

Politecnico di Torino

Politecnico di Torino

Master's Degree course in Computer Engineering

A.Y. 2025/2026

CLOVER

Continuous Livestock Observation, Visualization, and Evaluation Resource

Supervisors:

Prof. Stefano Di Carlo
Prof. Alessandro Savino
Prof. Roberta Bardini
Dr. Lorenzo Martini
Alessia Leclercq

Candidate:

Salar Farahmand

Abstract

Daily milk yield is the central indicator of dairy productivity and animal welfare, and unexpected drops often signal disease, heat stress, or nutritional deficits. Reliable short-term forecasts of milk yield can therefore enable timely intervention and prevent economic loss. Precision livestock farming increasingly relies on low-cost behavioral and physiological sensors monitoring activity, rumination, core body temperature, and thermal indices. however, converting these signals into accurate next-day yield forecasts remains an open challenge, particularly for small-to-medium herds that lack the historical yield data automated milking systems normally provide.

When historical yield is available, tree-based and linear models normally achieve R^2 above 0.8–0.9, since yield is strongly autocorrelated day to day. Using behavioral and physiological sensors alone, however, results are considerably weaker, with reported R^2 values ranging from 0.30 to 0.77 depending on herd size and task formulation. The most directly comparable prior work, PLF-Mamba, addresses the same constrained scenario (10 cows, sensor-only, no yield history, next-day forecast) using a specialized reinforcement-learning-gated state-space architecture, reporting R^2=0.656. Such deep sequence models, however, introduce substantial computational complexity, opacity, and deployment costs.

This thesis analyzes whether a simpler, interpretable classical machine learning approach can match or exceed this performance. From four base sensor channels (core body temperature, temperature-humidity index, lying time, and feeding time), 57- and 77-feature sets were engineered via lags, rolling statistics, and interactions. Four models (Ridge, Random Forest, XGBoost, LightGBM) plus a mean-predictor baseline were evaluated across two cross-validation strategies (walk-forward, group k-fold) and three target formulations (next-day absolute yield, percentage change, absolute change), combined with hyperparameter tuning and feature selection for 60 configurations, each assessed under a chronological holdout and a Leave-One-Cow-Out (LOCO) test of generalization to an unseen cow, for 120 experiments in total.

Under the chronological holdout, the champion configuration (XGBoost, 77 engineered features) achieved R^2=0.6905 and MAE=2.48 kg for next-day yield, outperforming PLF-Mamba (R^2=0.656) and all baselines reported in that study (Transformer 0.640, LSTM 0.633, MLP 0.548) under an identical constraint, while using a simpler, more interpretable model class; the percentage-change and absolute-delta formulations showed negligible-to-negative gains over baseline. Under LOCO, every model failed to beat baseline on absolute yield, while a Random Forest on the relative-delta target modestly exceeded it (R^2=0.03), indicating the sensor-only

features forecast a monitored herd well but do not yet generalize to an unseen cow's absolute production level.

These results show that domain-informed feature engineering combined with classical tree ensembles constitutes a competitive, cheaper, and more interpretable alternative to specialized deep sequence models for sensor-only milk yield prediction in an already-monitored small herd. The study is limited to a single 10-cow herd, and its feature set intentionally excludes historical yield. Future work could pursue multi-herd validation, closing the individual-generalization gap, and markerless computer vision as an alternative to wearable sensors.

Acknowledgements

To my mother and my father, who gave me birth—the very opportunity that led me to love computer science.

Table of Contents

List of Figures VII

1 Introduction 1
- 1.1 Rise of Precision Livestock Farming . . . 1
- 1.2 Challenge of Environmental Heat Stress . . . 2
- 1.3 Biological Lag in Milk Production . . . 3
- 1.4 Small-Herd Constraint and Sensor-Only Forecasting . . . 4
- 1.5 Research Objectives and Methodology . . . 5

2 Background 7
- 2.1 Sensor Modalities and Wearable Technologies . . . 7
  - 2.1.1 Hardware Telemetry, Sensor Physics, and Power Constraints 8
  - 2.1.2 Anatomical Placements and Kinetic Modalities . . . 8
  - 2.1.3 Signal Processing and Algorithmic Classification . . . 9
  - 2.1.4 Commercial Benchmarks and Comparative Analysis . . . 10
- 2.2 Markerless Computer Vision for Behavioral Monitoring . . . 10
  - 2.2.1 Pose Estimation and Keypoint Tracking . . . 10
  - 2.2.2 Translating Kinematics into Actionable Behaviors . . . 11
- 2.3 Available Datasets for Precision Livestock Farming . . . 11
- 2.4 Time-Series Modeling Architectures . . . 13
  - 2.4.1 The Recurrent Era: RNNs and LSTMs . . . 13
  - 2.4.2 The Attention Mechanism and Transformers . . . 14
  - 2.4.3 State-Space Models and Selective Mamba . . . 14
- 2.5 Explainable AI (XAI) . . . 15
  - 2.5.1 Local Interpretable Model-agnostic Explanations (LIME) . . 15
  - 2.5.2 SHapley Additive exPlanations (SHAP) . . . 15
- 2.6 Data Imputation and IoT Signal Preprocessing . . . 16
  - 2.6.1 The Taxonomy of Missing Data in PLF . . . 16
  - 2.6.2 Imputation Methodologies . . . 16
  - 2.6.3 Signal Smoothing and Outlier Attenuation . . . 17

3 State of the Art 18
3.1 Environmental Stress, Biomarkers, and Health Detection 18
3.1.1 Deep Learning for Subclinical Lameness Detection 19
3.1.2 Systemic Biomarkers and Classical ML in Clinical Disease . 19
3.1.3 Predictive Modeling in Reproductive States 19
3.1.4 Synthesis: Biomarkers as Precursors to Environmental Heat Stress 20
3.2 Milk Yield Prediction: From Historical Data to Sensor-Only Constraints 20
3.2.1 Autoregressive Models Utilizing Historical Yield Data 21
3.2.2 The Sensor-Only Exogenous Forecasting Constraint 21
3.2.3 The PLF-Mamba Empirical Benchmark 22
3.3 Limitations of Current Approaches and This Thesis's Response . . . 23

4 Methodology 25
4.1 The MmCows Dataset and Bronze Layer Ingestion 25
4.2 The Silver Layer: Deterministic Extraction 27
4.2.1 Rule-Based Extraction from Hardware Geometry 27
4.2.2 Feature Orthogonality and "Idle Standing" 28
4.3 Target Formulation and Biological Normalization 30
4.4 Construction of the Temporal Feature Matrices 31
4.4.1 Non-Linear Interactions and Cyclical Seasonality 32
4.4.2 The Dimensionality Consideration 33
4.5 Algorithmic Architectures 33
4.5.1 Linear Baseline: Ridge Regression (L_2 Regularization) . . . 34
4.5.2 Bootstrap Aggregating: Random Forest Regressor 34
4.5.3 Extreme Gradient Boosting (XGBoost) 35
4.5.4 Light Gradient Boosting Machine (LightGBM) 35
4.6 MLOps: Systematic Hyperparameter Tuning 36
4.7 Cross-Validation Methodologies 36
4.8 Two Evaluation Strategies: Temporal and Individual Generalization 37
4.9 Evaluation Metrics and Interpretability (XAI) 38
4.10 Overview of the Experimental Design 39

5 Results and Discussion 40
5.1 Experimental Architecture and the Naive Baseline 40
5.2 Target Variable Analysis: The Signal-to-Noise Problem 41
5.2.1 The Success of Absolute Yield Forecasting 41
5.2.2 The Collapse of Delta Targets 41
5.2.3 Understanding the Forecasting Chasm 42
5.3 Champion Model Performance and SOTA Benchmarking 42

5.3.1 Matched Comparison: Small Herds without Yield History . . 43
5.3.2 The Autoregressive Performance Ceiling 43
5.4 Feature Space and Cross-Validation Dynamics 44
5.4.1 Feature Dimensionality (`mega_57` vs. `mega_77`) 44
5.4.2 Algorithmic Stability and Validation Splits 45
5.5 Explainable AI and Biological Interpretability 46
5.5.1 Global Feature Importance and Temporal Engineering . . . 46
5.5.2 SHAP Value Analysis: Directional Impact on Yield 46
5.6 Leave-One-Cow-Out Results: Generalization to an Unseen Cow . . 49
5.6.1 A Sharp Reversal for Absolute Yield 49
5.6.2 Why Absolute Yield Fails to Generalize 50
5.6.3 A Partial Success for Relative Change 50
5.6.4 Implications for the Sensor-Only Constraint 51

6 Secondary Contribution: Environmental Sensor Data Quality Assessment 52
6.1 Experimental Setup and Data Characteristics 52
6.2 Hardware Validation: Are the Two Boxes Measuring the Same Thing? 53
6.3 Barn Density and Air Quality: A Comparative Analysis 54
6.4 Contextualizing Readings Against Welfare Thresholds 55
6.5 Relevance to the Broader Thesis 56

7 Future Work and Alternative Methodologies 57
7.1 Alternative Methodology: Markerless Visual Extraction 57
7.1.1 Stage 1: Object Detection and Localization 58
7.1.2 Stage 2: Micro-Behavioral Classification 58
7.1.3 Stage 3: Visual Cow Identification 58
7.1.4 Multi-View Fusion (RGBm) and Performance Trade-offs . . 59
7.2 Advanced Behavioral Classification Models 60
7.3 Markerless Computer Vision Integration 61

8 Conclusion 62
8.1 Empirical Validation and the SOTA Benchmark 62
8.2 Generalization to an Unseen Cow: A Necessary Boundary 63
8.3 The Value of Deterministic Feature Engineering and XAI 64
8.4 Agricultural Democratization and Forward Trajectory 64

Bibliography 66

List of Figures

2.1 The hierarchical pipeline of markerless computer vision in PLF, from raw video to the same feeding and lying behaviors targeted by the wearable-sensor pipeline in this thesis. 12

4.1 The Composite AI Pipeline: A fully automated, leakage-proof architecture mapping raw hardware telemetry into an interpretable early warning forecasting system. 26

4.2 Empirical Geofencing: Spatial distribution of UWB telemetry points strictly associated with ground-truth feeding behaviors. This density map was used to derive the exact bounding box coordinates for the Silver Layer extraction. 29

4.3 Micro-Level Behavioral Signals: The square-wave comparison of UWB feeding coordinates against Ankle-IMU lying state. The absence of overlap shows an apparent lack of contemporaneous activity between the two modalities at the second-by-second level. . 29

4.4 Macro-Level Daily Trade-off: The wide scatter distribution shows that feeding and lying hours are not perfectly inverse over a 24-hour period. 30

5.1 Relative feature importance (gain) for the champion XGBoost model, illustrating the dominance of 3-day rolling core-body-temperature volatility and lying/feeding duration extremes over raw daily telemetry. 47

5.2 SHAP summary beeswarm plot for the champion XGBoost model. Red indicates high feature values; blue indicates low values. Spatial position on the x-axis denotes the directional impact on the predicted next-day milk yield. 48

6.1 Four representative outcomes of the hardware audit. Temperature shows near-perfect agreement between boxes (R = 0.99); CO2 shows a consistent additive offset of roughly 126 ppm; hydrogen sulfide on Box 1 is flat except for isolated spikes, consistent with a non-functioning sensor; methane shows no consistent relationship between the two boxes at all. 54
6.2 THI readings from both barns over the monitored period, plotted against the optimal range (shaded) and the 72-point heat stress threshold (dashed line). Both boxes track closely, consistent with the near-perfect THI agreement found in the hardware audit. 56

7.1 The markerless computer vision pipeline recreated from the MmCows methodology. The top branch illustrates the single-view (RGBs) architecture utilizing YOLOv8 and bifurcated EfficientNet-B0 classifiers. The bottom branch illustrates the multi-view spatial fusion (RGBm). 59

Chapter 1

Introduction

The convergence of precision agriculture and advanced data analytics has established a new paradigm in livestock management, fundamentally shifting how farmers monitor and optimize animal welfare [1]. Within the dairy sector, environmental heat stress is a pervasive and economically severe challenge [2]. As climate change drives global temperatures upward, dairy cattle—animals biologically prone to heat sensitivity due to the high metabolic output required for digestion and lactation [3] encounter increasingly hostile environmental conditions.

Daily milk yield acts as the primary indicator of both production and welfare in modern dairy management. Sudden fluctuations or declines in yield function as early warning signals for underlying complications, including metabolic disease, environmental heat stress, or acute nutritional deficits. Developing a reliable, short-term forecasting system is therefore essential. Such a system allows herd managers to take proactive measures, minimizing financial losses and protecting animal health before irreversible damage occurs.

