2025-STM5IPL- FINAL REPORT

Mapping of Dairy Cow Head Movement in Relation to Methane Sensors using Deep Learning

A Comparative Analysis of CNN and Transformer Architectures for Occluded Tracking

Author: Judy Thanh Uyen Nguyen
Date: February 2026
Host Organisation: DEECA /Agriculture Victoria Research
University/Course: La Trobe University / Master of Data Science

TABLE OF CONTENTS

Abstract

Livestock methane emissions are a notable contributor to climate change, yet scalable phenotyping remains a bottleneck for genetic selection programs limited by the high cost of respiration chambers. While "sniffer" sensors in milking parlours offer a high-throughput solution, their accuracy is potentially compromised by variable animal behaviour and head positioning relative to the sampling inlet. This study presents a computer vision framework to optimise these sensor readings. By correlating continuous head-tracking data with methane concentration spikes, we aim to identify the optimal behavioural window for capturing eructation events. We evaluate the performance of YOLOv8 (CNN) versus RT-DETR (Transformer) in detecting animal heads under severe occlusion. The resulting spatial-temporal data was mapped against the proposed methane sampling window to confirm biological alignment, establishing a validated protocol for distinguishing feeding events from non-feeding idling.

1. Introduction

Climate change is significantly driven by greenhouse gases (Gerber et al., 2010), with ruminants accounting for approximately 18% of all human-caused GHG emissions (Steinfeld et al., 2006). Enteric methane (CH4) is a primary component, with a Global Warming Potential (GWP) significantly higher than CO2. With global livestock consumption projected to double by 2050 (Stanford Woods Institute, 2025), there is an urgent need for mitigation strategies that align with production demands.

Current strategies to reduce methane emissions generally fall into two categories: management/nutrition and genetics. Nutritional interventions (e.g., 3-NOP, improved forage) deliver immediate reductions but require continuous input (Kinley et al., 2020). Conversely, genetic selection offers a permanent, cumulative solution requiring no ongoing intervention. However, to effectively select for these traits, we require accurate, high-throughput methods to phenotype methane emissions in large commercial populations.

The ‘gold standard’ for measuring methane is the Respiration Chamber (RC), which encloses the animal in a sealed environment to capture total gas flux (Global Research Alliance on Agricultural Greenhouse Gases [GRA], n.d.). While highly accurate, RCs are expensive, labour-intensive, and limit animal mobility, making them unsuitable for large-scale phenotyping. The Greenfeed system offers a more flexible alternative, using an automated head-chamber to measure gas flux during voluntary visits. However, capital costs remain a barrier to herd-wide deployment (GRA, n.d.).

Other scalable alternatives exist but come with trade-offs. Laser Methane Detectors (LMD) allow for remote measurement but often provide only a 'snapshot' of emissions rather than a full profile (Jenkins et al., 2024). Similarly, Mid-Infrared Spectroscopy (MIR) can predict methane output based on milk fatty acid profiles (Gerber et al., 2010). While scalable, MIR remains an indirect proxy rather than a direct measurement of gas. Consequently "Sniffer" systems, sensors installed directly in milking parlour feed bins, are increasingly favoured due to their low cost and ability to capture direct emission data during routine milking (Pfau, 2024).

Despite their scalability, sniffers face a critical limitation: they measure gas concentration (ppm) rather than absolute emission volume, making them highly sensitive to the animal's head position relative to the sensor. If a cow lifts her head or moves laterally during an eructation (burp), the sensor may miss the peak entirely or record a diluted value. This project aims to address this validation gap using computer vision. The primary objective is to pinpoint the time window where the cow's head is consistently submerged in the feed bin. Since methane concentration dilutes rapidly with distance, maximizing 'Head-Down' time directly correlates with the most accurate and undiluted sensor measurements.

2. Methodology

2.1 Data Acquisition and Preprocessing

The data collection setup is designed to capture direct methane emissions during the simultaneous milking and feeding cycle. The facility utilises automated feed bins integrated into the milking bail, with methane sensors mounted directly above each bin feed. The measurement strategy hypothesizes that sensor readings will be maximal when the cow’s head is lowered into the bin (concentrating the eructation plume).

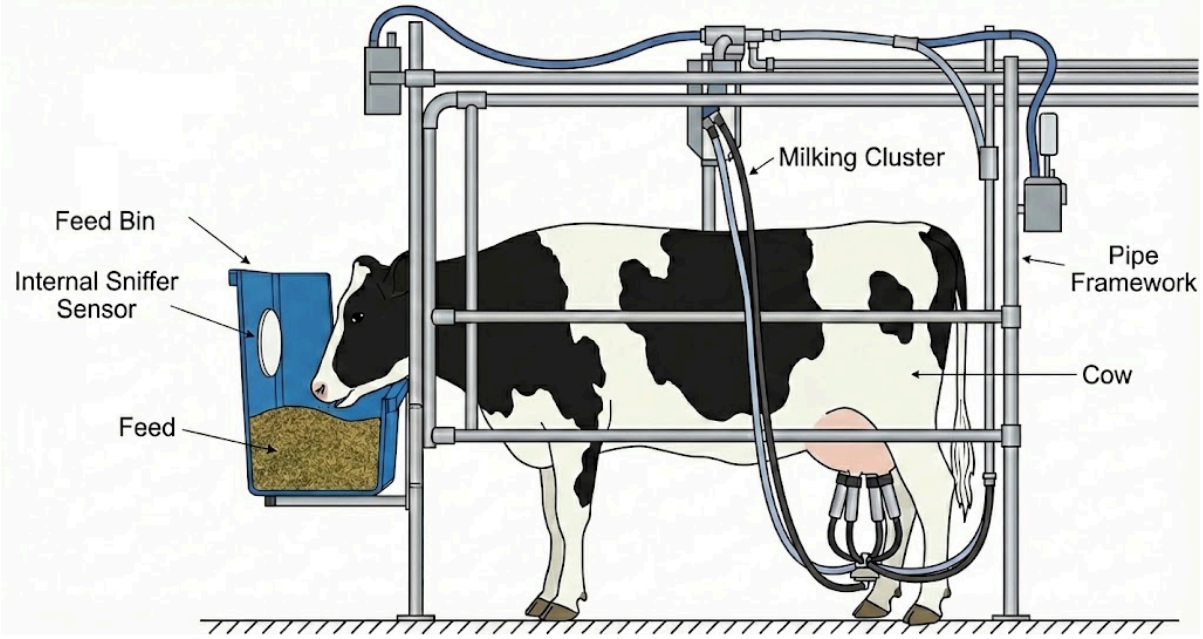

Figure 1: Profile of milking parlour sensor setup. Generated by Gemini (Google, 2026).

