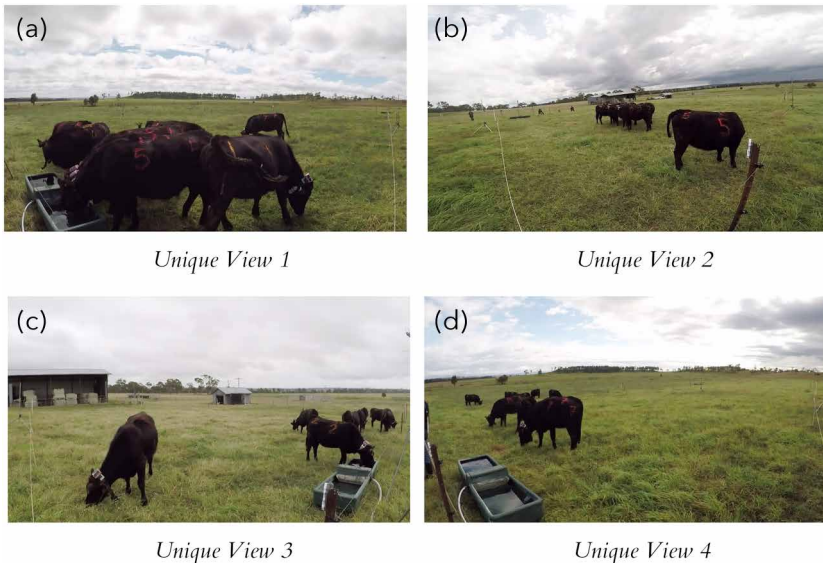


Figure 7 Sample frames from the CVB dataset showing different views captured by the cameras.

uncontrolled environment. This raises some challenges for the machine learning models used for prediction, as the models could get biased in the detection toward the most frequent classes. We solve this by using random subsampling techniques to enhance the classification performance across all the behavior classes and reduce the bias in the data.

6.2 Methodology

Here we describe the detection methodology pipeline as shown in Fig. 9, step by step, namely cows' detection using object detection models, pose estimation to extract intermediate context-aware features, and lastly, behavior classification models and results.

6.3 Animal detection using a fine-tuned object detection model

The first step for processing the frames in the dataset is to detect cows in the frames. We have chosen single-stage detector YOLOv8 due to its compelling balance of speed and accuracy. While training our own object detection

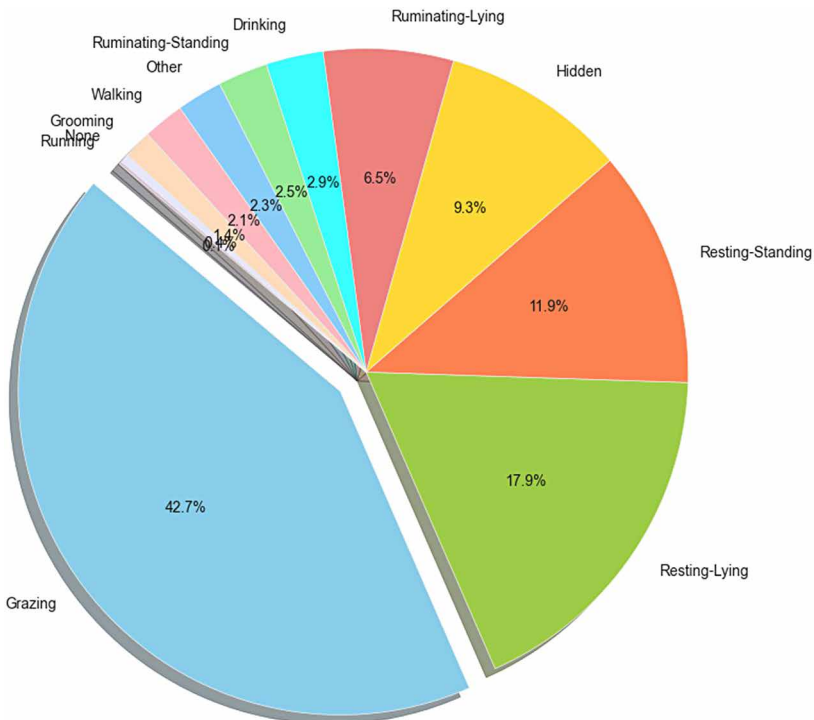


Figure 8 Behavior labels class distribution in the CVB dataset. Note the imbalance in the classes. Data subsampling and augmentation are usually used to mitigate the effect of this imbalance on the machine learning classification models.

model from scratch offers a high degree of customization, there are reasons for our choice of opting for a pre-trained solution like YOLOv8. Our limited dataset, primarily containing images of the same cattle, would make training a generalized model extremely challenging. Considering these constraints and YOLOv8's proven performance in object detection, leveraging a pre-trained model becomes the most practical and efficient choice for our project. All YOLO models are trained on the COCO: Common Objects in Context (COCO) dataset. This dataset offers 80 different classes, including a class: cattle. Considering these, we chose a pre-trained model. Figure 10 shows an example of the model's performance on one image from our dataset. The model detects the cows in the image by generating bounding boxes as well as detection confidence values.