1.1 Rise of Precision Livestock Farming

In recent years, Precision Livestock Farming (PLF) has transitioned from a theoretical framework into a practical reality [4]. Traditional herd management depended heavily on manual observation to identify distressed or sick animals. This approach is subjective, labor-intensive, and inherently unscalable. PLF replaces this reactive model with a system focused on proactive, individualized forecasting [1].

Driven by the expansion of the Internet of Things (IoT), affordable wearable sensors, and reliable wireless communication protocols, continuous monitoring of individual animals within large herds is now feasible. Modern PLF architectures rely on a multimodal sensor network that runs continuously inside harsh barn environments, generating high-frequency streams of physical activity, rumination,

core body temperature, and ambient thermal data [5]. The specific hardware behind these sensors, and how their raw signals are turned into usable behavioral data, is described in detail in Chapter 2.

Acquiring this telemetry is only the initial phase, however. The primary computational challenge in PLF involves translating massive, noisy, high-frequency sensor data into accurate, deployable forecasts of future milk yield. Continuous IoT streaming introduces datasets prone to signal drift, missing values, and the need to synchronize disparate sampling frequencies, among other practical obstacles.

Additionally, the utility of this data is currently constrained by farm size and capital infrastructure. Large industrial operations generally utilize automated milking systems (AMS) and leverage historical yield data to train predictive models. Small and medium-scale farms often lack this equipment. Since milk production exhibits strong autoregressive properties, models with access to current yield data can compute short-term future yields with high accuracy.

For smaller, traditional, or pasture-based farms unable to finance individual daily milking sensors, historical yield data remains unavailable. These operations therefore rely on advanced data science techniques to derive actionable insights solely from behavioral and physiological sensor inputs. Addressing this algorithmic gap is essential for democratizing PLF technology and represents the central computational focus of this research.

1.2 Challenge of Environmental Heat Stress

Dairy cattle function optimally within a defined thermal comfort zone, biologically identified as the thermoneutral zone, which generally ranges from 5°C to 25°C depending on breed and lactation stage. When environmental conditions—standardly quantified by the Temperature-Humidity Index (THI)—exceed this threshold, the animal experiences heat stress [6].

In precision agriculture and predictive modeling, environmental severity is not evaluated by ambient temperature alone. THI serves as a bioclimatic metric measuring the combined oppressive effect of dry-bulb temperature and relative humidity on the animal's capacity to dissipate heat [6]. A standard formulation of THI utilized in dairy science is defined as:

$$\text{THI} = (1.8 \times T_{db} + 32) - [(0.55 - 0.0055 \times RH) \times (1.8 \times T_{db} - 26)] \tag{1.1}$$

where T_{db} is the dry-bulb temperature in Celsius and RH is the relative humidity percentage. Once this threshold is crossed, the animal's metabolic equilibrium is disrupted through several physiological pathways: blood flow is redirected away from the udder to aid heat dissipation, directly impeding milk synthesis [7], while core body temperature rises and triggers broader cellular stress.

These physiological shifts are accompanied by clear behavioral adaptations. Feed intake decreases substantially—a response termed heat-induced anorexia—as the animal attempts to reduce the internal metabolic heat generated by rumen fermentation [8], pushing the cow into a Negative Energy Balance (NEB). Resting behavior also degrades: standing time increases sharply as the animal maximizes body surface area exposed to cooling airflow [9], a pattern known as "idle standing" that deprives the animal of rest and accelerates exhaustion. It is precisely these two measurable signals—core body temperature and the balance between feeding and lying time—that this thesis uses as the biological bridge between environmental conditions and milk yield.

1.3 Biological Lag in Milk Production

It is important to note that the manifestation of these stress responses in milk production is not instantaneous. The metabolic pathways controlling continuous lactation are highly buffered, resulting in a documented biological lag between the initial environmental stressor and the subsequent drop in yield [10, 11].

This latency occurs because milk synthesis requires a continuous supply of glucose, amino acids, and fatty acids derived from digestion. When heat stress causes a cow to curtail feed intake, the rumen still processes previously ingested material. The biological mechanics of digestion, microbiome fermentation into Volatile Fatty Acids (VFAs), and absorption through the rumen wall require 24 to 48 hours before nutrients circulate to the mammary gland [8]. Consequently, a severe heatwave on Tuesday (marked by immediate CBT spikes and feeding reduction) may not result in a measurable yield deficit until Thursday. During this 48-hour window, the mammary gland draws on circulating nutrients and mobilized body fat, temporarily obscuring the stress event from daily production metrics.

This latency introduces a complex time-series forecasting problem. The relationship between environmental hostility at time t and milk yield at time $t + n$ is non-linear and temporally shifted. While immediate weather conditions and behavioral distress are observable, their exact quantitative impact on future milk production is masked by the delayed physiological response. Recording a high THI value today does not automatically translate into an actionable prediction of tomorrow's yield without advanced feature engineering and lagged temporal modeling [12].

Reactive management strategies—intervening only after a yield drop is recorded in the parlor—are inefficient and lead to unrecoverable economic losses. Once mammary secretory cells biologically downregulate production due to nutrient starvation and heat stress, restoring yield to the pre-stress baseline can take weeks [8, 7]. Therefore, deploying machine learning algorithms that map this biological

lag to generate an actionable 24- to 48-hour early warning window is crucial. This approach transitions herd management to a preventative model, enabling targeted cooling and nutritional interventions before financial losses materialize.

1.4 Small-Herd Constraint and Sensor-Only Forecasting

When historical milk yield is provided as a feature, existing predictive models easily achieve high accuracy. This is driven by the strong day-to-day autoregressive nature (autocorrelation) of continuous milk production [12]. In time-series forecasting, milk yield at time t naturally embeds the cumulative biological effects of the cow's genetics, parity, days in milk, and baseline health. Providing this yield as an input variable mathematically simplifies the prediction of yield at time $t + 1$. Machine learning algorithms place heavy weight on this historical anchor, largely bypassing the subtle environmental or behavioral fluctuations occurring concurrently [12]. Leveraging this data, tree-based and linear models routinely achieve R^2 scores between 0.80 and 0.90 for short-horizon forecasts [12].

This reliance on historical yield, however, restricts real-world applicability. The fundamental constraint addressed in this thesis is forecasting next-day milk yield using behavioral and physiological sensors *alone*. This constraint directly supports the research objective of agricultural scalability. While industrial mega-dairies utilize Automated Milking Systems (AMS) to log precise per-milking yields, the capital expenditure required is prohibitive for small-to-medium and pasture-based operations [13]. Wearable IoT sensors (e.g., UWB tags and IMU collars), in contrast, are highly affordable and scalable. Developing an effective sensor-only methodology is therefore critical to democratizing predictive AI for smaller agricultural operations.

Under this strictly restricted sensor-only regime, the predictive task becomes highly non-linear, and current literature reflects more modest outcomes. Lacking the historical yield anchor, models struggle to map noisy behavioral shifts directly to delayed production outputs. The most comparable prior work, the PLF-Mamba framework [14], operates under identical constraints—a 10-cow herd, sensor-only inputs, no yield history, and a strict next-day horizon—but relies on a computationally heavy, reinforcement-learning-gated deep sequence architecture to do so. Chapter 3 examines this and related approaches in full detail.

This thesis questions the necessity of such opaque, computationally heavy deep learning models. Deep sequence architectures demand substantial overhead (e.g., GPU acceleration) and function as "black boxes," offering no biological explanation for predicted yield drops. The central research question investigates whether classical, interpretable machine learning algorithms (such as XGBoost and LightGBM)

can match or exceed SOTA deep sequence performance through rigorous feature engineering. By mapping deterministic behaviors and evaluating them through explainable AI frameworks, this methodology aims to provide predictive accuracy and biological transparency without requiring industrial milking infrastructure.

1.5 Research Objectives and Methodology

The core objective of this research is to demonstrate that accessible, highly interpretable tree-based machine learning ensembles can match or outperform computationally intensive deep sequence architectures (such as PLF-Mamba [14]) when supplied with robust, engineered features. In precision agriculture, accuracy must be coupled with biological interpretability; managers need transparent explanations for predicted yield drops to implement the correct interventions.

The methodology utilizes the MmCows dairy herd dataset [15], a synchronized 10-cow repository. The study isolates four primary multimodal sensor channels: Core Body Temperature (CBT), Temperature-Humidity Index (THI), lying time, and feeding time. Rather than relying on deep learning algorithms to automatically parse latent temporal patterns, this research champions an explicit, deterministic feature engineering pipeline. Applying strict hardware physics, the pipeline constructs comprehensive 57- and 77-feature spaces (`mega_57` and `mega_77`). These feature sets actively model the biological lag via temporal lags, rolling statistical windows (e.g., 3-day THI moving averages), and biological interaction terms to capture intra-day volatility.

To ensure statistical validity, an extensive experimental matrix of 60 configurations was systematically evaluated. This infrastructure benchmarked the core hypothesis across the following constraints:

- **Four Machine Learning Algorithms:** XGBoost (the primary gradient boosting model), LightGBM, Random Forest (a bagging baseline), and Ridge Regression (a regularized linear baseline).
- **Two Cross-Validation Strategies:** Group K-Fold (isolating specific cows to prevent identity-based data leakage) and Walk-Forward Chronological Holdout (training on past data to predict unseen future data, simulating real-world deployment).
- **Automated Hyperparameter Tuning:** Implementing Randomized Search algorithms bounded strictly within cross-validation folds to prevent optimization leakage.

Beyond this internal validation choice, the evaluation design distinguishes between two complementary questions: how well the model forecasts tomorrow's

yield for a cow it is already monitoring, and how well the same modeling approach generalizes to a cow it has never observed at all. The first question is addressed through a chronological holdout, where the dataset is split by date so that each cow's earlier days form the training set and her most recent days form an unseen test set, directly reflecting the deployment scenario of an early-warning system running on a herd already under observation. The second question is addressed through a Leave-One-Cow-Out (LOCO) evaluation, where each of the ten cows is held out in turn, the model is trained on the remaining nine, and tested exclusively on the excluded individual, measuring whether the learned relationship between behavior, environment, and yield transfers to an animal the model has never encountered. Both evaluation strategies are applied across the full 60-configuration grid described above, bringing the total number of experiments to 120. The two are never merged into a single ranking, since a chronological result and a Leave-One-Cow-Out result answer different questions; they are reported and discussed separately throughout this thesis.

Additionally, this study evaluated three distinct target formulations for the forecasting horizon: next-day absolute yield ($t + 1$), absolute change (raw kilogram difference), and relative percentage shock. The evaluation of a percentage-change target introduces a systematic metric largely absent from surveyed literature. Predicting a relative percentage shock provides a standardized measure for environmental stress, circumventing the natural variance in absolute yield caused by genetics and days in milk.

Although delta-style targets were extensively engineered and tested, the empirical models struggled to outperform naive mean-predictor baselines under the strict sensor-only constraints. The specific algorithmic results of these configurations, the SHAP interpretability analysis of the successful absolute-yield models, and the theoretical constraints preventing the models from capturing the volatile percentage-shock signal are analyzed comprehensively in the discussion and results chapters.

Chapter 2

Background

This chapter introduces the core technologies underpinning this research: the sensor hardware used to monitor dairy cattle, the markerless computer vision techniques that represent an emerging alternative to wearables, the deep learning architectures commonly used for time-series modeling, the explainable AI methods used to interpret model behavior, and the signal processing techniques required to handle noisy, incomplete IoT data. Rather than reviewing how these tools have been applied to the specific research problem of this thesis, the aim here is to explain what each technology is and how it functions, providing the reader with the technical vocabulary necessary to follow the state-of-the-art discussion and methodology that follow.

2.1 Sensor Modalities and Wearable Technologies

Table 2.1 summarizes the primary sensor modalities deployed in modern Precision Livestock Farming (PLF) architectures.

The initial wave of PLF integration was fundamentally dominated by the rapid development and deployment of commercial wearable sensors [4]. Because the manual observation of cattle by ethologists is unscalable for massive commercial herds, wearable sensors form the foundational hardware layer of modern dairy data engineering. By continuously capturing the physical movements of the animal, these devices successfully map raw kinetic energy into distinct, actionable biological behaviors.

Table 2.1: Primary Sensor Modalities in Precision Livestock Farming

Modality	Hardware Type	Placement	Biological Targets
Kinematic	3-Axis IMU	Neck/Leg	Activity, rumination, lying
Spatial	UWB Tracker	Ear/Collar	Geospatial mapping, feeding
Physiological	Ingestible Bolus	Reticulum	Core Body Temp (CBT), pH
Environmental	Weather Station	Barn	THI, microclimate

2.1.1 Hardware Telemetry, Sensor Physics, and Power Constraints

At the core of almost all commercial PLF wearables are Micro-Electromechanical Systems (MEMS), specifically Inertial Measurement Units (IMUs). A standard commercial dairy tag typically relies on a 3-axis accelerometer, which measures the proper acceleration (gravitational and dynamic kinetic forces) across the x, y, and z axes. More advanced research-grade tags employ 6-axis IMUs that incorporate gyroscopes to measure angular velocity, providing a deeper understanding of rotational postures.