Raw video footage was captured using wide-angle GoPro cameras positioned at a high angle, located several metres in front of the milking bail. This vantage point establishes a clear visual distinction: heads are visible during idling (head-up) but significantly occluded by the bin during feeding (head-down).

To ensure model robustness, data was collected from four distinct sessions representing different bail setups, times of day (AM/PM), and lighting conditions (Figure 2). The source footage includes recordings from February 3rd, 4th (2024), and February 5th (2025), ensuring the model generalises across temporal and environmental variations.

Figure 2: Diversity of data collection setups. The dataset includes four distinct video sources captured across different dates and times (AM/PM), utilising GoPro Hero 5 and Hero 6 cameras. This variability in lighting, camera angle, and bail configuration prevents the model from overfitting to a single visual environment. (Source IDs: GH020088, GH020129, GP05905, GP020047).

Raw footage was segmented using LosslessCut to extract 2-minute clips capturing balanced behavioural variability ('head-down' vs. 'head-up') without re-encoding. Footage was subsequently standardised via Handbrake to optimise file size for web-based annotation.

2.2 Annotation Strategy

Initial testing failed due to domain shift; pastoral-trained models (Figure 3) did not generalise to the low-contrast, occluded parlour environment. Consequently, the models frequently misclassified the entire bail structure as a cow head even if the head is not visible, yielding false positives (Figure 3). To address this, a custom dataset was curated.

Figure 3: (Top) A representative sample from the public 'Cow Head' dataset, characterized by idealized conditions: high-contrast illumination, unobstructed views, and uniform backgrounds. (Bottom) Failure cases of pre-trained models in the target milking parlour environment. Left: A generic off-the-shelf detector fails to localise the subject due to severe occlusion. Right: The Segment Anything Model 3 (SAM 3) exhibits inconsistent performance, producing False Negatives (missed detections) and misclassifying the cow's body as the head. Collectively, these failures highlight the necessity of training custom models.

Due to the low-contrast environment and severe occlusion, CVAT's temporal annotation features were utilised. By observing animal movement over time, annotators could infer head positions in frames with near-zero visibility, a task near impossible using static image labelling tools. Furthermore, CVAT's Track Mode with linear interpolation was employed to streamline the process. Keyframes were annotated every 10–20 frames, utilising software interpolation for intermediate coordinates. This not only accelerated the workflow but ensured smoother, more consistent bounding box trajectories for the training data.

A critical deviation from standard annotation protocols was implemented to accommodate severe occlusion. While traditional object detection guidelines typically dictate labelling only

the visible portion of an object ('modal' annotation), this project employed an 'amodal' annotation strategy. When the cow’s head lowered into the feed bin, the bounding box was drawn to encompass the estimated full extent of the head, effectively labelling the occluding bin structure as the target during feeding events.

This approach was chosen to exploit the YOLO format’s coordinate system (x,y,w,h). By maintaining a bounding box that represents the head's true physical position rather than just its visible fragments, the spatial centre point (x, y) remains distinct and stable even during occlusion. However, this strategy introduces a risk of 'contextual bias,' where the model might overfit to the static features of the feed bin (e.g., metal bars) rather than the animal itself. To mitigate the risk of false positives in empty bails, a specific negative sampling strategy was employed (detailed below).

2.3 Data Synchronisation Scripting

A key challenge was the format incompatibility between the annotation tool and training pipeline. CVAT was utilised to label video data, but the export function generated only coordinate text files, omitting the corresponding image frames. A custom Python script using the OpenCV library **(Script A.1)** was developed to resolve this issue.

This script programmatically scanned the exported label directory to identify annotated frame indices. It then processed the raw source video, extracting the corresponding high-resolution frames and pairing them with their respective label files. This automated approach ensured synchronisation between the visual data and ground-truth labels, reconstructing a valid YOLOv8-compliant dataset.

This data structure was specifically prioritised because the selected model implementations (for both YOLOv8 and RT-DETR) natively support this format. This unification eliminates the need for redundant conversion to COCO JSON format, facilitating a direct and efficient comparative analysis between the CNN-based and Transformer-based architectures.

2.4 Environment and Hardware Setup

To facilitate local training and ensure reproducibility, the training environment was configured using Python 3.12 within a specialised Conda environment. The Ultralytics library was employed to deploy the YOLOv8 and RT-DETR architectures. The training process was accelerated using NVIDIA CUDA integration.

2.5 Data Augmentation Pipeline

To enhance model robustness within the low-light, high-occlusion environment, a two-stage augmentation strategy was implemented, combining offline pre-processing with real-time training augmentations.

2.5.1 Offline Augmentation (Roboflow)

A targeted augmentation pipeline was first applied via Roboflow to physically expand the training set. Since the dataset contained only one class ("Cow Head"), class imbalance was not an issue. The following transformations were generated to triple the effective dataset size (15,810 training images):

- **Random Horizontal Flip & Rotation (-12° to +12°):** Simulates the variable head angles of feeding cattle, as well as accounting for minor variances in camera mounting orientation (roll) across different bails.
- **Brightness (-15% to +15%) & Exposure (-10% to +10%):** Critical for this environment, these augmentations train the model to recognize features regardless of the variable lighting conditions in the barn.
- **Blur (up to 1.5px) & Noise (up to 0.1%):** Simulates the sensor grain and motion blur common in low-light surveillance footage.

Additionally, ~180 "background-only" images (empty stalls) were incorporated as negative samples **(Script A.1b)**. These null examples are critical for reducing false positives, forcing the model to distinguish between the static bail infrastructure and the animal itself.

Once augmentation by Roboflow was complete, the final dataset of 18,062 images was split into Training, Validation, and Test sets following an approximate 88:8:4 split ratio (15,810, 1,494, and 758 images, respectively)

2.5.2 Online Augmentation (Ultralytics Pipeline)

Complementing the offline strategy, the Ultralytics training pipeline applied online (dynamic) augmentations directly on the GPU during the training loop **(Script A.3)**. Unlike the static Roboflow dataset, these transformations are applied stochastically for each batch, ensuring the model generalises to diverse variations of the input data. A few notable augmentations are:

- **Mosaic Augmentation:** Applied (p=1.0) to encourage scale-invariance by synthesizing images from four random crops. This prevents overfitting to spatial context (e.g., assuming central positioning) and forces the model to recognize feature representations regardless of pixel size, essential for distinguishing peripheral cattle in adjacent stalls. Mosaic was disabled for the final 10 epochs.
- **HSV (Hue, Saturation, Value) Modulation:** High variance in Saturation (0.7) and Value (0.4) was applied to mimic drastic lighting shifts (day vs. night) and changing exposure levels. A conservative Hue shift (0.015) was retained to accommodate minor white-balance fluctuations (e.g., colour temperature changes from natural to artificial light) while ensuring the semantic colour integrity of the cattle (e.g., preventing a black cow from appearing artificially green or purple).
- **Occlusion Simulation:** Random Erasing (p=0.4) was utilised to simulate partial occlusion, training the model to infer head position even when obscured by feed bins

and parlour architecture.