6.4 Pose estimation

Recent work on vision-based cattle behavior analysis relies on black-box models (McDonagh et al., 2021). These models, based on convolutional deep neural networks, attempt to directly classify behaviors (standing, grazing, etc.) from raw images. While black-box models can achieve reasonable accuracy, they suffer from limitations, namely, lack of interpretability as well as sensitivity to variations in the data. Black-box models are data-driven and learn to identify patterns directly from raw images. These patterns can be overly sensitive

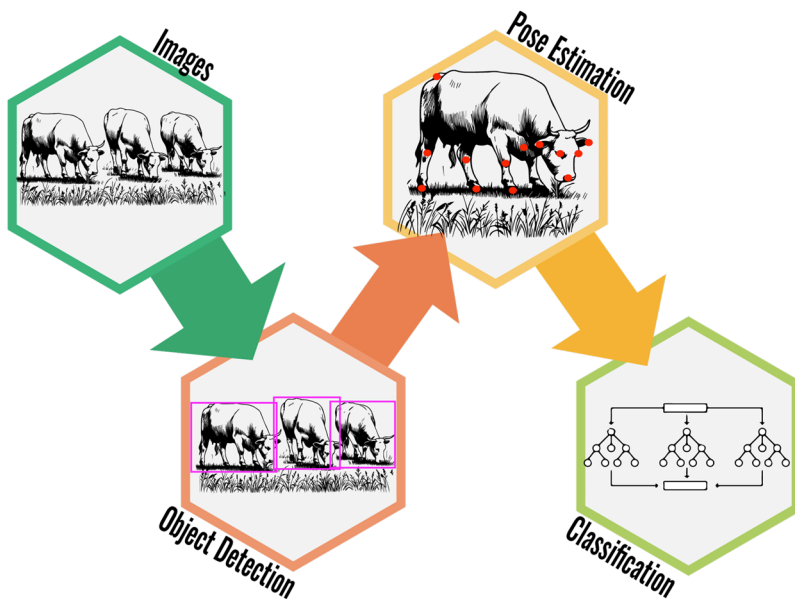


Figure 9 Cattle behavior detection using a post estimation pipeline.

to factors like lighting conditions, background clutter, and camera angles. Variations in these factors can cause the model's performance to degrade in real-world scenarios that differ from the training data. For instance, a model trained on images captured in a brightly lit barn might struggle to accurately classify cattle behaviors in low-light evening footage (Fuentes et al., 2023).

In this study, we extracted cattle pose as intermediate features. For this step, we utilized the ViTPose model (Xu et al., 2022). ViTPose is specifically designed and extensively tested on animal pose datasets in diverse 'in-the-wild' environments, as it has the highest AP (average precision) on the AP-10K dataset. This aligns well with the potential variability in our video footage. Importantly, ViTPose's generalization capability eliminates the need for fine-tuning the model across different environments. Using ViTPose, we extracted 17 key point coordinates for cattle (e.g. locations of joints, nose, etc.) with a confidence score. See Fig. 11 for an example of the extracted keypoints.

Before feeding the extracted keypoints to the machine learning classification model, we performed data normalization. Normalization is a known data preprocessing technique widely used in machine learning to transform data into a common scale. It ensures that all features contribute proportionally to the learning process by bringing them to a common scale. In our case, since we are dealing with spatial coordinates (X-Y), we used Min-Max scaling for normalizing our pose data since it preserves the linear relationships between the points.

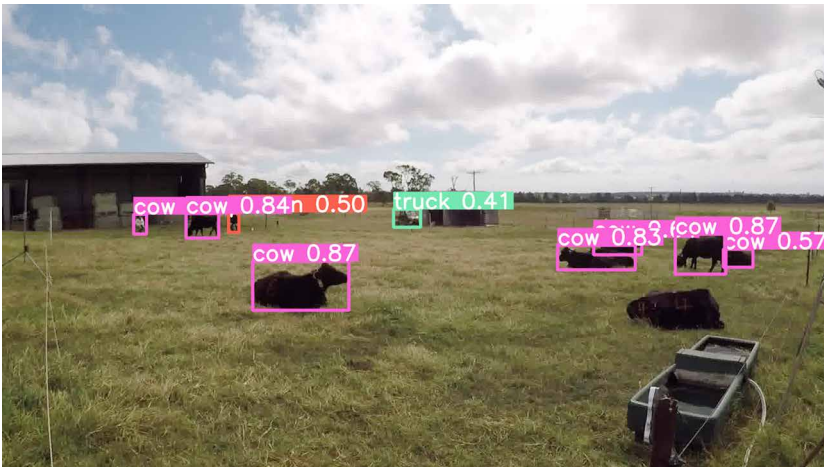


Figure 10 Example of cow detection model performance using a fine-tuned YOLOv8 model. The bounding boxes show the detected cows with the model's confidence displayed at the top of every bounding box.