To accurately capture high-frequency micro-behaviors, such as the rhythmic jaw movements of rumination, these accelerometers typically sample data at frequencies ranging between 10 Hz and 50 Hz. However, this high sampling rate introduces severe power and bandwidth constraints. Because it is unfeasible to frequently replace batteries on thousands of active cattle, commercial sensors aggressively employ data compression and duty-cycling algorithms at the edge (on the device itself). Rather than transmitting continuous raw data, the tags aggregate kinetic forces into proprietary epochs (e.g., 15-minute or 2-hour windows) and transmit these compressed summaries to local barn receivers via low-power radio-frequency (RF) protocols, Bluetooth Low Energy (BLE), or Ultra-Wideband (UWB) telemetry.

2.1.2 Anatomical Placements and Kinetic Modalities

Leading commercial systems utilize distinct anatomical placements, as the physical mounting point strictly dictates which behavioral modalities can be accurately captured. The industry is primarily bifurcated into neck, leg, and ear placements:

- **Neck and Collar Tags:** Systems such as the SCR Heatime (AI24) and

Afimilk AFICollar are mounted tightly to the cow's neck. This positioning provides an extremely high signal-to-noise ratio for head-centric movements. By tracking the cyclical raising and lowering of the head, combined with high-frequency vibrations from the jaw, these collars excel at detecting active grazing, feeding durations, and rumination cycles [16].

- **Leg and Ankle Tags:** Conversely, systems designed to measure absolute posture and locomotion, such as IceRobotics CowAlert, utilize leg-mounted IMUs. By tracking the horizontal versus vertical orientation of the sensor relative to the earth's gravitational pull, leg sensors extract highly accurate lying and standing bout durations [17]. This data is critical for animal welfare, as deviations in lying times and step counts are the primary leading indicators of lameness.

- **Ear Tags:** Emerging systems like the CowManager SensOor are affixed to the animal's ear like a traditional identification tag. The ear acts as an amplification pendulum for head movements, allowing the IMU to accurately capture the subtle twitches and rhythmic motions associated with rumination, while simultaneously utilizing an infrared thermistor to estimate superficial tissue temperature [18].

2.1.3 Signal Processing and Algorithmic Classification

To translate the raw x, y, and z accelerometer streams into discrete behaviors, commercial systems employ highly customized signal processing pipelines. First, the raw time-series data is segmented into rolling windows. Within each window, time-domain features—such as the Root Mean Square (RMS), variance, zero-crossing rate, and peak-to-peak amplitude—are extracted. For rumination detection, frequency-domain features are highly valuable; by applying Fast Fourier Transforms (FFT), systems can isolate the 2.0 to 3.0 Hz rhythmic frequencies that biologically correspond to the cow's chewing cycles.

Modern analytical frameworks have largely transitioned from heuristic thresholding to supervised machine learning paradigms. For example, the Afimilk AFICollar utilizes Linear Discriminant Analysis (LDA) to classify feeding times with a balanced precision of 0.88 [16]. Leg-mounted sensors are also being evaluated with advanced time-series classifiers; recent models leverage transformations like ROCKET (Random Convolutional Kernel Transform) to map erratic step-count volatility, achieving over 85% accuracy in predicting the need for corrective hoof trimming [19].

2.1.4 Commercial Benchmarks and Comparative Analysis

Table 2.2 provides an expanded summary of the prevailing commercial technologies, highlighting their primary modalities, physical placements, and the underlying algorithmic approaches utilized to map hardware physics to biological states.

Table 2.2: Expanded Overview of Commercial Wearable Technologies in Dairy PLF

System Name	Placement	Primary Modalities	Learning Type / Models
SCR Heatime / AI24	Collar/Neck	General activity, rumination, estrus, distress	Unsupervised / Logistic Regression
Afimilk AFICollar	Collar/Neck	Feeding times, rumination (jaw movements)	Supervised / Linear Discriminant Analysis
IceRobotics CowAlert	Leg/Ankle	Lying/standing duration, step count, lameness	Supervised / Thresholds & ROCKET
CowManager SensOor	Ear Tag	Rumination, eating, superficial temperature	Supervised / Decision Trees

Taken together, these systems illustrate the practical hardware and algorithmic vocabulary that recurs throughout the rest of this thesis. How well these commercial approaches actually serve the specific forecasting problem tackled in this research—and where they fall short—is examined critically in Chapter 3.

2.2 Markerless Computer Vision for Behavioral Monitoring

An alternative to physical wearable sensors is markerless computer vision. Vision-based pose estimation circumvents the hardware vulnerabilities of IoT tags, such as battery depletion, sensor drift, and physical damage caused by the barn environment. Using standard overhead cameras, computer vision architectures can theoretically measure lying and feeding durations non-invasively.

2.2.1 Pose Estimation and Keypoint Tracking

The first step in markerless behavioral analysis is multi-animal pose estimation. Open-source deep learning frameworks now allow researchers to track animal kinematics by plotting skeletal keypoints (e.g., nose, spine, hooves) across sequential video frames.

DeepLabCut (DLC) [20] advanced this field by showing that architectures pre-trained on large human datasets (like ImageNet) can be fine-tuned for animal poses via transfer learning. DLC applies ResNet and MobileNet convolutional backbones to generate probability maps for specific body parts, maintaining spatial accuracy with limited training data. However, in dense, occluded commercial barns, standard DLC frequently struggles with identity assignment between animals.

To address this, **SLEAP** (Social LEAP Estimates Animal Poses) [21] introduced custom UNet architectures optimized for multi-animal tracking. SLEAP handles both top-down and bottom-up inference techniques, enabling it to resolve identity swaps and occlusions in crowded pens while maintaining inference speeds above 800 frames per second on consumer GPUs [21].

2.2.2 Translating Kinematics into Actionable Behaviors

Skeletal keypoints are merely raw spatial coordinates; they must be translated into discrete behaviors like active rumination or idle standing. Tools such as **SimBA** (Simple Behavioral Analysis) [22] process this pose data to compute hundreds of spatiotemporal features, including angular velocities, Euclidean distances between keypoints, and bounding-box convex hulls.

SimBA then applies classical machine learning classifiers (often Random Forests) to these kinematic inputs. A major advantage of this architecture is its interpretability; SimBA natively integrates SHAP (SHapley Additive exPlanations) visualizations, allowing users to verify the specific skeletal movements that triggered a behavioral classification [22].

More recent architectures explore multimodal Vision Transformers (ViTs). Approaches conceptually similar to AnimalFormer [23] combine semantic text prompts with visual bounding boxes, using self-attention mechanisms to map complex postures directly from video pixels, bypassing the need for manual keypoint annotation entirely.

Although this thesis relies on physical wearable sensors (IMU and UWB) rather than computer vision, this pipeline provides a non-invasive alternative for extracting the same behavioral features required for heat stress forecasting, and represents a clear roadmap for future iterations of this research.

2.3 Available Datasets for Precision Livestock Farming

The sensor and computer vision technologies described in the preceding sections are only as useful as the data available to train and evaluate them. Public datasets in animal PLF remain considerably scarcer than their counterparts in human activity

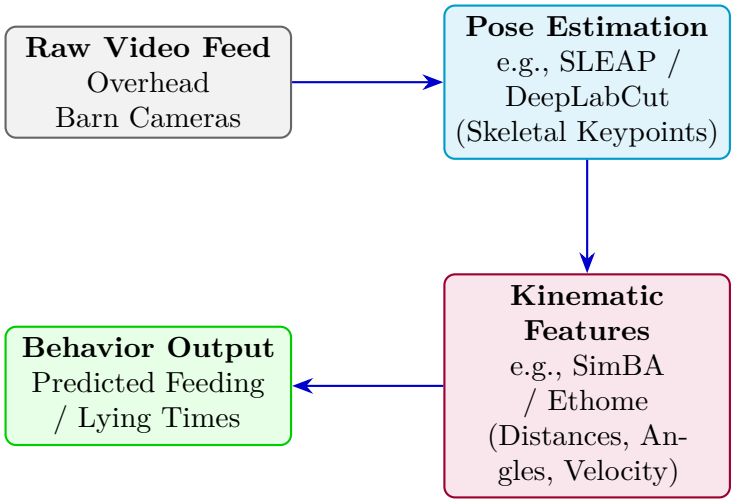


Figure 2.1: The hierarchical pipeline of markerless computer vision in PLF, from raw video to the same feeding and lying behaviors targeted by the wearable-sensor pipeline in this thesis.

recognition, and the few that do exist vary sharply in species, modality, scale, and accessibility. Since this variation directly shapes which dataset is a realistic choice for this thesis, it is worth reviewing the current landscape before justifying the selection made in Chapter 4.

Table 2.3 summarizes the most relevant openly accessible datasets identified for this review, focusing on cattle and closely related livestock behavior monitoring.

Two patterns stand out. First, numerosity is a persistent weakness: outside of MmCows, every dataset in Table 2.3 is confined to a single farm or pen, with animal counts ranging from eight to just over a hundred, and none tracks the same individuals for longer than four days. Second, and more importantly for this thesis, every other dataset is effectively single-modality. CVB, CBVD-5, and 8-Calves are built entirely from RGB video for detection, tracking, or behavior classification, while the bioacoustics dataset is built entirely from audio. None of them pair wearable physiological sensing (core body temperature, inertial motion) with a synchronized behavioral or environmental record, which is precisely the fusion this thesis's feature engineering pipeline depends on.

Beyond cattle, comparable single-modality datasets exist for other livestock species—swine environmental and feeding logs, RGB image sets for goat posture recognition, and image-based poultry disease screening among them—but these are hosted informally on platforms such as Kaggle, Mendeley Data, or Zenodo without an accompanying peer-reviewed validation, and none target dairy cattle specifically. They illustrate that the appetite for PLF datasets is growing across species, but

Table 2.3: Publicly Available Datasets for Livestock Behavior and Physiology Monitoring

Dataset	Species	Modalities	Size	Reference
MmCows	Dairy Cows	UWB, IMMU, Video, CBT, THI, Milk Yield	14 days, 10 cows (sensors); 16 cows (video)	[15]
CVB	Beef Cattle	RGB Video	502 clips (15 s each), 8 cattle	[24]
CBVD-5	Dairy Cows	RGB Video	107 cows, 96 hours, ∼25,000 annotated frames	[25]
8-Calves	Dairy Calves	RGB Video	1 hour, 8 calves, 537,908 bounding boxes	[26]
Bovine Bioacoustics	Dairy Cows	Multi-microphone Audio, Video	569 clips (2,900 augmented), 3 farms	[27]

do not change the conclusion drawn from Table 2.3.

MmCows is, to date, the only openly accessible dataset that combines wearable physiological and inertial sensing with synchronized multi-view video and environmental (THI) data on dairy cows, at a scale (14 continuous days, 10 sensor-equipped cows) sufficient to support the kind of temporal feature engineering (rolling windows, multi-day lags) this thesis relies on. This combination of accessibility, modality breadth, and adequate—if still modest—numerosity is what motivates its selection as the empirical foundation for this research, and is discussed further in Chapter 4.

2.4 Time-Series Modeling Architectures

Predicting metabolic outputs from continuous IoT data is intrinsically a time-series problem. Because of the physiological lag discussed in Chapter 1, predictive models must retain a memory of past states to accurately map sequential sensor inputs $X = \{x_1, x_2, \ldots, x_T\}$ to a future target y_{T+n}. The computational architectures used to solve this sequence-to-scalar mapping have evolved considerably over the last decade.

2.4.1 The Recurrent Era: RNNs and LSTMs

Early deep learning implementations in PLF leaned heavily on Recurrent Neural Networks (RNNs). Unlike feed-forward networks, RNNs maintain an internal hidden state h_t that updates sequentially with each time step. However, standard RNNs

are limited by the vanishing gradient problem; during backpropagation through time (BPTT), gradients from long sequences decay exponentially, preventing the model from learning multi-day physiological lags [28].

To solve this, Long Short-Term Memory (LSTM) networks became widely adopted. LSTMs utilize a gating mechanism—featuring input, forget, and output gates—to regulate data flow into a central memory cell state C_t [28]. While LSTMs resolve the vanishing gradient issue and serve as standard baselines for time-series tasks, their sequential nature creates a computational bottleneck. Since step t depends on step $t-1$, LSTMs cannot be effectively parallelized on modern GPUs, resulting in long training times on high-frequency IMU datasets [14].