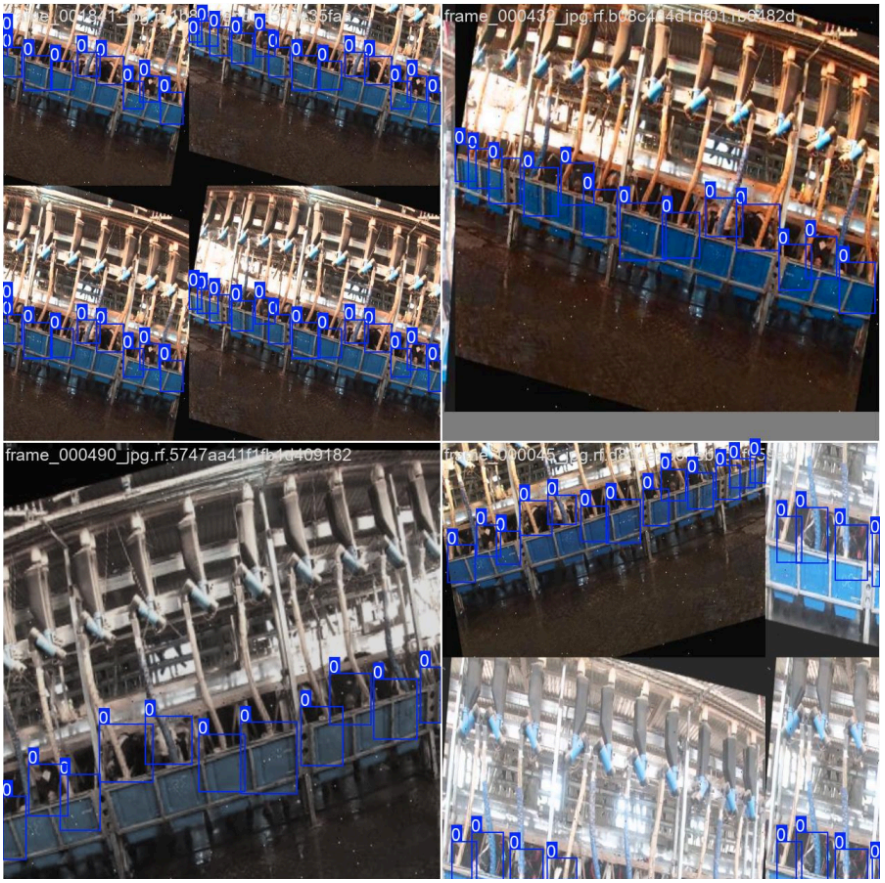

Figure 4: Visualisation of the dual-stage augmentation pipeline. The sample demonstrates the combined effects of independent Geometric Rotation (visible in all images) followed by Mosaic Augmentation (stitching four images, In quadrants 1 & 4). Pixel-level photometric modulations are also visible, specifically Hue/Saturation shifts (Quadrants 1 & 2) and Luminance/Contrast adjustments (Quadrants 3 & 4).

2.6 Model Architectures and Training Configuration

To evaluate the trade-off between inference speed and occlusion robustness, two distinct architectures were implemented: YOLOv8n and RT-DETR. Both architectures were trained using the Ultralytics framework, initialised with COCO-pretrained weights to leverage transfer learning.

1. **YOLOv8n (You Only Look Once: Nano):** A single-stage Convolutional Neural Network (CNN) optimised for real-time inference. It extracts hierarchical feature maps in a single forward pass, relying on inductive bias to prioritise local features (shapes and edges) for rapid object localisation.

2. **RT-DETR (Real-Time Detection Transformer):** A hybrid architecture utilising a Transformer encoder to capture global context via self-attention mechanisms. This model was selected to test if analysing the entire image simultaneously could improve detection under occlusion, effectively inferring the head's location based on the position of the body.

Optimisation Protocol:

Both models were optimised using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, a final learning rate fraction of 0.01, and Nesterov momentum set to 0.937. A linear warmup phase was applied for the first 3 epochs to stabilise gradients. To refine feature localisation in the final training stages, Mosaic augmentation was systematically disabled for the last 10 epochs.

Architecture-Specific Hyperparameters:

- YOLOv8n: Training was executed for 100 epochs with a batch size of 16. To mitigate overfitting on the static background features of the milking parlour, a Dropout rate of 0.2 was explicitly applied. An early stopping patience of 15 epochs was enforced.
- RT-DETR: The model was scheduled for 100 epochs (executed in 2 stages due to interruptions). Due to the high memory cost of the Transformer's self-attention mechanism, the batch size was reduced to 4. The Early Stopping patience was set to 20 epochs to account for the slower convergence characteristic of Transformer architectures.

Input Specifications:

All input media was resized to a standard resolution of 640x640 pixels.

2.7 Post-Processing and Temporal Analysis

A custom three-stage pipeline was developed to translate raw frame-by-frame detections into actionable behavioural data **(Script A.5)**. The spatial coordinates of each detection were serialised into a structured dataset (CSV format) containing the Frame Number, Track ID, and the centroid coordinates (x, y) of every detected head.

2.7.1 Multi-Object Tracking (ByteTrack)

The first stage utilised the track method with the ByteTrack algorithm. The implementation of a tracker was essential for short-term temporal stability. By utilising Kalman filtering, the system smoothed the bounding box trajectories, reducing the high-frequency 'jitter' inherent in frame-by-frame detection. Furthermore, this temporal association prevented sensor flicker (where a detection momentarily vanishes), ensuring that a cow contributed consistent spatial data to the herd index even if its specific Track ID fragmented over long durations.