6.5 Dataset synchronization for pipeline integration

Our pipeline relies solely on the bounding boxes generated by our object detection algorithm – YOLOv8. Therefore, we dynamically matched the annotations after our algorithm had identified cattle within a frame. For dynamically matching the labels with annotations, we employed a two-step technique:

- 1 Center point calculation: We started by calculating the center points of both: key points derived from the pose data of detected cattle and bounding boxes from the manually annotated CVB dataset;
- 2 Proximity-based matching with nearest neighbors: We used the nearest neighbors algorithm (Cover and Hart, 1967) to match each set of key points with the closest corresponding annotation, based on their center points. This method ensures that the matches are meaningful and adhere to a specified distance threshold while preventing incorrect matches.

The synchronization process is designed to bridge the gap between manual annotations and the real-time requirements of the pipeline.

6.6 Behavior classification

The last step is behavior classification, which was the main goal of this study. We have experimented with three popular machine learning models, namely:

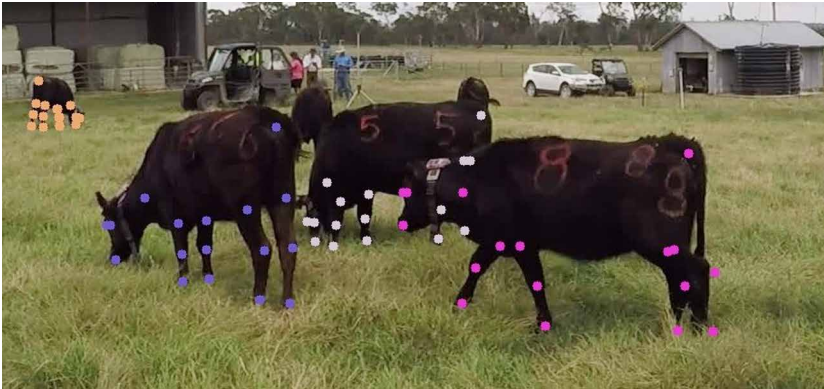


Figure 11 Pose estimation example showing extracted keypoints on the cows, which are intermediate features to be fed to the classification model.

support vector machines (SVM), random forests, and multi-layer perceptron (MLP). For all our experiments, we have used a 10-fold cross-validation strategy.

The best performing model was random forest, with an overall accuracy of 97% and an F1 score of 95%. Figure 12 shows the confusion matrix for the random forest classifier results on our dataset.

On the contrary, SVM and MLP did not perform well in terms of classification results, with an accuracy of 75% for SVM with RBF (Radial basis function kernel) Kernel and 80% for the MLP model. Since this is a 12-class classification problem, the intrinsic binary nature of SVM could have affected the performance, making it not the best model for multi-class classification. Although MLP performed better than SVM, it is still not as good as random forest. The reason could be that random forests are better at finding patterns in small datasets with low-dimensional feature spaces. Random forest proves to be highly adaptable, effectively handling a diverse range of behaviors. This robustness underscores its suitability for classification tasks involving complex datasets with multiple labels.

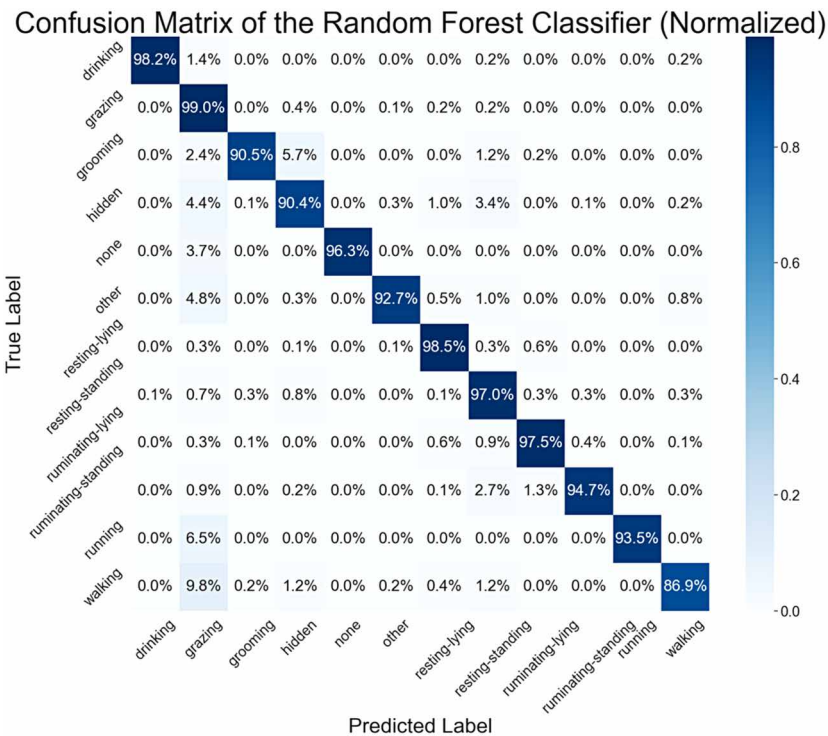


Figure 12 Confusion matrix for the random forest classifier, which was the best-performing model on our dataset with an average overall accuracy of 97%.