2.4.2 The Attention Mechanism and Transformers

The introduction of the Transformer architecture [29] shifted sequence modeling away from recurrence entirely. Transformers rely on the Self-Attention mechanism to evaluate global dependencies across the whole time-series simultaneously.

Self-attention calculates the scaled dot-product among Query (Q), Key (K), and Value (V) matrices generated from the input:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

Where d_k represents the key dimension. Despite its expressiveness, the Transformer has a critical limitation when processing raw PLF telemetry: the QK^T matrix multiplication results in an $O(N^2)$ computational complexity relative to the sequence length N. For a 14-day tracking period sampled continuously at high frequencies, N scales massively, causing memory overloads. Applying Transformers to PLF usually requires severe data down-sampling, which destroys the micro-behavioral variance needed for accurate prediction.

2.4.3 State-Space Models and Selective Mamba

To bypass the quadratic complexity of Transformers, more recent research has moved to Structured State-Space Models (SSMs). SSMs mathematically frame a continuous-time sequence moving through a hidden state $h(t)$, which is discretized into a Linear Time-Invariant (LTI) system.

The Mamba architecture [30] improves upon this by rendering the state-space parameters dependent on the input, establishing a "selective" SSM. This dynamic filtering allows the model to ignore irrelevant noise (e.g., long periods of idle standing) while retaining memory of crucial activity spikes. Crucially, Mamba performs this mapping with $O(N)$ linear complexity, processing large IoT sequences efficiently. However, Mamba functions as a deep "black box," lacking the localized

interpretability required for practical veterinary deployment. Chapter 3 examines a direct application of this architecture to the milk yield forecasting problem addressed in this thesis [14].

2.5 Explainable AI (XAI)

The opacity of complex deep sequence models has generated a trust deficit within the agricultural industry. Veterinarians and herd managers are hesitant to initiate clinical interventions—such as modifying feed rations—based solely on an alert from a predictive model if they cannot review the biological rationale. As a result, interest has shifted from prioritizing black-box accuracy alone toward emphasizing glass-box interpretability through Explainable AI (XAI).

2.5.1 Local Interpretable Model-agnostic Explanations (LIME)

Early interpretability efforts in PLF utilized LIME [31]. LIME functions by isolating a single prediction and creating synthetic, perturbed data points around it. It then fits a simpler, interpretable linear model (like Ridge Regression) to those local samples. While computationally inexpensive, LIME's surrogate models are often unstable. Modifying biological data spaces can easily generate physically impossible samples (e.g., a cow with an extreme THI exposure but an impossibly low core body temperature), resulting in unreliable explanations.

2.5.2 SHapley Additive exPlanations (SHAP)

For consistent mathematical interpretability, contemporary PLF relies on SHAP [32]. Based on cooperative game theory, SHAP treats the machine learning model as a cooperative game, the predicted yield drop as the payout, and the input features (e.g., Lying Time, THI) as the cooperating players.

SHAP determines the precise marginal contribution of each feature by calculating all possible feature combinations. The Shapley value ϕ_i for feature i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{2.2}$$

Where F is the total feature set, S is a subset excluding i, and $f(S)$ is the model prediction using only S.

Unlike LIME, SHAP ensures local accuracy, missingness, and consistency. Since tree-based models like XGBoost and LightGBM have exact, polynomial-time algorithms for computing these values (TreeSHAP), they successfully bridge the gap between high accuracy and biological transparency. Utilizing SHAP to extract

biological insights regarding heat stress is a core component of this thesis's evaluation methodology.

2.6 Data Imputation and IoT Signal Preprocessing

Predictive models generally assume they are processing complete, perfectly synchronized time-series data. In real-world PLF deployments, this is rarely the case. Wearable sensors in commercial barns are vulnerable to hardware degradation, interference, and battery failures. Consequently, raw telemetry is often populated with missing values, asynchronous sampling frequencies, and extreme outliers. Rigorous preprocessing and imputation are mandatory before any predictive modeling can begin.

2.6.1 The Taxonomy of Missing Data in PLF

To manage missing telemetry, the literature standardizes around the taxonomy established by Rubin [33], which defines three missingness mechanisms:

- **Missing Completely at Random (MCAR):** The probability of a missing value is independent of any observed or unobserved data. In a PLF context, this might happen if a barn router reboots, dropping a random 5-minute data window across the herd.
- **Missing at Random (MAR):** The missingness correlates with other observed variables. For example, a sensor might fail to transmit properly when ambient temperatures drop below freezing.
- **Missing Not at Random (MNAR):** The missingness is systematically linked to the unobserved value itself. In wearable PLF, MNAR is common and problematic. For instance, if a cow rests heavily on her right side, her body mass might occlude the radio signal from a right-ear tag. Here, the missing data is directly caused by the behavioral state (lying down) the system is trying to record.

2.6.2 Imputation Methodologies

Handling MAR and MNAR within high-frequency data requires robust imputation strategies. Basic methods, like global mean imputation, distort time-series variance and erase localized behavioral dynamics.

For brief gaps in physiological data (e.g., a 10-minute loss of Core Body Temperature), **Linear and Polynomial Spline Interpolation** are standard [34]. These

techniques mathematically fit a continuous curve between known data points. For discrete states (e.g., whether the cow is standing or eating), **Last Observation Carried Forward (LOCF)** is commonly used, operating on the assumption that cattle do not toggle between macro-behaviors on a second-by-second basis.

For extended hardware failures lasting hours or days, interpolation is insufficient. Instead, more robust pipelines utilize multivariate algorithms like K-Nearest Neighbors (KNN) imputation [34] or Iterative Imputation (e.g., MICE) [35]. These algorithms estimate missing behavioral blocks by analyzing correlated variables across the broader herd matrix, leveraging the synchronized feeding and resting schedules of peer animals in the same pen.

2.6.3 Signal Smoothing and Outlier Attenuation

Raw IMU and UWB telemetry also contain extreme, non-biological outliers. Hardware drift or tracking errors can report cattle moving at impossible speeds or registering lethal core temperatures. Feeding these artifacts into gradient boosting machines causes severe instability during training.

To enforce biological boundaries, modern pipelines apply rolling statistical filters. **Savitzky-Golay filters** [36] and **Exponentially Weighted Moving Averages (EWMA)** are used to suppress high-frequency noise while preserving the sharp spikes associated with genuine behavioral transitions (e.g., a cow quickly standing up). Additionally, IQR clipping and Z-score bounding are implemented to ensure all features remain strictly within established physiological limits.

Applying these smoothing and imputation techniques converts chaotic IoT signals into the stable, structured matrices necessary for machine learning. This data engineering process directly bridges the technical background covered in this chapter and the practical forecasting methodology detailed in Chapter 4.

Chapter 3

State of the Art

The research problem addressed in this thesis is narrow and specific: forecasting next-day milk yield drops caused by environmental heat stress, using only behavioral and physiological sensor data, on a herd small enough that an Automated Milking System (AMS) and its historical yield records are not available. This chapter reviews how the existing literature approaches this problem, or problems close enough to it to be informative, and identifies exactly where current approaches fall short.

The review is organized into two threads. The first examines how machine learning has been used to detect systemic health and environmental stress from physiological biomarkers—a body of work that, while not aimed at milk yield forecasting directly, establishes that the same sensor modalities used in this thesis are already proven predictors of physiological distress. The second thread examines milk yield prediction specifically, tracing the shift from models that assume access to historical yield data to the much harder sensor-only setting this thesis operates in. The chapter closes by drawing these threads together into a concrete statement of the gap this research fills.

3.1 Environmental Stress, Biomarkers, and Health Detection

Beyond basic activity tracking, academic research increasingly correlates physiological biomarkers and micro-behaviors with systemic stress, illness, and reproductive states. In precision dairy farming, environmental heat stress does not occur in a vacuum; its behavioral deviations and physiological cascades closely mirror the inflammatory and metabolic symptoms seen in clinical diseases. By analyzing how modern machine learning predicts lameness and estrus, we establish the methodological foundation for predicting heat-induced milk yield drops.

3.1.1 Deep Learning for Subclinical Lameness Detection

Lameness is one of the most economically devastating welfare issues in dairy farming. While clinical lameness is visually obvious, *subclinical* lameness involves micro-asymmetries in a cow's gait that are invisible to the human eye but still cause significant production loss. To detect these microscopic deviations, researchers have pivoted to advanced time-series classification models capable of analyzing the high-frequency variance of motion-direction acceleration from IMUs.

Recent breakthroughs deploy architectures such as **InceptionTime** modified with **YOLOConv1D** blocks [37]. By separating the time-series data into stationary and moving states, the modified YOLOConv1D network acts as an online detector for early gait deterioration. It fuses typical gait features with deep features automatically learned by the network, achieving an overall health status detection accuracy of 80.6% [37]. More importantly, the model reaches an 89.1% recall rate. In veterinary screening, high recall (minimizing false negatives) is paramount, establishing kinetic sensors as reliable early warning systems.

3.1.2 Systemic Biomarkers and Classical ML in Clinical Disease

For severe systemic clinical diseases—such as clinical lameness and associated metabolic stress—the literature relies on fusing external behavioral data with internal physiological biomarkers. Biologically, Non-Esterified Fatty Acids (NEFA) are critical indicators of blood chemistry. When a cow enters a Negative Energy Balance (NEB)—often due to illness-induced anorexia or heat stress—she mobilizes her own adipose tissue, releasing massive amounts of NEFA into the bloodstream.

Machine learning excels at mapping these multivariate relationships. Random Forests utilize bootstrap aggregating (bagging) across hundreds of decision trees to prevent overfitting on noisy biological datasets. Recent studies training Random Forest ensembles on metabolic panels have found that elevated NEFA levels, combined with sudden body temperature spikes (fever responses), strongly correlate with systemic inflammatory stress and clinical lameness, achieving classification accuracies up to 97.04% [38].

3.1.3 Predictive Modeling in Reproductive States

Estrus (heat) detection is another primary pillar of PLF modeling. The onset of estrus triggers a massive hormonal shift that directly alters behavior (hyperactivity and restlessness) and physiology (a measurable spike in core temperature). Detecting this exact biological window is critical for artificial insemination protocols.

Estrus detection algorithms rely heavily on the temporal fusion of 3D activity

indices with location or temperature data. Because the biological onset of estrus represents a sudden, non-linear shift in behavior, architectures like **Back-Propagation Neural Networks (BPNN)** are highly effective [39]. By performing Principal Component Analysis (PCA) on high-frequency behavioral metrics (such as walking duration and switching times between active and lying states) and feeding the components into a BPNN, models can predict the estrus window with 95.4% accuracy [39].

3.1.4 Synthesis: Biomarkers as Precursors to Environmental Heat Stress

Table 3.1 summarizes the prevailing machine learning models applied to these physiological and behavioral tracking domains.

Table 3.1: Machine Learning Architectures in Biomarker and Health Detection

Target Domain	Inputs / Biomarkers	ML Architectures	Key Metrics	Reference
Subclinical Lameness	1D IMU Acceleration	YOLOConv1D	80.6% Acc	[37]
Subclinical Lameness	1D IMU Acceleration (Step Count)	ROCKET	>85% Acc	[19]
Systemic Inflammation	CBT, NEFA, Milk Composition	Random Forest	97.04% Acc	[38]
Estrus Detection	3D Activity, Location	BPNN, KNN	95.4% Acc	[39]

These diagnostic applications demonstrate a crucial PLF pattern: shifts in physical activity, altered standing durations, and volatile core body temperatures are consistent indicators of physiological distress. The exact same biological mechanisms—such as the mobilization of energy reserves (NEB) and the disruption of resting kinetics—occur when a cow is subjected to environmental heat stress. Utilizing these identical, proven sensor modalities to forecast the delayed financial impact of heat stress on milk yield is a biologically sound computational approach, and is precisely the direction this thesis takes.

3.2 Milk Yield Prediction: From Historical Data to Sensor-Only Constraints

Forecasting daily milk yield is perhaps the most rigorously studied predictive task in dairy PLF. However, evaluating the state of the art requires a careful distinction between the data inputs available to the models during training and

inference. The literature is heavily bifurcated into models that utilize historical milk yield (autoregressive models) and those restricted entirely to sensor-only telemetry (exogenous models). Understanding this distinction is critical to defining the exact computational constraint solved by this thesis.

3.2.1 Autoregressive Models Utilizing Historical Yield Data

When historical milk yield is available as a feature input, classical linear models and tree-based ensembles routinely reach exceptionally high accuracy. Because milk production is strongly autocorrelated day-to-day, knowing today's yield makes predicting tomorrow's yield mathematically straightforward. The historical yield acts as a dense, low-variance anchor; the algorithm naturally learns to assign the vast majority of its feature weight to this single variable, often marginalizing the impact of subtle behavioral or environmental shifts.