2.7.2 Scene-Specific Spatial Calibration

To normalise data across varying camera setups, a dynamic thresholding strategy was employed. A manual Feeding threshold was calibrated for each video, defining the specific pixel coordinate of the feed rail (typically 550px–900px) **(Script A.2)**. This boundary provided a consistent biological reference point to distinguish "head-down" (feeding) states from "head-up" (non-feeding) states. A frequency distribution analysis of the vertical centroid positions confirmed the validity of this binary classification, revealing a distinct bimodal distribution between the two behavioural states **(Script A.12)**

2.7.3 Derivation of the Herd Feeding Index

Finally, a Herd Feeding Index was computed by calculating the arithmetic mean of the vertical ‘y’ centroids of all detected subjects per frame. To mitigate high-frequency sensor noise, a 30-frame rolling average filter was applied to the time series, effectively smoothing the data to isolate underlying behavioural trends **(Script A.6, A.7)**.

3. Results and Discussion

3.1 Model Performance Comparison

To assess model performance, we utilised Mean Average Precision (mAP), the standard metric for object detection. Specifically, we report mAP@50-95, which averages precision across ten Intersection over Union (IoU) thresholds ranging from 0.50 to 0.95.

While standard mAP@50 was calculated, it resulted in a ceiling effect, with both models achieving near-perfect scores. This suggests that despite frequent visual obstructions (e.g., bars, gates), the models could robustly identify the presence of the animal, aided by the single-class nature of the task which eliminates classification ambiguity.

However, the mAP@50-95 metric proved to be the necessary discriminator. The obstructions made precise boundary delineation difficult; by enforcing stricter IoU thresholds, mAP@50-95 revealed that YOLOv8 was significantly more effective at inferring the correct object extent under partial occlusion compared to RT-DETR.

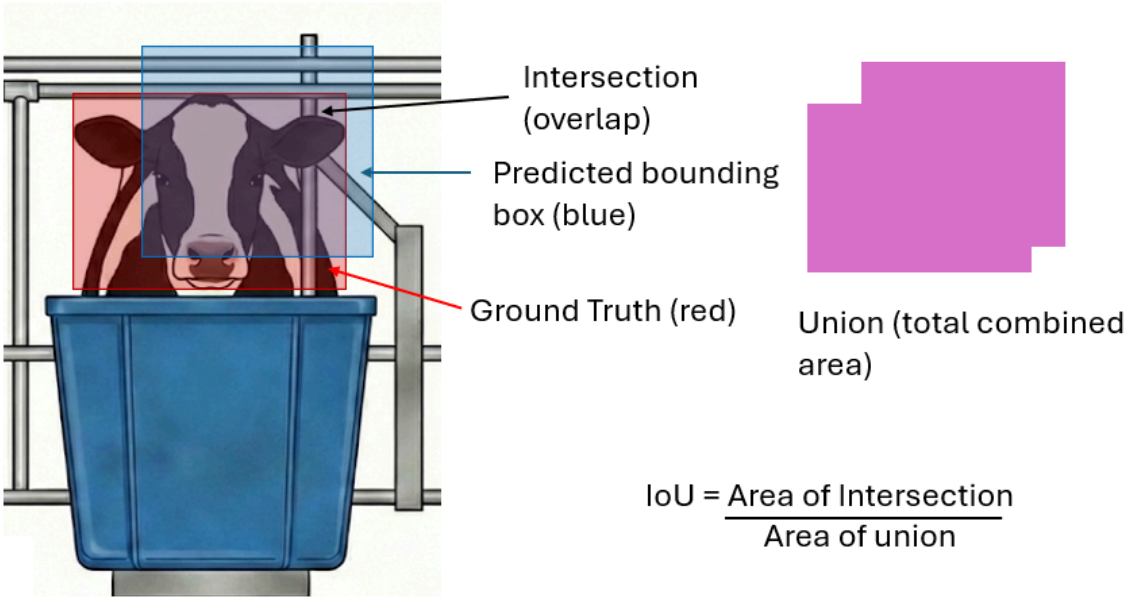

Figure 5: Visual representation of the Intersection over Union (IoU) metric applied to livestock detection. The Ground Truth (Red) represents the precise manual annotation of the cow's head. The Predicted Bounding Box (Blue) represents the model's output. The Union (Purple) illustrates the total combined area of both boxes. The IoU score is calculated by dividing the area of overlap (Intersection) by the total area (Union), penalising predictions that are misaligned or oversized relative to the ground truth.

Contrary to the initial hypothesis that the Vision Transformer would outperform due to global context awareness, the YOLOv8 model demonstrated superior performance.

- **YOLOv8:** Achieved a peak mAP@50-95 of 97.0% at epoch 100 with an inference speed of 190ms (CPU). The model demonstrated exceptional stability, converging steadily over 100 epochs with no signs of overfitting.
- **RT-DETR:** Peaked at an mAP@50-95 of 87.8% at epoch 56 and suffered from significant latency, averaging ~3071ms per image. Crucially, the model exhibited severe training instability, with accuracy dropping as low as 44% at Epoch 15, indicating a failure to robustly converge.

3.2 Quantitative Error Analysis (Confusion Matrix)

While mAP metrics provide a high-level summary of performance, they do not fully capture the operational risk of "hallucinations" in a precision livestock setting. To quantify this, we analysed the Confusion Matrix for both models on the validation set (Figure 6).