7 Case study 3: implementing a 3D pose estimation system for cow behavior tracking

This Section highlights a research collaboration between the Swedish University of Agricultural Sciences and Sony Nordic Sweden, aiming at implementing a multi-camera system and developing 3D pose estimation for cows at the Swedish Livestock Research Centre. A system capable of producing continuous and reliable cow pose estimation in real time. We will summarize the different steps from concept to implementation, detailing the challenges, lessons realized so far, and highlighting the potential of such a system.

A key challenge in developing 3D Pose for cows is the need for high-quality training and validation datasets for the detectors. Open datasets, like the Microsoft COCO dataset with 1346 images including cows comprising 3914 individuals, lack anatomical landmark annotations essential for precise pose estimation, emphasizing the importance of creating domain-specific data for animal science research. The choice between 2D and 3D pose estimation depends on, among other, system maturity, research objectives and project resources. In this specific project, the decision to apply a 3D approach was driven by its availability and the biological need for accuracy to analyse and interpret complex behaviors linked to spatial and diverse resource usage.

7.1 Adaptation and implementation of anatomical landmarks

With the purpose of exploring animal behaviors from a different perspective i.e. by utilizing CV, pose estimation accuracy stood as a key element. Thus, mandating a methodology for annotations striving for simplicity and precision. The definition of the anatomical landmarks to be adopted for pose estimation was chosen to balance biological relevance with consistency in manual annotation. Although a baseline of 37 anatomical landmarks was initially defined (Fig. 13), annotations started with a subset of 12 keypoints and increased gradually to a total of 24 keypoints. We attempted to speed up the annotation work using 30% of synthetically generated images (Fig. 14) and managed to create a dataset comprising over 3000 images and 250 000 annotated keypoints, 80% of which were synthetically generated. As for the model performance using HRNet, 23 out of 24 joints reached over 80% accuracy (Fig. 15).

Additionally, 3D pose estimation relies on the object detector's performance on multiple synchronized sources. Using a dataset comprising over 12 000 images and 80 000 annotations, we reached an AP of over 75% with a 78% recall using a You Only Look Once X (YOLOX) object detector despite recurrent occlusions.

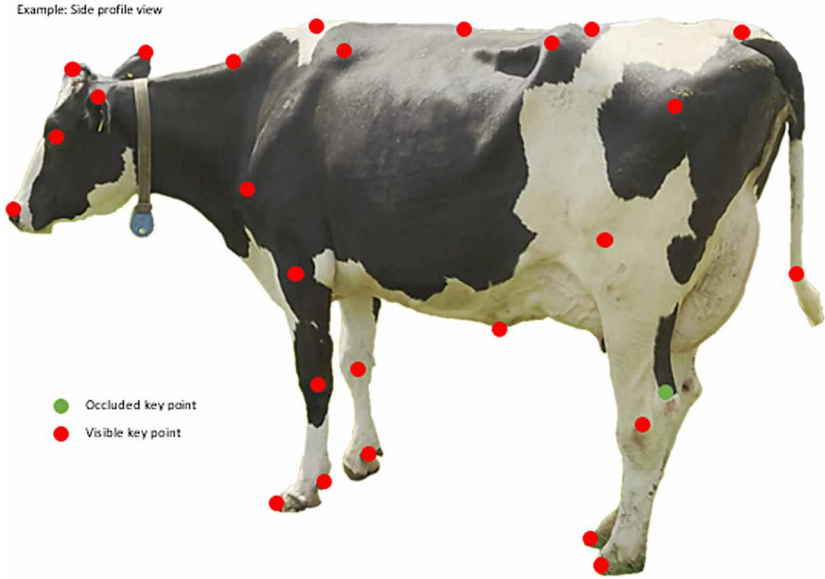


Figure 13 An overview of 37 anatomical landmarks that were initially chosen and defined to be adopted for pose estimation.



Figure 14 An example of a synthetically generated image used for model training.

7.1.1 Focus on observable features

- Select keypoints with anatomical relevance that can be annotated with precision.
- Limit the number of keypoints to a strict minimum.