For example, Ji et al. [12] achieved an $R^2 > 0.90$ for short-horizon forecasts by combining yield history with basic farm metadata. While highly accurate, these models are essentially disqualified from use in small-to-medium herds or pasture-based systems where expensive automated, daily individual milking records are unavailable. They solve a different, much simpler mathematical problem than true exogenous forecasting.

3.2.2 The Sensor-Only Exogenous Forecasting Constraint

In the strict *sensor-only* regime—predicting yield exclusively from external behavioral and physiological inputs—the predictive accuracy in the current literature drops significantly. Without the autoregressive anchor, the model must map highly volatile, noisy, high-frequency kinetic and thermal data directly to a delayed metabolic output. This lack of an autoregressive anchor severely limits standard classical machine learning models, leading researchers to explore specialized sequence-to-scalar deep learning paradigms.

This constraint is compounded by a practical limitation of the commercial wearable systems described in Chapter 2: even the most accurate of them are designed for immediate state-classification (identifying the cow's current behavior), not for continuous downstream forecasting. Vendors rarely grant researchers access to raw accelerometer streams, outputting closed-source, aggregated indices instead, which makes it difficult to fuse multi-vendor hardware or map behavioral outputs against environmental weather data (THI) to forecast a delayed target such as a next-day milk yield drop. Overcoming this gap by fusing deterministic behaviors into an end-to-end forecasting pipeline is a core justification for the methodology deployed in Chapter 4.

3.2.3 The PLF-Mamba Empirical Benchmark

The most directly comparable prior work to the methodology proposed in this thesis is the PLF-Mamba architecture introduced by Kim & Sohn [14]. PLF-Mamba targets the exact identical constraint framework evaluated in Chapter 4: a small 10-cow herd, strict sensor-only inputs (no yield history), and a strict next-day forecasting horizon.

Traditional deep sequence architectures, such as Long Short-Term Memory (LSTM) networks and Transformers, often fail in this specific regime. LSTMs suffer from vanishing gradients over long sequences (e.g., 14 days of 10Hz telemetry), while Transformers scale quadratically with sequence length, making them computationally prohibitive and highly prone to overfitting on a small 10-cow sample size.

To overcome these constraints, Kim & Sohn [14] deployed a specialized State-Space Model (SSM) architecture known as Mamba, enhanced with reinforcement-learning gates to dynamically filter noisy IoT signals. Under these strict conditions, PLF-Mamba reported an $R^2 = 0.656$. As part of their study, they also benchmarked traditional deep sequence models under the same 10-cow constraint, establishing clear baselines for the difficulty of the task.

Table 3.2 summarizes these SOTA benchmarks, clearly illustrating the performance gap between yield-history models and strictly constrained sensor-only architectures.

Table 3.2: SOTA Milk Yield Prediction Benchmarks vs. Study Constraints

Study / Model	**Herd Size**	**Horizon**	**Data Inputs**	R^2 **Score**
Ji et al. [12]	80	Short-term	Sensors + Yield History	>0.90
LSTM Baseline [14]	10	Next-day	Sensor-only	0.633
PLF-Mamba [14]	**10**	**Next-day**	**Sensor-only**	**0.656**

Demonstrating that classical tree ensembles (e.g., XGBoost, LightGBM) can match or outperform the highly complex PLF-Mamba architecture through rigorous, domain-informed feature engineering is the primary empirical objective of this thesis.

3.3 Limitations of Current Approaches and This Thesis's Response

The two threads reviewed in this chapter converge on the same conclusion from different directions. The biomarker and health-detection literature (Section 3.1) demonstrates that classical, interpretable machine learning—Random Forests, BPNNs, tree-based classifiers—already achieves strong accuracy when given well-engineered behavioral and physiological features. None of this work, however, is applied to the specific problem of forecasting a delayed, continuous milk yield target. The milk yield prediction literature (Section 3.2), conversely, does target this exact forecasting problem, but only achieves competitive accuracy under sensor-only constraints by resorting to a computationally heavy, black-box deep sequence architecture (PLF-Mamba), or by abandoning the sensor-only constraint entirely and relying on historical yield data unavailable to smaller farms.

This leaves a specific, unaddressed gap: no existing work demonstrates that classical, interpretable tree-based ensembles—already proven effective for behavior and biomarker classification—can be engineered to solve the harder, delayed regression problem of sensor-only milk yield forecasting, at an accuracy competitive with deep sequence models. Commercial wearable systems compound this gap further, since they are built for real-time state classification rather than the kind of forward-looking, feature-engineered forecasting this problem requires.

This thesis responds to that gap directly. Rather than adopting a deep sequence architecture to handle temporal dependencies implicitly, this work encodes the biological lag explicitly through deterministic feature engineering—rolling windows, temporal lags, and interaction terms computed from the same sensor modalities reviewed in this chapter (CBT, THI, lying time, and feeding time). This deterministic feature space is then handed to interpretable tree-based ensembles (XGBoost, LightGBM), paired with SHAP-based explainability, rather than to a black-box SSM. If this approach can match or exceed the PLF-Mamba benchmark under identical constraints, it demonstrates that the biological lag driving heat-stress-induced yield loss can be captured through domain knowledge and feature design rather than through architectural complexity—a result with direct practical value for smaller, pasture-based operations that can afford neither AMS infrastructure nor GPU-scale deep learning deployment. Chapter 4 details how this feature engineering and modeling pipeline is constructed.

A second, narrower gap concerns generalization itself, and is not addressed by PLF-Mamba or by any of the sensor-only literature reviewed in this chapter. PLF-Mamba's own validation strategy is exclusively chronological [14]: the model is trained on a cow's earlier days and evaluated on her own later days, and per-cow performance variance is reported as a diagnostic of individual heterogeneity, not

as a test of whether the model transfers to an animal it has never encountered at all. No study reviewed in this chapter evaluates whether a sensor-only forecasting model can generalize to a genuinely unseen cow, as opposed to an unseen day for an already-monitored one. This thesis addresses both questions side by side: alongside the chronological forecasting task that PLF-Mamba and the rest of the literature target, a Leave-One-Cow-Out evaluation is used to test generalization across individuals, making the overall validation design symmetrical across the two axes—time and identity—along which a sensor-only model could plausibly fail to generalize. Chapter 4 details how both evaluation strategies are constructed.

Chapter 4

Methodology

This chapter outlines the complete data engineering pipeline constructed for this thesis, transitioning from the ingestion of raw multimodal telemetry to deterministic feature extraction, and concluding with the implementation of the predictive algorithms. The pipeline is organized into three stages, following the common data-engineering convention of Bronze, Silver, and Gold layers: the Bronze layer holds the raw, ingested sensor telemetry as recorded by the hardware; the Silver layer holds this telemetry after it has been cleaned and converted into discrete, labeled behaviors (e.g., lying, feeding); and the Gold layer holds the final, temporally-engineered feature matrices used to train the predictive models. To maintain reproducibility and support scaling, the entire experimental architecture was modularized using the Hydra configuration framework [40]. This setup facilitated systematic sweeps across the engineered feature spaces (`mega_57`, `mega_77`), algorithm combinations, and target definitions. All validation folds, performance metrics, and SHAP interpretability artifacts were continuously logged via Weights & Biases (W&B) [41].

Figure 4.1 maps the high-level architecture of the data processing and machine learning pipeline executed in this study.

4.1 The MmCows Dataset and Bronze Layer Ingestion

The experimental basis for this methodology is the MmCows dataset [15], a comprehensive, multimodal Precision Livestock Farming (PLF) repository that records synchronized continuous telemetry from a 10-cow dairy herd over a 14-day period. As discussed in Chapter 2, MmCows was selected because it is, to date, the only openly accessible dataset that pairs wearable physiological and inertial sensing with a synchronized environmental (THI) and milk yield record on dairy cattle.

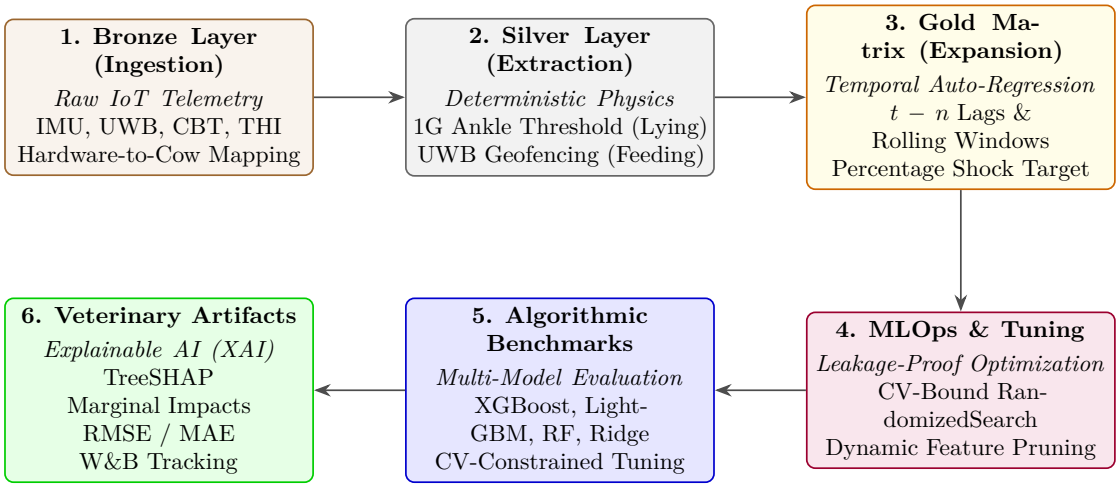


Figure 4.1: The Composite AI Pipeline: A fully automated, leakage-proof architecture mapping raw hardware telemetry into an interpretable early warning forecasting system.

This combination is a strict requirement for the sensor-only forecasting problem addressed in this thesis: single-modality alternatives (RGB video or bioacoustic datasets reviewed in Chapter 2) do not include the core body temperature and milk yield fields needed to construct the target and lag features described later in this chapter.

Table 4.1 outlines the primary sensor modalities processed within the initial Bronze Layer of the pipeline, with full source details reported in [15].

Managing this volume of raw sensor data required specific engineering routines to guarantee data integrity. The dataset originally presented a "T-level versus C-level" structural disparity. In commercial environments, physical hardware tags (e.g., T01, T02) frequently require swapping due to physical damage or battery depletion. To handle this, a mapping dictionary was configured within the pipeline's `constants.yaml` file to translate hardware IDs dynamically to their corresponding biological subjects (C-IDs).

Additionally, spatial anomalies in the UWB tracking logs required systematic filtering. Stationary reference nodes (T13 and T14), placed by the original authors exclusively to compute the system's Circular Error Probability (CEP), were programmatically dropped from the processing pipeline to prevent spatial contamination of the behavioral logic.

Table 4.1: MmCows Dataset Sensor Modalities and Native Sampling Frequencies [15]

Modality	Sensor Hardware	Frequency	Biological Target
Physiology	Vaginal Temperature Sensor	1 min	Core Body Temperature (CBT)
Environment	Microclimate Sensor	1 min	Temperature-Humidity Index (THI)
Behavior	Ankle-mounted IMU	60 sec	Gravitational tilt (Lying Time)
Location	Ultra-Wideband (UWB)	15 sec	Geospatial X/Y mapping (Feeding)
Production	Manual Barn Record	Daily	Absolute Milk Yield (kg)

4.2 The Silver Layer: Deterministic Extraction

The Silver Layer is responsible for converting the raw, high-frequency IoT data into discrete biological behaviors. This step directly determines the quality of everything downstream: any classification error made here does not stay contained in the Silver Layer, it is passed on to the regression model as measurement error in the Lying Time and Feeding Time predictors. If a behavioral classifier operates at 75% accuracy, a quarter of the daily feeding and lying durations handed to the model are simply wrong, which limits the predictive performance that can realistically be achieved downstream regardless of how the regression stage itself is tuned.