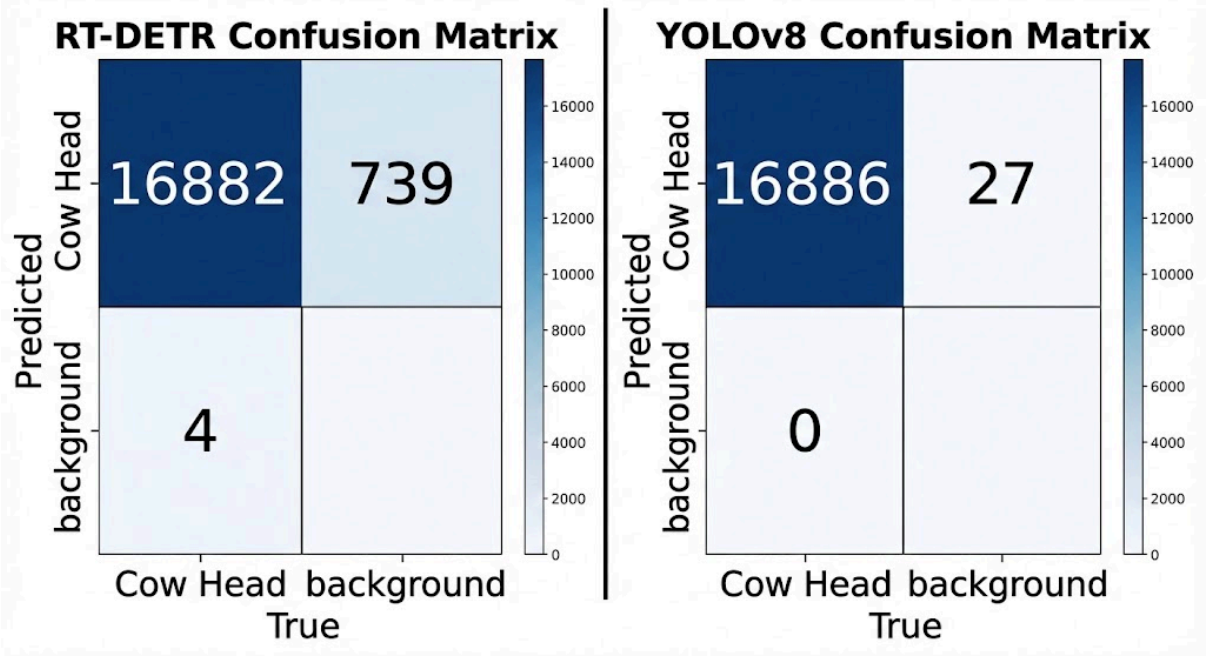

Figure 6: Confusion Matrix for RT-DETR and YOLOv8

The RT-DETR model exhibited a systematic deficiency, generating 739 False Positives (identifying background infrastructure as cattle). With a precision of 95.8%, the model essentially "hallucinated" a cow in approximately 4% of its positive detections. In a production environment, this tendency would lead to thousands of invalid data points being logged as "feeding events," effectively corrupting the methane dataset with background noise.

In contrast, YOLOv8 generated only 27 False Positives across the entire validation set, demonstrating a 96% reduction in false alarms compared to RT-DETR.

- **Precision:** 99.8% (Virtually zero "ghost" detections).
- **Recall:** 100% (No missed feeding events).

This result validates the effectiveness of the negative sampling strategy, confirming that the CNN architecture successfully learned to ignore background movement. This metric is critical for the project's logic script; the near-zero False Positive rate ensures that sensor data is logged only when an animal is physically present and feeding.

3.3 Qualitative Failure Analysis

A qualitative analysis of failure cases **(Script A.4, Script A.11)** reveals a critical difference in how the models interpret the milking parlour environment. As shown in Figure 7, RT-DETR 'hallucinated' objects in the upper infrastructure, misclassifying a ventilation unit as a 'Cow Head' while failing to detect the actual animal below. In contrast, YOLOv8 successfully located the animal despite extreme occlusion where the head was not strictly visible.

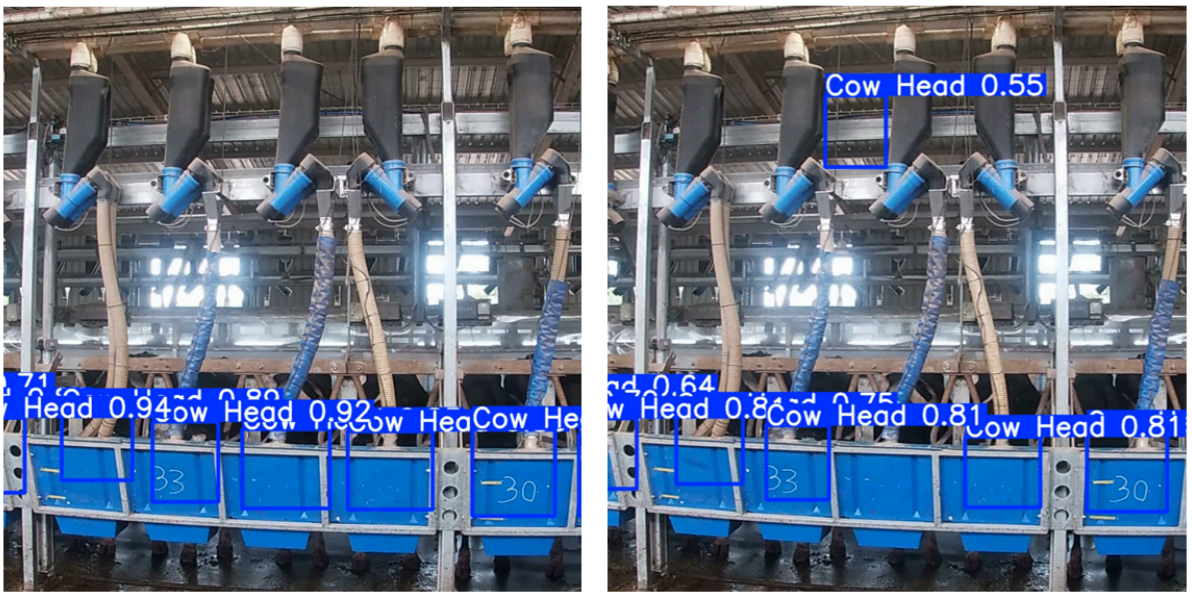

Figure 7: Qualitative comparison of detection performance under severe occlusion. Source: Video GH040088, timestamp 04:32 (Left) YOLOv8: The model demonstrates robust contextual inference, correctly placing the bounding box on the feed bin despite the animal's head being visually obscured. (Right) RT-DETR: The model suffers from a severe misclassification, confusing the upper ventilation system for a cow head (False Positive) while missing the actual animal in bail 32 (False Negative).

This disparity suggests that the CNN-based YOLOv8 effectively leveraged 'inductive biases', specifically locality (neighbouring pixels are related). By implicitly learning that 'cow' features are always surrounded by specific textures (feed bins, feeding tubes), YOLOv8 correctly rejected illogical detections in the upper gantry where the background (ceiling, pipes) did not match this learned context. In contrast, the Transformer—lacking this neighbourhood constraint—often mistook the global similarity of metal pipes for feed rails.

The observed performance gap aligns with findings by Dosovitskiy et al. (2021), who noted that Vision Transformers lack these inherent spatial priors and typically require massive datasets to learn spatial relationships from scratch. With a dataset of only ~18,000 images, the RT-DETR model struggled to establish stable feature representations, whereas YOLOv8 efficiently extracted robust local features (shapes and edges) from the limited data. Given this superior stability, combined with inference speeds 16x faster than RT-DETR on standard hardware, YOLOv8 remains the only viable candidate for the project's real-time deployment phase.