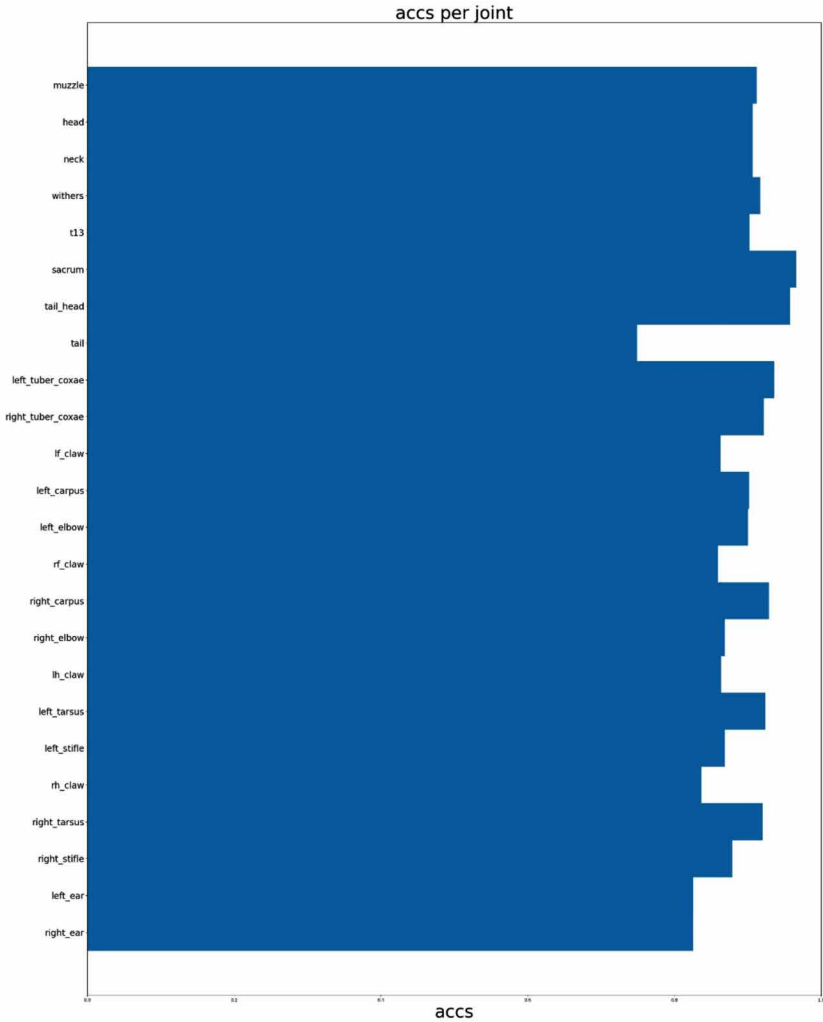


Figure 15 A bar chart highlighting model (HRNet) performances for 24 specific joints using the validation dataset.

- Consider the relevance of a keypoint in relation to the camera views and occlusion.
- Budget time for annotation.

7.2 Exploration phase: early adaptation and development

At the initial phase, the area to be covered by the multi-camera system comprised one voluntary milking group (23 × 29 m) with access to 62 cubicles,

22 individual feed bunks, 2 automated feed dispensers, and 2 mechanical rotating brushes. The group consists of an average population of 55 lactating cows.

As a proof of concept, a small setup based on 7 RGB cameras covering an area including 14 cubicles and the smart gate for the voluntary milking system was set up. The dominant tasks, at this stage, were to identify suitable models for both object detection and pose estimation and build up datasets for model training. This included: (1) capturing images from a wide range of angles, locations, and situations using various devices, such as phones, cameras, and even drones, (2) selecting and archiving the training material, (3) annotating, (4) training the models. It should be noted that this phase required a considerable amount of manual labor. At this point, 4 out of 12 keypoints were retained for evaluating 3D cow pose (Fig. 16). The resulting 3D pose was rudimentary, but sufficient for monitoring several behaviors, including detection of posture transition in cubicles.

7.3 Learning phase: misalignment between expectations and practical outcomes

At this stage, system capabilities were often overestimated, leading to a misalignment between expectations and practical outcomes. At one point, 42 RGB cameras were deployed, generating vast amounts of data requiring significant effort to process and manage. Without clear research goals, much of the data collected was not optimized and redundant, creating additional workload and inefficiencies. This highlighted the importance of aligning

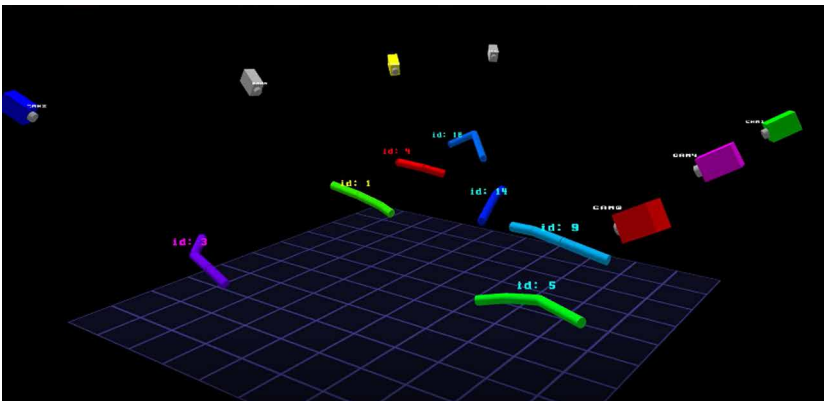


Figure 16 3D evaluation of cow posture using 4 keypoints of the back line (October 2021).

system capabilities with precise objectives to streamline efforts and focus on meaningful outcomes.