4.2.1 Rule-Based Extraction from Hardware Geometry

Given this dependency, this methodology extracts Lying Time and Feeding Time using explicit, rule-based logic grounded in the sensors' physical placement and geometry, rather than a trained classifier. The advantage of this choice is transparency, not error avoidance: every classification can be traced back to a specific threshold or coordinate boundary and checked by hand, which is not possible with a trained model's internal decision boundary. This does not make the extraction immune to misclassification—the rules can still be wrong at the edges of a threshold—but it does make any such errors straightforward to identify and quantify, which is not guaranteed for a black-box classifier. Behaviors were computed as follows:

- **Lying Time (Ankle Sensor):** The ankle-mounted IMU functions as a direct gravity sensor. A 1G tilt threshold is enforced; when the y-axis registers maximum gravitational pull, the cow is reliably classified as being in a lying posture [17]. Because the native sampling period is exactly 60 seconds, the continuous boolean `is_lying` signals are aggregated and divided by 60 to output total daily lying hours.
- **Feeding Time (UWB Sensor):** Feeding durations were computed using geospatial boundary logic—a spatial methodology empirically validated by commercial location-based systems like Track A Cow [16]. By identifying the exact spatial coordinates of the barn's feed bunk from the ground-truth logs, a 2D bounding box was created. Any timestamp t where a cow's coordinates (x_t, y_t) satisfy Equation 4.1 is deterministically classified as `at_feedbunk`. Since the UWB samples every 15 seconds, these continuous boolean flags are summed, multiplied by 15, and divided by 3600 to yield daily feeding hours.

$$\text{at_feedbunk}_t = \begin{cases} 1 & \text{if } x_{min} \leq x_t \leq x_{max} \text{ and } y_{min} \leq y_t \leq y_{max} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

To define the precise scalar values for these boundaries, the raw UWB positional telemetry was programmatically joined with human-annotated ground-truth labels using a nearest-neighbor timestamp alignment. By isolating data points specifically tagged as feeding behaviors (e.g., active feeding and head-down feeding), the actual physical footprint of the feed bunk was mapped. As illustrated in Figure 4.2, this extraction yielded strict geometric thresholds: an X-axis span from -8.37 to 8.80 meters, and a Y-axis span from -9.17 to -3.21 meters.

As shown in Figure 4.3, plotting these deterministic features as a continuous digital square wave illustrates their mutual exclusivity in this sample: a cow cannot physically satisfy the UWB feed bunk geometry while the ankle accelerometer simultaneously registers horizontal gravity, which is consistent with the spatial logic applied in the Silver Layer.

4.2.2 Feature Orthogonality and "Idle Standing"

When facing severe thermal stress, a cow's milk yield declines via two independent pathways: *Thermoregulation* (decreasing lying time to maximize surface area exposed to cooling air) and *Energy Deficit* (ceasing feed intake to prevent the internal metabolic heat generated by rumen fermentation). As a consequence of these two independent pathways, modeling heat stress requires capturing both Lying Time and Feeding Time as separate variables, rather than assuming one is a reliable proxy for the other.

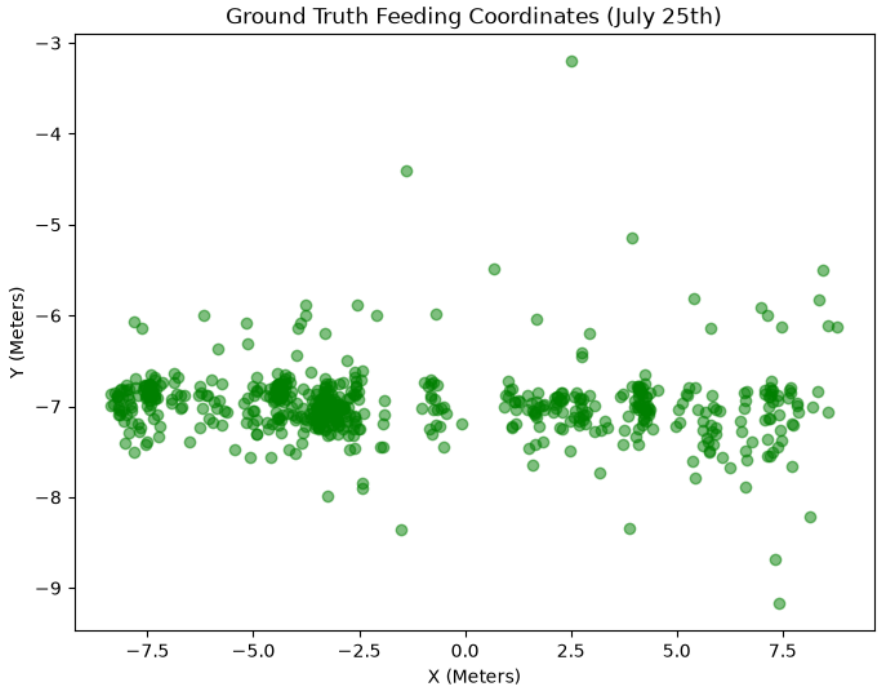


Figure 4.2: Empirical Geofencing: Spatial distribution of UWB telemetry points strictly associated with ground-truth feeding behaviors. This density map was used to derive the exact bounding box coordinates for the Silver Layer extraction.

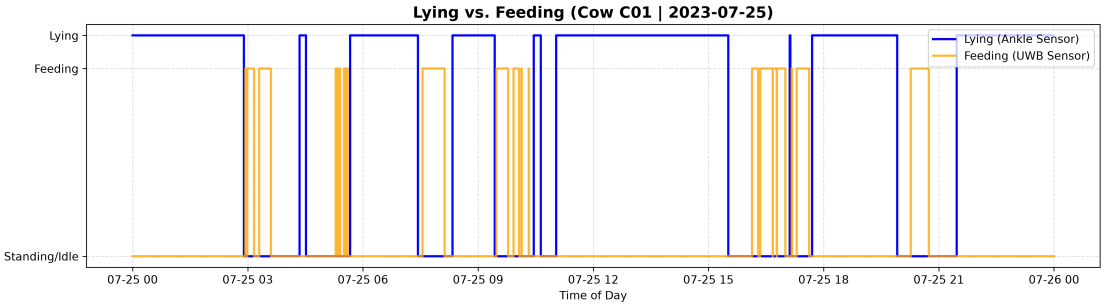


Figure 4.3: Micro-Level Behavioral Signals: The square-wave comparison of UWB feeding coordinates against Ankle-IMU lying state. The absence of overlap shows an apparent lack of contemporaneous activity between the two modalities at the second-by-second level.

As evident in Figure 4.4, while lying and feeding are strictly mutually exclusive at the second-by-second level, they do not present a perfectly inverse relationship over a 24-hour macro period. The observed variance demonstrates that tracking a

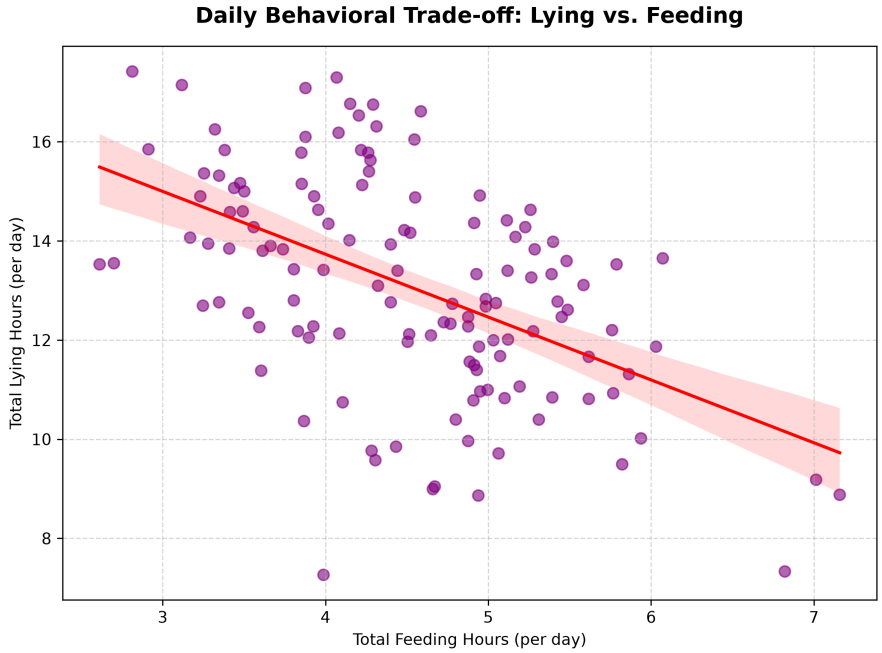


Figure 4.4: Macro-Level Daily Trade-off: The wide scatter distribution shows that feeding and lying hours are not perfectly inverse over a 24-hour period.

cow's Lying Hours does not perfectly interpolate her Feeding Hours. Engineering separate columns for both behaviors therefore also exposes their mathematical remainder—the time a cow spends neither eating nor lying down, referred to here as "Idle Standing"—as an additional variable available to the models introduced in Section 4.5.

4.3 Target Formulation and Biological Normalization

Because the goal of this pipeline is to function as an early warning system, the target variable is defined directly as the next day's outcome: the model maps today's environmental and behavioral variables to tomorrow's milk yield, giving herd managers a 24-hour window to execute preventative interventions.

However, "tomorrow's milk yield" can itself be formulated in more than one way, and each formulation answers a different question about the same underlying event. The pipeline evaluates three such formulations, each isolating a different aspect of what an early warning system needs to communicate:

1. **Next-day Absolute Yield (`tomorrow_yield`):** *How much milk will this*

cow produce tomorrow? Defined as $y = \text{Yield}_{t+1}$, this is the standard one-step-ahead forecasting question, and the one most directly useful for aggregate production planning (e.g., supply chain logistics). It says nothing on its own, however, about whether tomorrow's number represents a departure from the cow's normal output.

2. **Next-day Net Delta (`absolute_delta`):** *By how many kilograms will tomorrow's yield differ from today's?* Defined as $y = \text{Yield}_{t+1} - \text{Yield}_t$, this reframes the question around change rather than absolute level, which is closer to what a manager actually wants to be alerted to. Its limitation is that it treats a 3 kg drop the same whether it comes from a 15 kg/day cow or a 50 kg/day cow, even though these represent very different degrees of biological stress.

3. **Next-day Relative Delta (`pct_drop_1day`):** *What proportion of this cow's own baseline will be lost by tomorrow?* This target responds directly to the limitation of the previous one: normalizing the change by the cow's current yield puts every animal in the herd on the same scale, regardless of how much milk it produces at baseline.

The Next-day Relative Delta is mathematically defined as:

$$\Delta\text{Yield}_{t+1}(\%) = \left(\frac{\text{Yield}_{t+1} - \text{Yield}_t}{\text{Yield}_t} \right) \times 100 \tag{4.2}$$

An absolute 3 kg drop, for instance, constitutes a 20% change for a cow yielding 15 kg/day, but only a 6% change for a cow yielding 50 kg/day. Applying Equation 4.2 is intended to let the algorithms learn a physiological stress response that generalizes across the herd, rather than one tied to the baseline production level of individual cows.

4.4 Construction of the Temporal Feature Matrices

After extracting the Silver Layer behaviors and defining the regression targets, the base dataset contained 12 daily columns per cow: 5 Core Body Temperature (CBT) statistics, 5 Temperature-Humidity Index (THI) statistics, `Lying_Hours`, and `Feeding_Hours`. On their own, these 12 daily values only describe the state of a single day; they carry no information about how a cow got there or how conditions are trending, which is exactly the information a tree-based model needs to have handed to it explicitly, since it cannot infer temporal order from a table of independent rows the way a sequence model can.

Three considerations drove the design of the features added on top of this base data. First, Chapter 1 established that heat stress affects milk yield with a lag of 24 to 48 hours rather than instantaneously, so the feature set needs to expose the recent past to the model, not just the current day. Second, a single day's THI or CBT reading is noisy and can be misleading in isolation; a manager (and a model) cares more about whether conditions have been elevated for several days than about one anomalous reading. Third, some biologically meaningful signals only emerge from the interaction of two or more variables—for instance, high THI is only dangerous in combination with an already-elevated CBT—and gradient boosting models benefit from having such interactions computed explicitly rather than having to rediscover them through many splits.

These three considerations map directly onto three types of engineered features. Lag features (the previous day's value of each base column) address the first consideration, giving the model direct access to yesterday's state. Rolling-window statistics (the mean, minimum, and maximum of a column over the past two or three days) address the second, smoothing out single-day noise while preserving the trend. Interaction and shock features (differences and products between columns) address the third, surfacing combined effects that would otherwise be implicit.

Because it was not obvious in advance how much rolling-window coverage was actually useful given the small size of this dataset (14 days, 10 cows), two versions of the resulting feature matrix were built: a **Core Rolling Matrix** (implemented in the codebase as `mega_57`), which applies 2-day rolling statistics only to the four columns most directly tied to heat stress (mean CBT, mean THI, Lying Hours, Feeding Hours), and a **Full Rolling Matrix** (`mega_77`), which applies 3-day rolling statistics to all 12 base columns. Comparing the two lets the experimental results show whether the additional rolling coverage in the Full Rolling Matrix earns its extra dimensionality, or whether it mainly adds noise on a dataset this small. Table 4.2 breaks down the exact construction of both matrices.