3.4 Limitations and Operational Constraints

3.4.1 Inference Challenges at the Periphery

Inference (at standard Confidence: 0.25, IoU: 0.70) across wide-field recordings revealed consistent performance degradation at the edges of the frame. While central detections were robust, the extreme outer bails exhibited three distinct error patterns driven by the wide viewing angle:

1. **False Negatives (Perspective Compression):** High-angle occlusion and reduced pixel density at the boundaries rendered key features indistinguishable from the background. Notably, RT-DETR outperformed YOLOv8 here, leveraging global attention to infer presence from herd patterns despite compression.
2. **Duplicate Detections (Morphological Ambiguity):** The oblique perspective often triggered double-counting, where the dorsal curve of a cow's shoulder visually mimicked the curvature of a head ("feature mimicry").
3. **Contextual Hallucination:** The model occasionally flagged empty stanchions as cattle. This is an artifact of the amodal labelling strategy, where the model detected the context (the bail structure) rather than the object (the cow).

3.4.2 Dataset Constraints: The Necessity of Random Splitting

The use of a random stratified split introduces a degree of temporal autocorrelation between training and test sets. However, this strategy was prioritised over a session-based split due to the subjective nature of amodal annotation. Since the "Ground Truth" for an occluded head is an educated human estimate rather than a visible fact, a distinct test session introduces the risk of inter-session variability—where slight inconsistencies in annotation style could artificially penalise the model. Random splitting ensured the model was evaluated against a consistent logic, measuring its ability to learn the specific spatial constraints of the project. To mitigate data leakage risks, manual verification of inference video streams was conducted to confirm that high mAP scores translated into stable, persistent tracking behaviour.

3.4.3 The Operational Advantage of Spatial Overfitting

While lack of generalization to new environments is typically considered a failure in machine learning, for this specific industrial application, spatial overfitting is an operational advantage. The GoPro cameras are installed in fixed locations within a stationary milking parlour. By "overfitting" to this specific environment, the model effectively memorizes the valid "feeding zones," utilising the static background features (pipes, gates) as a robust spatial filter. This reduces false positives in non-relevant areas (walkways, rafters), prioritising site-specific reliability over portability.

3.4.4 Variable Frame Rate Artifacts

Post-hoc metadata analysis revealed frame rate inconsistencies (~30fps vs. ~60fps) between sessions. Since the current aggregation pipeline processes data frame-by-frame rather than normalising to a time-base, high-framerate sessions are effectively 'time-dilated' in the final index. This results in temporal smearing, where behavioural events in these sessions appear delayed relative to the herd average. Future iterations of the pipeline should implement explicit time-resampling (e.g. normalising all inputs to 1 Hz) prior to aggregation to eliminate this variance.

3.5 Operational Feasibility: The Herd Feeding Index

To determine if a consistent pattern exists in collective head movement, the YOLOv8 pipeline was applied to continuous milking sessions to generate a custom metric termed the

'Herd Feeding Index' defined as the arithmetic mean of the vertical (y) centroid positions for all detected animals, calculated over each frame. This aggregation smooths individual behavioural variance, revealing the underlying spatiotemporal rhythm of the herd's feeding activity.

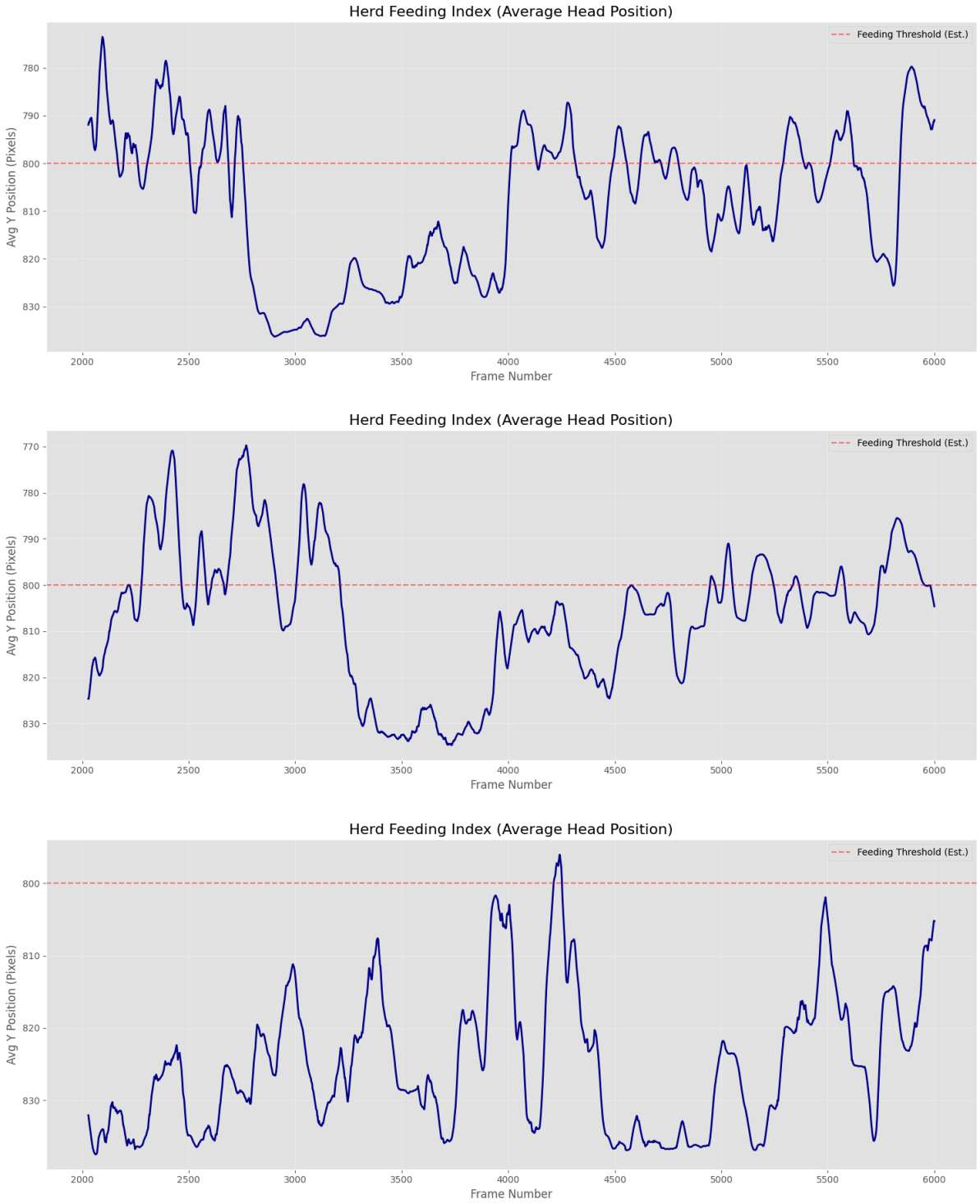

Figure 8: Temporal profile of the Herd Feeding Index across three distinct sessions. The y-axis represents the inverted average centroid height (higher values = head down/feeding). The "Feeding Horizon" (dashed line) indicates the estimated

threshold where the muzzle enters the bin zone. (Video acquisition rate: 30 fps; the displayed range represents approximately 130 seconds of activity).