Camera placement posed another challenge. Determining optimal positions involved frequent adjustments through trial and error, as it was difficult to predict the best locations for comprehensive coverage and model performance (Fig. 17). It further emphasized the need for diverse data at this stage, rather than a focus on specific behaviors or areas of interest. As a result, our understanding of how to balance model capabilities with camera setup was significantly improved. During this phase, data generation relied on post-processing videos, which required considerable time and resources for data creation and management. To address this challenge, methods for generating pose data in real-time were developed, thus reducing the need for videos for sanity check purposes.

7.4 Initial research phase: balancing research goals with system demands

At this stage, the focus was on optimizing the system to address specific research goals while balancing cost and efficiency. In contrast to early phases, where large areas were monitored, we reduced coverage to specific regions of interest (ROIs). Optimization of camera placement and system development enabled the integration of cameras in the coverage of several ROIs, reducing hardware requirements while maintaining functionality. By refining the



Figure 17 Example of camera position to cover a region of interest. The images illustrate a complex environment with multiple challenges, including occlusion from barn infrastructure and neighboring cows.

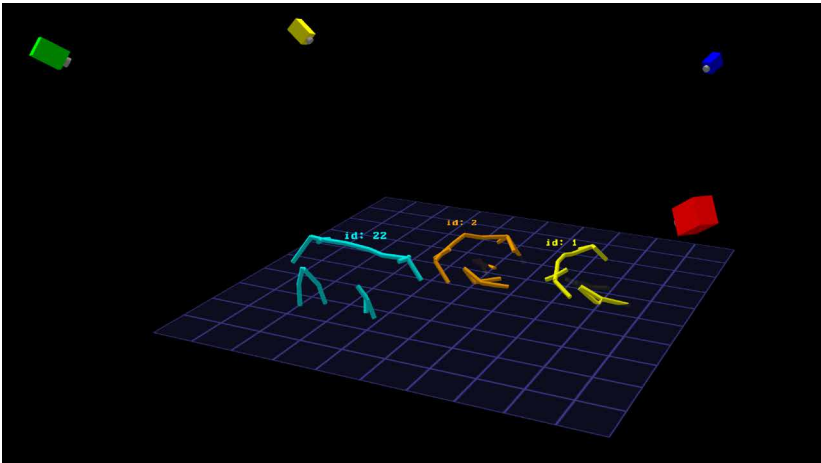


Figure 18 3D evaluation of cow posture using 24 keypoints, including the back line, head, and legs (December 2024).

setup, computational demands were reduced, and manual intervention was minimized, making the process more efficient and purpose-driven. The number of keypoints used in analysis was increased to 24 (Fig. 18), enhancing the system's ability to capture subtle movements and improving accuracy in behavior tracking. Through iterative development and ongoing annotation, the system's alignment and performance have steadily improved. Increasing correlations between automated outputs and manual observations suggest that the system is now sufficiently refined to support reliable and meaningful research.

8 Conclusion

Computer vision offers a robust, non-invasive approach for monitoring cattle health by extracting detailed information on posture, locomotion, body condition, and social interactions from continuously collected visual data. With high-resolution cameras and advanced image-processing algorithms, CV systems can detect subtle gait abnormalities and measure weight changes without the need for physical contact. Compared to wearable sensors such as accelerometers, CV eliminates the challenges of fitting and maintaining equipment on each animal, which can be labor-intensive and stressful if frequent recalibration or battery replacement is needed. It also avoids issues like sensor loss or malfunction, which might go unnoticed and compromise data integrity. In contrast to sound analysis, which is most effective for vocalization-based indicators (e.g. coughing or stress calls) but can be prone to ambient noise contamination, CV captures a broader range of health and welfare metrics (e.g. body shape, behavioral

interactions) through spatial and temporal analyses. However, implementing CV solutions in production environments requires reliable hardware, stable lighting, and potentially sophisticated machine learning models capable of handling occlusions (e.g. animals blocking one another) and variable backgrounds. Furthermore, the large data volumes generated by continuous video streams demand significant computational and storage resources, necessitating careful system design and ongoing technical maintenance.

As technology continues to reshape livestock production, research on CV algorithms for dairy cattle behavior monitoring offers a promising path toward improving livestock productivity, fostering sustainability, and protecting animal welfare. Such approaches have the potential to provide farmers, veterinarians, feed experts, slaughterhouse personnel, consumers, and other stakeholders with real-time insights into their herds, allowing for more efficient use of resources, reducing environmental impacts, and ensuring that animal health and welfare remain central concerns.

However, this is no simple task. Developing and refining advanced imaging techniques, robust machine learning models, and reliable analytical frameworks involves grappling with a wide range of challenges. The complexity lies in capturing subtle behavioral cues under diverse conditions and ensuring that any information derived is accurate, actionable, and ethically responsible. Achieving the highest levels of performance requires careful preparation, from extensive data collection to fine-tuning algorithms and integrating expert domain knowledge.

Yet technology alone will not suffice. Before these solutions progress from the research arena into commercial R&D, rigorous scientific validation through collaboration between engineers and animal experts is crucial. Demonstrating their reliability, generalizability, and measurable benefits under real, diverse farm conditions is a non-negotiable prerequisite. Only with this strong evidence can stakeholders confidently embrace these tools, thereby reinforcing trust and paving the way for meaningful improvements in both economic viability and animal welfare standards. Ultimately, it is the steadfast commitment to scientific rigor and transparency that will ensure these innovations deliver to their potential, ushering in a new era of sustainable and animal-centric livestock management.