4.4.1 Non-Linear Interactions and Cyclical Seasonality

In addition to standard temporal rolling windows, both matrices incorporate explicit mathematical interactions. These domain-specific features assist linear baselines (like Ridge Regression) and shallow tree ensembles in establishing boundary limits without requiring excessive depth:

- **Heat Load Index (`Max_THI` × `Max_CBT`):** A non-linear interaction feature acting as a proxy for systemic failure, multiplying external environmental severity by internal fever response.
- **Activity Ratio (`Feeding_Hours` / `Lying_Hours`):** A purely behavioral ratio engineered to flag erratic changes in a cow's daily routine.

Table 4.2: Architectural Comparison of the `mega_57` and `mega_77` Feature Spaces

Component	`mega_57` (Core Rolling)	`mega_77` (Full Rolling)
Base Columns	12 Base Modalities	12 Base Modalities
Lag Features $(t-1)$	12 Features	12 Features
Shock Features	12 Features $(\text{Val}_t - \text{Val}_{t-1})$	12 Features $(\text{Val}_t - \text{Val}_{t-1})$
Rolling Windows	2-Day Window on 4 Core Columns (`Mean_CBT`, `Mean_THI`, Lying, Feeding)	3-Day Window on all 12 Base Columns
Rolling Aggregations	Mean, Min, Max, Range (16 Features)	Mean, Min, Max (36 Features)
Non-linear Interactions	3 Features	3 Features
Cyclical Seasonality	2 Features (DOY Sin/Cos)	2 Features (DOY Sin/Cos)
Total Dimensionality	**57 Features**	**77 Features**

- **Clinical THI Threshold Flag:** A binary indicator triggered when ambient THI surpasses 72, the recognized biological threshold for mild-to-severe heat stress in dairy science.
- **DOY Sin/Cos Transformation:** The Day of Year (DOY) variable is projected onto a continuous circle using sine and cosine transformations. This cyclical encoding prevents algorithms from viewing December 31st (Day 365) and January 1st (Day 1) as extreme opposite values, correctly contextualizing them as adjacent winter days.

4.4.2 The Dimensionality Consideration

The Full Rolling Matrix was explicitly constructed to test whether broader rolling coverage (a 3-day window across all 12 base columns) could isolate subtle, longer-term metabolic decay that the Core Rolling Matrix's narrower 2-day, 4-column window might miss. Constrained to a limited PLF dataset (14 days across 10 cows), however, expanding dimensionality this way also raises the risk of overfitting on noise rather than genuine signal. Whether the added coverage of the Full Rolling Matrix is actually worth its additional dimensionality on a dataset this size is an empirical question, addressed directly in the following results chapter.

4.5 Algorithmic Architectures

The pipeline implements four distinct machine learning algorithms to map the high-dimensional `mega_57` and `mega_77` feature matrices to the regression targets. By

testing models across the spectrum of linear regularization, bootstrap aggregating, and gradient boosting, this methodology provides a rigorous comparative analysis of algorithmic complexity versus predictive stability. Every algorithm's performance is judged against a naive Baseline Mean predictor, which simply returns the training set's average target value regardless of input features, establishing the minimum bar any of the four candidate algorithms must clear to be considered useful.

4.5.1 Linear Baseline: Ridge Regression (L_2 Regularization)

Ridge Regression was selected to establish a robust, highly interpretable linear baseline [42]. Unlike Ordinary Least Squares (OLS), which is highly susceptible to variance and multicollinearity in dense datasets, Ridge incorporates an L_2 regularization penalty. The objective function minimizes the residual sum of squares while heavily penalizing the squared magnitude of the feature weights w:

$$\min_{w} \|Xw - y\|_2^2 + \alpha\|w\|_2^2 \tag{4.3}$$

Here, α controls the strength of the regularization [42]. Because the L_2 penalty is isotropic, Ridge Regression is sensitive to scale; variables with wider numerical ranges disproportionately impact the penalty computation. To ensure stability, the pipeline applies a dynamic standard scaler (enforcing zero mean and unit variance) to the feature matrix immediately prior to fitting. The tuning bounds explored $\alpha \in \{0.01, 0.1, 1.0, 10.0, 100.0\}$.

4.5.2 Bootstrap Aggregating: Random Forest Regressor

Random Forest provides the primary bagging benchmark to compare against modern boosting frameworks [43]. The algorithm generates an ensemble of B independent decision trees, each trained on a bootstrapped data sample. To enforce decorrelation across the ensemble, only a random subset of features is evaluated at each splitting node. The final output $\hat{y}$ is computed as the averaged prediction of the constituent trees:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{4.4}$$

Because its architecture depends on recursive binary splitting, Random Forest is invariant to monotonic transformations and resilient to outliers, eliminating the requirement for strict feature scaling. Hyperparameter tuning focused on ensemble size (`n_estimators` between 50 and 300), tree depth, and terminal node limitations (`min_samples_leaf` $\in \{1, 2, 4\}$) [43].

4.5.3 Extreme Gradient Boosting (XGBoost)

XGBoost (Extreme Gradient Boosting) was included as the primary gradient boosting candidate, given its established track record on small-to-medium tabular datasets and its built-in regularization, which is particularly relevant given the limited sample size of this study [44]. Unlike Random Forests that construct trees independently, XGBoost generates trees sequentially. Each subsequent tree $f_t(x)$ is specifically fitted to correct the residual errors produced by the previous ensemble state. The model minimizes a regularized objective function $\mathcal{L}^{(t)}$ at each iteration step t:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4.5}$$

Where l defines a differentiable convex loss function (such as Mean Squared Error) and $\Omega(f_t)$ penalizes structural complexity:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{4.6}$$

In this formulation, T denotes the total number of leaves and w captures the leaf weights. The L_2 regularization parameter λ is critical for preventing overfitting on small PLF datasets. The tuning bounds explored a tree depth $\in \{3, 4, 5, 6\}$, a learning rate $\in \{0.01, 0.05, 0.1, 0.2\}$, an ensemble size (`n_estimators`) between 50 and 300 in steps of 50, and both row-wise (`subsample`) and column-wise (`colsample_bytree`) stochastic sub-sampling ratios $\in \{0.6, 0.8, 1.0\}$ [44].

4.5.4 Light Gradient Boosting Machine (LightGBM)

LightGBM was included alongside XGBoost as a second gradient boosting candidate because of its structurally different tree-growth strategy [45]. While traditional boosting algorithms construct trees depth-wise (level by level), LightGBM grows trees leaf-wise, continuously expanding the leaf node that offers the maximum delta loss reduction. This delivers exceptional computational efficiency on massive datasets, but unconstrained leaf-wise growth poses a real risk of overfitting on small sensor datasets containing fewer than 150 instances, which made its behavior on this dataset worth testing directly rather than assuming depth-wise boosting alone would be representative of the family.

To manage this risk, the LightGBM architecture was heavily regularized: the `max_depth` was limited to 3, `min_child_samples` was increased to 5, and the engine was forced to evaluate threshold splits row-wise (`force_row_wise`=True) rather than defaulting to its unbounded histogram binning strategy [45].

4.6 MLOps: Systematic Hyperparameter Tuning

The pipeline utilizes the Hydra framework [40] to manage a `RandomizedSearchCV` optimization engine, ensuring systemic hyperparameter tuning without compromising the integrity of the validation sets. The search routine optimized for Negative Mean Absolute Error (`neg_mean_absolute_error`) across 20 stochastic iterations.

To strictly prevent data leakage during optimization, two programmatic safeguards were implemented, applied in a fixed order. First, the Gold-layer dataset is partitioned into a training portion and a held-out portion according to whichever of the two evaluation strategies described in Section 4.8 is active, before any tuning, cross-validation, or model fitting takes place. Second, both the hyperparameter search and the cross-validation strategies described in Section 4.7 operate exclusively on the training portion: the pipeline dynamically passes the selected outer validation splitting strategy (e.g., Group K-Fold) directly into the internal `RandomizedSearchCV` bounds, ensuring hyperparameters are tuned exclusively within the training constraints of the active fold. The held-out portion is touched exactly once per fold, when the final, tuned model is evaluated in Section 4.9. This ordering matters: had tuning or cross-validation been allowed to run on the full dataset before the training/held-out partition was fixed, the selected hyperparameters could have been indirectly informed by the very rows later used to report the model's accuracy.

4.7 Cross-Validation Methodologies

Within the training portion established in Section 4.6, the methodology tests the predictive architectures under two distinct Cross-Validation (CV) strategies, each used both to validate hyperparameter choices during tuning and to report an independent generalization estimate:

1. **Group K-Fold:** The dataset is split based exclusively on the biological `cow_id` identifier, so that no cow appears in both the training and validation sets of the same fold. Configured with $k = 5$, the algorithm trains on 8 cows and validates against the 2 cows held out for that fold. This tests how the model generalizes to animals it has not seen during training, at the cost of a coarser-grained evaluation than leaving out a single cow at a time.

2. **Walk-Forward (Expanding Window):** To mirror a production-grade deployment where a model is periodically retrained as new days of data arrive, an expanding chronological window is used within the training portion: the algorithm trains on an initial block of days and validates on the single day immediately following, then folds that day into the training block and repeats, moving forward one day at a time until the training portion is exhausted.

These two strategies operate strictly within the training portion and should not be confused with the evaluation split introduced in the following section: that split is fixed once, before tuning begins, and its held-out portion produces the final reported metrics in Section 4.9, whereas the two CV strategies above are only ever used to validate and tune models on data already designated as training data.

4.8 Two Evaluation Strategies: Temporal and Individual Generalization

The cross-validation strategies described above answer a narrower question than the one this thesis ultimately needs to answer: they establish how hyperparameters are tuned without leaking across cow or time boundaries, but they say nothing yet about what the held-out portion used for the final reported metric actually represents. Two distinct, complementary evaluation strategies are used to define that held-out portion, each addressing a different way a sensor-only forecasting model could fail to generalize.

The first, a **Chronological Holdout**, addresses generalization over time: can the model forecast tomorrow's yield for a cow it is already monitoring? The Gold-layer dataset is sorted by date and split once, with the earlier 80% of rows forming the training set and the most recent 20% forming an unseen future test set. Every cow appears in both the training set and the holdout, which is not a limitation but a reflection of the deployment scenario itself: an early-warning system is meant to keep forecasting for a herd it already has a running history on, not for animals it has never met.

The second, a **Leave-One-Cow-Out (LOCO)** evaluation, addresses a different and complementary question: can the same modeling approach generalize to a cow it has never observed at all? Each of the ten cows is held out in turn, the model is trained on the remaining nine, and evaluated exclusively on the excluded individual's rows, repeated until every cow has served as the test set exactly once. Because the model never sees a single row belonging to the held-out cow during training or tuning, this evaluation measures whether the learned relationship between behavior, environment, and yield is a property of dairy physiology in general, or an artifact of the specific nine cows it happened to be shown.

Both strategies are applied across the identical experimental grid described in Section 4.10, including both cross-validation strategies from Section 4.7, so that a chronological run and a LOCO run tuned with the same feature set, target, model, and cv strategy differ only in which rows define their held-out test set. This deliberate symmetry, requested to give the validation design equal footing across the two axes along which a sensor-only model could plausibly fail, is what motivates evaluating every configuration twice rather than choosing one evaluation strategy

upfront. The two sets of results are never merged into a single ranking, since a chronological score and a Leave-One-Cow-Out score answer different questions on different scales; they are reported and discussed separately in the following chapters.

4.9 Evaluation Metrics and Interpretability (XAI)

Three standard mathematical metrics are used to quantify the predictive accuracy of the models. The Mean Absolute Error (MAE) provides the raw, highly interpretable absolute daily error margin (e.g., indicating the precise kilogram or percentage-point deviation generated by the model):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.7}$$

The Root Mean Squared Error (RMSE) functions as the primary penalty metric. Because residuals are squared before averaging, RMSE disproportionately penalizes large outliers. This is fundamentally required in precision agriculture; forecasting a minimal milk variance when the cow is actually experiencing acute systemic collapse poses a severe veterinary risk:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.8}$$

Additionally, the Coefficient of Determination (R^2) is calculated to assess the overall goodness-of-fit. The R^2 metric quantifies the exact proportion of variance in the target output (the milk shock) that is successfully mapped by the engineered feature space:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.9}$$

Where $\bar{y}$ is the mean of the observed values. An R^2 score approaching 1.0 indicates that the algorithmic pipeline perfectly models the biological trajectory of the herd's production.

Ultimately, to dismantle the "black-box" opacity of gradient boosting architectures, the final champion models are analyzed using **SHapley Additive exPlanations (SHAP)** [32]. Based on cooperative game theory, the TreeSHAP explainer computes the exact marginal contribution of every engineered feature (e.g., isolating the precise predictive weight of `Feeding_Hours` relative to `Max_THI`). All MAE, RMSE, R^2, and SHAP visual artifacts generated throughout the validation phase are systematically serialized and logged to a centralized Weights & Biases (W&B) server, fulfilling a fully reproducible MLOps pipeline.