3.5.1 Behavioural Sensitivity and Social Facilitation

The resulting time-series data (Figure 8) reveals that the system successfully digitised the collective behaviour of the herd. Rather than random noise, the data exhibits synchronised "waves" of activity where the herd collectively transitions above and below the feeding threshold. This strongly suggests the system is capturing Allelomimetic Behaviour (social facilitation), where cattle are driven to feed simultaneously by herd instinct.

Notably, this synchronisation effect appeared consistently stronger in smaller herd configurations (~10-12 animals) compared to wide-field recordings (~18-19 animals). This variance is hypothesized to be a compound effect of both biological and technical factors:

1. **Signal Dilution:** In wider bails, the "Average Centroid" calculation is more susceptible to outliers. If peripheral detection fails (as noted in Section 3.4.1) or if the wider angle distorts the vertical position of outer cows, these errors introduce noise that dampens the collective "wave" signal.
2. **Social Fragmentation:** Biologically, larger groups may exhibit less cohesive synchronisation compared to smaller, tighter clusters, resulting in more sporadic feeding intervals.

The ability of the computer vision model to quantify this group dynamic along with its variations confirms its operational feasibility. It provides a reliable, high-resolution Boolean trigger (Feeding/Not Feeding) that can be timestamp-synced with methane sensor logs to filter out background noise.

3.5.2 Temporal Variability and Protocol Validation

Aggregating eight recorded sessions (Figure 9) **(Script A.8)** revealed signal dampening due to variable feeding onset times. The high variance between individual sessions (grey lines) confirms that narrow, static sampling windows would frequently capture non-feeding noise.

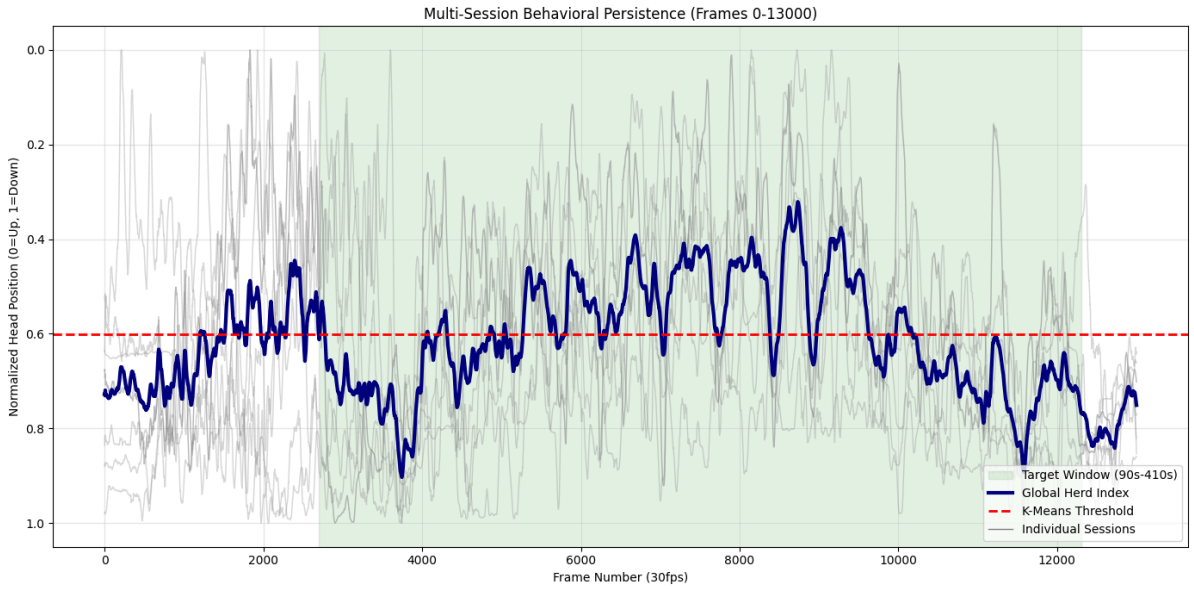

Figure 9: Temporal profile of the Global Herd Feeding Index (averaged over 8 sessions) covering approximately 6 minutes and 40 seconds, overlaid with the Almasi et al. (2025) sampling window (Green Zone; 90–410s). (Gray Lines): Individual sessions exhibit high temporal variability ("jitter"). (Blue Line): The global average reveals a pattern: an initial feeding bout, followed by a mid-session lull in activity (Frames 6,000–10,000), and a secondary feeding resurgence (Frame 11,500+).

The Computer Vision analysis confirms that the 90–410s window is essential not for static alignment, but for temporal sufficiency; it anchors on the initial 'Head-Down' onset and spans the subsequent mid-session oscillations, ensuring the capture of multiple eructation events despite natural behavioural volatility.

4. Future Work

While the framework validates CNN efficacy, observed anomalies suggest potential for downstream refinements:

- **Active Learning Loop:** Using the model to generate pseudo-labels for unannotated footage **(Script A.9, A.10)**—followed by human verification—would rapidly correct perspective-based peripheral errors.
- **Hyperparameter Optimisation:** Automated parameter sweeping could resolve the convergence instability observed in Transformer architectures (RT-DETR) and further refine YOLOv8 precision.
- **Sensor Fusion:** Integrating the real-time "Feeding State" boolean into the sensor could autonomously pause sampling during idling.

Conclusion

This study successfully validated the use of CNN-based computer vision to operationalise sniffer sensors in milking parlours. While RT-DETR struggled with the limited dataset and severe occlusion, YOLOv8 demonstrated the necessary robustness and speed for real-time deployment. The identification of peripheral limitations and the implementation of specific "Feeding Horizon" calibration protocols provide a clear roadmap for integrating this technology into commercial genetic selection programs.