9 Where to look for further information

- Guzhva, O., Ardö, H., Nilsson, M., Herlin, A., & Tufvesson, L. (2018). Now you see me: Convolutional neural network based tracker for dairy cows. *Frontiers in Robotics and AI*, 5, 107.
- Kroese, A., Högberg, N., Vicuna, E. D., Berthet, D., Fall, N., Alam, M., & Tamminen, L. M. (2025). Evaluating the automated measurement of

abnormal rising and lying down behaviours in dairy cows using 3D pose estimation. *Smart agricultural technology*, 101205.

- Högberg, N., Berthet, D., Alam, M., Nielsen, P. P., Tamminen, L. M., Fall, N., & Kroese, A. (2025). Exploring pose estimation as a tool for the assessment of brush use patterns in dairy cows. *Applied Animal Behaviour Science*, 106746.
- Feng, Z., Karaskova, M., & Mahmoud, M. (2023, September). Open-sheep-face: A comprehensive application for sheep face analysis and pain estimation. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 1–3). IEEE.

10 Future trends in research

The future of CV in dairy cattle behavior research is poised to transform the field significantly. Advancements in CV technology, driven by rapid developments in AI and hardware, are set to enable more precise, non-invasive, and continuous monitoring of livestock. This progress will likely shift the paradigm in animal behavior research from labor-intensive, manual observation methods to fully automated systems capable of providing real-time insights.

One key trend is the increasing adoption of more advanced DL models, such as transformer-based architectures, tailored to analyze complex animal behaviors in diverse environments. These models are being integrated with emerging imaging techniques like 3D and thermal imaging, offering richer datasets that can capture subtle movements, posture changes, and physiological indicators such as body temperature. This will allow researchers to identify patterns and anomalies in behavior with improved accuracy, improving our understanding of animal welfare and health dynamics.

Another promising direction is the use of markerless pose estimation and facial expression analysis to assess cattle comfort, stress, and pain. These techniques will enhance the granularity of behavioral studies, linking specific visual cues to physiological and emotional states. The application of these methods in real-world farming settings will likely expand, supported by improvements in hardware robustness and algorithm resilience against environmental variability, such as lighting changes or obstructions in barns.

In the broader field of animal behavior research, the development of CV technologies will facilitate the study of behavior across larger populations and over extended periods, which was previously impractical. This will yield deeper insights into group dynamics, social structures, and long-term behavioral trends. The ability to gather data at a scale without disturbing natural behavior patterns will significantly enhance the ecological validity of research findings.

However, these advancements also bring challenges. The complexity of CV models requires careful preparation, from collecting diverse and representative

datasets to designing interpretable systems that can explain predictions. Ethical considerations, such as data privacy and the implications of surveillance technologies, must be addressed. Furthermore, proper scientific validation under varied real-world conditions will be crucial before these tools transition to widespread commercial use. In addition to this, the need for standardized datasets is at its highest to ensure model interpretability and validation of technologies under diverse farm conditions. Future research should focus on developing robust, scalable, and ethically sound CV systems that can be seamlessly integrated into existing farm infrastructures. Collaborative efforts between researchers from different disciplines, industry stakeholders, and policymakers will be crucial in driving innovation and ensuring the successful adoption of these technologies in dairy farming.

In conclusion, the future of dairy cattle behavior monitoring lies in the synergistic application of CV with AI and biological process knowledge, paving the way for a more efficient, sustainable, and welfare-oriented dairy industry.

11 References

- Alvarez, J. R., et al. (2018). Body condition estimation on cows from depth images using Convolutional Neural Networks. *Computers and Electronics in Agriculture*, 155, 12–22.
- Barrier, A. C., et al. (2012). Parturition progress and behaviours in dairy cows with calving difficulty. *Applied Animal Behaviour Science*, 139(3–4), 209–217.
- Berckmans, D. (2017). General introduction to precision livestock farming. *Animal Frontiers*, 7(1), 6–11.
- Benaissa, S., et al. (2020). Calving and estrus detection in dairy cattle using a combination of indoor localization and accelerometer sensors. *Computers and Electronics in Agriculture*, 168, 105153.
- Borchers, M. R., et al. (2017). Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *Journal of Dairy Science*, 100(7), 5664–5674.
- Buhrmester, V., et al. (2021). Analysis of explainers of black box deep neural networks for computer vision: a survey. *Machine Learning and Knowledge Extraction*, 3(4), 966–989.
- Chang, A. Z., et al. (2022). *A multi-sensor approach to calving detection*. Information Processing in Agriculture.
- Cockburn, M. (2020). Application and prospective discussion of machine learning for the management of dairy farms. *Animals*, 10(9), 1690.
- Collins, B., et al. (2008). Towards scalable dataset construction: an active learning approach. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12–18, 2008, Proceedings, Part I 10* (pp. 86–98). Springer Berlin Heidelberg.
- Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

- Fadul, M., et al. (2017). Prediction of calving time in dairy cattle. *Animal Reproduction Science*, 187, 37–46.
- Feng, X., et al. (2019). Computer vision algorithms and hardware implementations: a survey. *Integration*, 69, 309–320.
- Fuentes, A., et al. (2020). Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information. *Computers and Electronics in Agriculture*, 177, 105627.
- Fuentes, A., et al. (2023). Multiview monitoring of individual cattle behavior based on action recognition in closed barns using deep learning. *Animals*, 13(12), 1-21
- Gao, G., et al. (2023). UD-YOLOv5s: recognition of cattle regurgitation behavior based on upper and lower jaw skeleton feature extraction. *Journal of Electronic Imaging*, 32(4), 043036–043036.
- García, R., et al. (2020). A systematic literature review on the use of machine learning in precision livestock farming. *Computers and Electronics in Agriculture*, 179, 105826.
- Gong, C., et al. (2022). Multicow pose estimation based on keypoint extraction. *PLoS One*, 17(6), e0269259.
- Guzhva, O. & Siegford, J. M. (2022). The unintended (and unconsidered) consequences of PLF: ethical and social considerations of PLF running the farm. In T. Bahanzi, et al. (Eds.), *Practical precision livestock farming* (pp. 383–396). Wageningen Academic.
- Guzhva, O., et al. (2016). Feasibility study for the implementation of an automatic system for the detection of social interactions in the waiting area of automatic milking stations by using a video surveillance system. *Computers and Electronics in Agriculture*, 127, 506–509.
- Guzhva, O., et al. (2018). Now you see me: Convolutional neural network based tracker for dairy cows. *Frontiers in Robotics and AI*, 5, 107.
- Kang, X., et al. (2021). Features extraction and detection of cow lameness movement based on thermal infrared videos. *Transactions of the Chinese Society of Agricultural Engineering*, 37(23), 169–178.
- Kroese, A., et al. (2024). 3D pose estimation to detect posture transition in free-stall housed dairy cows. *Journal of Dairy Science*, 107, 6878–6887. <https://doi.org/10.3168/jds.2023-24427>
- Kuncheva, L. I., et al. (2022, December). A benchmark database for animal re-identification and tracking. In *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)* (pp. 1–6). IEEE.
- Li, G., et al. (2021). Practices and applications of convolutional neural network-based computer vision systems in animal farming: a review. *Sensors*, 21(4), 1492.
- Liseune, A., et al. (2021). Leveraging sequential information from multivariate behavioral sensor data to predict the moment of calving in dairy cattle using deep learning. *Computers and Electronics in Agriculture*, 191, 106566.
- Lidfors, L. M., et al. (1994). Behaviour at calving and choice of calving place in cattle kept in different environments. *Applied Animal Behaviour Science*, 42(1), 11–28.
- Lodkaew, T., et al. (2023). CowXNet: an automated cow estrus detection system. *Expert Systems with Applications*, 211, 118550.
- Mar, C. C., et al. (2023). Cow detection and tracking system utilizing multi-feature tracking algorithm. *Scientific Reports*, 13(1), 17423.
- McDonagh, J., et al. (2021). Detecting dairy cow behavior using vision technology. *Agriculture*, 11(7), 675.

- Miller, G. A., et al. (2020). Using animal-mounted sensor technology and machine learning to predict time-to-calving in beef and dairy cows. *Animal*, 14(6), 1304–1312.
- Nir, O., et al. (2018). 3D computer-vision system for automatically estimating heifer height and body mass. *Biosystems Engineering*, 173, 4–10.
- Norton, T., et al. (2019). Precision livestock farming: building ‘digital representations’ to bring the animals closer to the farmer. *Animal*, 13(12), 3009–3017.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Rudin, C. & Radin, J. (2019). Why are we using black box models in AI when we don’t need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2), 1–9.
- Saar, M., et al. (2022). A machine vision system to predict individual cow feed intake of different feeds in a cowshed. *Animal*, 16(1), 100432.
- Stygar, A. H., et al. (2021). A systematic review on commercially available and validated sensor technologies for welfare assessment of dairy cattle. *Frontiers in Veterinary Science*, 8, 634338.
- Wang, J., et al. (2023). Open pose mask R-CNN network for individual cattle recognition. *IEEE Access*, vol. 11, (pp. 113752-113768), doi: 10.1109/ACCESS.2023.3321152.
- Wathes, C. M., et al. (2005). Is precision livestock farming an engineer’s daydream or nightmare, an animal’s friend or foe, and a farmer’s panacea or pitfall? *Precision Livestock Farming*, 5, 33–46.
- Wei, X. S., et al. (2016). Scalable algorithms for multi-instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(4), 975–987.
- Xu, Y., et al. (2022). Vitpose: simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35, 38571–38584.
- Zia, A., et al. (2023). Cvb: a video dataset of cattle visual behaviors. arXiv preprint arXiv:2305.16555.