References

Almasi, F., Williams, S. R. O., Vander Jagt, C. J., Marett, L. C., Jacobs, J. L., & Pryce, J. E. (2025). Repeatability and heritability of dairy cow methane concentration using sniffer sensors. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics*, 26, 146–149.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Gerber, P., Key, N., Portet, F., & Steinfeld, H. (2010). Policy options in addressing livestock's contribution to climate change. *Animal*, 4(3), 393–406. https://doi.org/10.1017/S1751731110000133

Global Research Alliance on Agricultural Greenhouse Gases. (n.d.). *Livestock emissions measurement*. Retrieved December 1, 2025, from https://globalresearchalliance.org/research/livestock/collaborative-activities/livestock-emissions-measurement/

Google. (2026). *Gemini* [Large language model]. https://gemini.google.com

Jenkins, B., Herold, L., de Mendonça, M., Loughnan, H., Willcocks, J., David, T., Ginns, B., Rock, L., Wilshire, J., & Avis, K. (2024). Breeding for reduced methane emissions in livestock. *ClimateXChange*. https://doi.org/10.7488/era/5569

Kinley, R. D., Martinez-Fernandez, G., Matthews, M. K., de Nys, R., Magnusson, M., & Tomkins, N. W. (2020). Mitigating the carbon footprint and improving productivity of ruminant livestock agriculture using a red seaweed. *Journal of Cleaner Production*, 259, 120836. https://doi.org/10.1016/j.jclepro.2020.120836

Lv, W., et al. (2024). DETRs Beat YOLOs on Real-time Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pfau, A. (2024, September 3). *How can we measure methane emissions from commercial farms?* University of Wisconsin–Madison Division of Extension. https://dairy.extension.wisc.edu/articles/how-can-we-measure-methane-emissions-from-commercial-farms/

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Stanford Woods Institute for the Environment. (2025, November 21). *Meat's environmental impact*. Stanford University. https://woods.stanford.edu/news/meats-environmental-impact

Steinfeld, H., Gerber, P., Wassenaar, T., Castel, V., Rosales, M., & de Haan, C. (2006). *Livestock's long shadow: environmental issues and options*. Food and Agriculture Organization of the United Nations.

Van Soest, P. J. (1994). *Nutritional ecology of the ruminant* (2nd ed.). Cornell University Press.

Appendix A: Software and Code Registry

All source code developed for this project is available at:
[github.com/Jyuudie/methane-validation-cv]

ID	***Script Name***	***Pipeline Stage***	***Functionality Description***
A.1	*sync_frames.py*	***Data Prep***	***Data Synchronisation:*** *Iterates through CVAT annotation exports, matches them to raw video frames, and extracts high-resolution image-text pairs to create a valid YOLO/COCO dataset.*
A.1b	*generate_negative_labels.py*	***Data Prep***	*Negative Sample Generation: Automates the creation of empty annotation text files for background images, enabling the negative mining strategy.*
A.2	*find_threshold.py*	***Calibration***	***Spatial Calibration:*** *A GUI utility that loads a reference frame and allows the user to manually identify the pixel coordinate (y) of the feed rail. This establishes the "Feeding Horizon" for specific camera angles.*

A.3	*train_models.py*	***Training***	***Model Training Loop:*** *Configures and initiates the training for YOLOv8 and RT-DETR. Includes the custom augmentation pipeline (Mosaic, MixUp) and hyperparameter definitions (Dropout=0.2, Epochs=100).*
A.4	*visualize_inference.py*	***Visualisation***	***Qualitative Verification:*** *Runs the trained model on sample video footage with `stream=True` to prevent memory overflows. Generates annotated video outputs with bounding box overlays to visually verify tracking stability*
A.5	extract_tracks.py	***Inference***	***Data Generation:*** *Applies ByteTrack to raw video footage, extracting temporal coordinate data (x,y) for every identified animal and serialising it into CSV format.*
A.6	aggregate_sessions.py	***Aggregation***	***Data Merging:*** *Iterates through a directory of individual session CSVs. Computes the spatial mean (average y-position of the herd) per frame and merges all sessions into a single `Master_Herd_Feeding_Index.csv` matrix.*

A.7	*plot_herd_index.py*	***Analysis***	***Single-Session Analysis:*** *Visualises the feeding behaviour of a single milking session. Applies a rolling average (Window=30) to smooth sensor noise and plot the head position against the calibrated threshold.*
A.8	*plot_validation_summary.py*	***Validation***	***Biphasic Validation:*** *The final analytical script. Applies Min-Max Scaling to normalise pixel coordinates across different camera resolutions. Uses K-Means Clustering (k=2) to determine thresholds and generates the "Spaghetti Plot" to prove biological alignment.*
A.9	*extract_inference_frames.py*	***Data Ingestion***	***Frame Extraction:*** *A preprocessing utility that converts raw video footage into a sequence of static images with configurable sampling rates (e.g., 5fps), preparing the data for the auto-labelling pipeline.*
A.10	*generate_cvat_preannotations.py*	***Active Learning***	***Automated Labelling:*** *Uses the trained model to predict bounding boxes on the frames generated by* ***Script A.9****, creating pre-annotated text files for rapid correction in CVAT.*

A.11	*run_rtdetr_inference.py*	***Inference***	***Transformer Inference:*** *The dedicated execution script for the* ***RT-DETR*** *architecture. It loads the Vision Transformer weights and runs inference with configuration parameters specific to the attention-mechanism architecture.*
A.12	*plot_bimodal_distribution.py*	***Analysis***	***Statistical Validation:*** *Generates a frequency histogram of all head positions. The resulting Bimodal Distribution (two distinct peaks) statistically confirms the existence of two discrete behavioural states (Feeding vs. Idling), validating the use of clustering algorithms.*

Appendix B: Data Acquisition Log

The following video sessions were processed to generate the Master Herd Feeding Index (Figure 9).

Session ID	***Video Filename***	***Date***	***Conditions***	***Duration (Processed)***
S01	***GH050088***	***Feb 3, 2024***	***PM Milking / Hero 3 White***	***10m 48s***

S02	***GH080088***	***Feb 3, 2024***	***PM Milking / Hero 3 White***	***7m 21s***
S03	***GH010129***	***Feb 4, 2024***	***AM Milking / Hero 6 Middle***	***10m 48s***
S04	***GH060129***	***Feb 4, 2024***	***AM Milking / Hero 6 Middle***	***10m 48s***
S05	***GH030093***	***Feb 4, 2025***	***PM Milking / White 3***	***10m 39s***
S06	***GH040093***	***Feb 4, 2025***	***PM Milking / White 3***	***7m 53s***
S07	***GH040088***	***Feb 3, 2024***	***PM Milking / Hero 3 White***	***8m 14s***
S08	***GH050093***	***Feb 4, 2025***	***PM Milking / White 3***	***10m 45s***