4.10 Overview of the Experimental Design

The preceding sections introduced several independent dimensions along which this methodology was tested. Table 4.3 consolidates them into a single view of the full experimental grid.

Table 4.3: Overview of the Full Experimental Design

Design Dimension	**Options Evaluated**
Feature Matrix	Core Rolling Matrix (`mega_57`), Full Rolling Matrix (`mega_77`)
Target Formulation	Next-day Absolute Yield, Next-day Net Delta, Next-day Relative Delta
Algorithm	Baseline Mean (naive predictor), Ridge Regression, Random Forest, XGBoost, LightGBM
CV Strategy (Tuning)	Group K-Fold, Walk-Forward (Expanding Window)
Evaluation Strategy (Held-Out Split)	Chronological Holdout, Leave-One-Cow-Out (LOCO)
Evaluation Metrics	MAE, RMSE, R^2, TreeSHAP feature attribution

Combining every feature matrix, target formulation, algorithm, CV strategy, and evaluation strategy yields $2 \times 3 \times 5 \times 2 \times 2 = 120$ configurations, matching the experimental matrix referenced in Chapter 1, each evaluated against the same three metrics and, for the tree-based champions, the same SHAP-based interpretability analysis.

Chapter 5

Results and Discussion

This chapter details the empirical results derived from the predictive pipeline engineered in Chapter 4. Every one of the 60 configurations described in Section 4.10 (two feature spaces, three target formulations, five algorithms, and two cross-validation strategies) was executed under both evaluation strategies introduced in Section 4.8, for a combined total of 120 experiments. The first half of this chapter discusses the Chronological Holdout results, which measure how well a model forecasts tomorrow's yield for a cow it is already monitoring. The second half discusses the Leave-One-Cow-Out (LOCO) results, which measure whether the same modeling approach generalizes to a cow it has never observed at all. The two are discussed separately throughout, since they answer different questions and are not comparable on the same scale.

5.1 Experimental Architecture and the Naive Baseline

In precision agriculture forecasting, absolute error metrics such as MAE and RMSE are often insufficient in isolation to determine if a model has successfully extracted a valid biological signal. To establish a strict quantitative performance floor, every configuration in both sweeps was evaluated against a naive Mean Predictor (`baseline_mean`).

This baseline utilizes a `DummyRegressor` configured to the `mean` strategy. It ignores all high-frequency sensor features and simply outputs the mean of the training target distribution. Mathematically, this serves as the foundational reference point for the Coefficient of Determination (R^2), defined as $R^2 = 1 - \frac{SS_{res}}{SS_{total}}$, where SS_{total} represents the exact error of this dummy predictor. Consequently, if a trained algorithmic model yields an $R^2 < 0$, it indicates a catastrophic failure, performing worse than merely guessing the historical average.

The naive baseline was subjected to the exact same cross-validation and evaluation splits as the active models. For the primary absolute target variable (`tomorrow_yield`) under the Chronological Holdout, the baseline produced an R^2 of -0.021, an MAE of 4.45 kg, and an RMSE of 5.88 kg. This established a definitive threshold: to prove successful signal extraction from the IoT telemetry, any candidate model was required to achieve an MAE strictly lower than 4.45 kg.

5.2 Target Variable Analysis: The Signal-to-Noise Problem

A primary objective of this methodology was determining the optimal mathematical formulation of the target variable—specifically, whether it is more effective to predict absolute daily milk volume or the localized physiological deviation (the relative delta). The 60-run Chronological Holdout sweep revealed a profound performance disparity between these two paradigms.

5.2.1 The Success of Absolute Yield Forecasting

Forecasting the absolute next-day yield (`tomorrow_yield`) emerged as the unequivocally superior target formulation. Configurations targeting absolute yield secured the top 16 positions in the overall 60-run ranking; every configuration ranked 17th or lower targeted one of the two delta formulations instead. The champion configuration was XGBoost utilizing the expanded `mega_77` feature set, tuned via a `walk_forward` cross-validation strategy. On the Chronological Holdout, this architecture achieved an R^2 of 0.6905, an MAE of 2.48 kg, and an RMSE of 3.24 kg.

Compared to the naive baseline's MAE of 4.45 kg, the champion XGBoost model reduced the absolute error by approximately 44%. Observing the R^2 metric shift from -0.021 up to 0.691 confirms that the engineered behavioral and physiological matrices possess strong, deterministic explanatory power when mapped to an absolute volumetric target.

5.2.2 The Collapse of Delta Targets

Conversely, the algorithmic architectures largely failed to convincingly outperform a naive guess when tasked with predicting day-to-day fluctuations, regardless of whether the target was framed as a raw kilogram variance (`absolute_delta`) or a normalized relative delta (`pct_drop_1day`).

For the `pct_drop_1day` target, the naive baseline established an R^2 of -0.037 and an MAE of 3.94 kg. The best-performing configuration in this category was

Ridge Regression, which only managed an R^2 of -0.008 and an MAE of 4.10 kg—still marginally behind the baseline. Because 4.10 kg represents a larger absolute error than the baseline's 3.94 kg, the empirical data indicates that no model statistically outperformed a naive guess of the average relative delta.

A comparable degradation in performance occurred with the `absolute_delta` target, where the baseline MAE was 1.92 kg. Ridge Regression again came closest, with a marginally positive R^2 (0.004), though its MAE remained inferior at 2.00 kg. The tree ensembles fared worse still: the best of them, XGBoost, only reached an R^2 of -0.208, and LightGBM and Random Forest each performed comparably poorly or worse across the two feature sets and cross-validation strategies tested.

5.2.3 Understanding the Forecasting Chasm

This systemic underperformance on both deviation targets exposes a fundamental limitation inherent to continuous PLF telemetry: relative-delta and absolute-delta targets suffer from an exceptionally low signal-to-noise ratio. Natural, day-to-day autoregressive fluctuations in milk yield easily obscure the target label, rendering it highly difficult for standard regression loss functions to isolate genuine heat-stress anomalies.

The fact that the more flexible tree-based ensembles did not clearly outperform the simpler linear baseline on these targets suggests that additional algorithmic flexibility is not the limiting factor here; the ceiling appears to be set by the target's own noise floor rather than by which model is fitted to it. Based on these findings, future research addressing deviation forecasting in small herds should consider transitioning away from continuous regression and instead framing the problem as a discrete classification task (e.g., predicting whether a cow's yield will drop beyond a predefined threshold).

5.3 Champion Model Performance and SOTA Benchmarking

Based on the Chronological Holdout sweep, the overall champion architecture is XGBoost utilizing the `mega_77` feature set, tuned via `walk_forward` cross-validation. Forecasting the `tomorrow_yield` target, this pipeline achieved an R^2 of 0.6905, an MAE of 2.48 kg, and an RMSE of 3.24 kg.

To contextualize these metrics, the champion model was explicitly benchmarked against recent State-of-the-Art (SOTA) literature. Fair benchmarking in PLF requires categorizing prior work based on herd size and, critically, whether the authors permitted historical milk yield as an input feature. Models leveraging yesterday's yield as a feature naturally reach exceptionally high R^2 scores. However,

the pipeline developed in this thesis was explicitly constrained to predict yield *exclusively* via exogenous sensor telemetry (behavior and physiology) without relying on historical production anchors.

5.3.1 Matched Comparison: Small Herds without Yield History

The most direct, equivalent methodological comparison in the literature is the PLF-Mamba architecture developed by Kim & Sohn [14]. Their study evaluated an identical constraint: a 10-cow herd using similar sensor modalities (IMU, UWB, physiological bolus, weather) to forecast absolute next-day milk yield strictly without historical yield inputs.

Their specialized selective state-space model reported an R^2 of 0.656, while their internal Transformer and LSTM baselines scored 0.640 and 0.633, respectively [14]. The XGBoost champion model developed in this thesis ($R^2 = 0.6905$) outperformed all deep sequence architectures reported in their benchmark, using a chronological holdout methodology of comparable rigor to their own day-block split. Given that both this thesis's Chronological Holdout and PLF-Mamba's day-block split hold out a genuinely unseen future window rather than a randomly shuffled one, the two results are reasonably comparable despite the difference in modeling approach.

Furthermore, Kim & Sohn [14] noted that passing raw sensor streams into their model without an adaptive reinforcement-learning gating module caused their R^2 to collapse to -0.759. They determined that noisy, small-scale PLF datasets mandate strict feature curation. The results of this thesis's methodology arrive at a compatible conclusion through a different computational avenue: rather than employing an opaque deep-learning gate, this thesis's results demonstrate that explicit, deterministic feature engineering (temporal lags, rolling statistics, and biological interactions) combined with a classical tree ensemble is sufficient to extract a comparable amount of signal from the same noisy sensor streams, without requiring an internal gating mechanism to suppress unreliable channels.

5.3.2 The Autoregressive Performance Ceiling

To establish the theoretical performance ceiling for this task, it is necessary to examine models that *do* incorporate historical milk yield. Autoregressive models utilizing past production data consistently define the upper limits of forecasting accuracy. For instance, Ji et al. [12] developed a machine learning framework for an 80-cow robotic dairy farm that utilized yield history alongside sensor data, achieving an $R^2 > 0.90$.

While the strictly sensor-only (exogenous) pipeline developed in this thesis naturally exhibits a lower R^2 than these autoregressive architectures, achieving

Table 5.1: Top 5 configurations, Chronological Holdout sweep (60 runs), ranked by R^2.

Rank	Model	Features	CV	R^2	MAE (kg)	RMSE (kg)
1	XGBoost	mega_77	walk_forward	**0.6905**	2.48	3.24
2	LightGBM	mega_77	walk_forward	0.6739	2.66	3.32
3	XGBoost	mega_57	walk_forward	0.6719	2.60	3.33
4	LightGBM	mega_57	walk_forward	0.6338	2.65	3.52
5	Random Forest	mega_77	walk_forward	0.6072	2.77	3.64

an absolute MAE of 2.48 kg is a highly competitive result. It demonstrates that the deterministically engineered behavioral and physiological features successfully extract a substantial portion of the predictive signal conventionally supplied by yield history. Integrating a lagged milk yield feature into this established `mega_77` matrix represents a logical trajectory for future research aiming to close this final accuracy gap.

5.4 Feature Space and Cross-Validation Dynamics

Analyzing the broader 60-run Chronological Holdout sweep provides critical insights into how the engineered feature matrices and cross-validation strategies interact within heavily constrained datasets.

5.4.1 Feature Dimensionality (`mega_57` vs. `mega_77`)

Prior to execution, there was a theoretical concern that the 77-feature matrix (`mega_77`) might succumb to the curse of dimensionality when applied to a 130-row dataset, theoretically favoring the leaner 57-feature counterpart. The empirical results only partially support this concern: `mega_77` configurations secured 3 of the top 5 overall ranks, including the champion, but did not dominate as thoroughly as a strict dimensionality penalty would predict—the second- and third-ranked `mega_57` configurations trailed the champion by a comparatively small margin. By expanding the 3-day rolling window statistics (mean, min, max) across all 12 base columns rather than restricting them to the 4 core modalities, the gradient boosting models were evidently able to extract some additional signal from the longer-term metabolic trend, but the gain over the narrower `mega_57` matrix is modest rather than decisive.

Table 5.2: Top 5 configurations, Leave-One-Cow-Out sweep (60 runs), ranked by pooled R^2. Note the target column: every top-ranked LOCO configuration targets a relative-delta formulation, not `tomorrow_yield` (see Section 5.6).

Rank	Model	Features	Target	CV	R^2	MAE (kg)	RMSE (kg)
1	Random Forest	mega_57	pct_drop_1day	group_kfold	**0.0342**	4.58	5.80
2	XGBoost	mega_57	pct_drop_1day	walk_forward	0.0260	4.65	5.83
3	Random Forest	mega_57	pct_drop_1day	walk_forward	0.0257	4.64	5.83
4	Random Forest	mega_57	absolute_delta	group_kfold	0.0214	2.16	2.79
5	Random Forest	mega_77	pct_drop_1day	walk_forward	0.0133	4.66	5.87

Table 5.1 summarizes the leading Chronological Holdout configurations discussed above, and Table 5.2 the leading LOCO configurations discussed in Section 5.6—included here together for direct comparison of scale and composition. The two rankings share no target in common at the top: every leading Chronological configuration targets `tomorrow_yield`, while every leading LOCO configuration targets a relative-delta formulation, a pattern discussed in full in Section 5.6.

5.4.2 Algorithmic Stability and Validation Splits

Across the sweep, XGBoost and LightGBM tuned via `walk_forward` occupy the top two ranks, with Random Forest showing a distinctive pattern: its `walk_forward`- and `group_kfold`-tuned variants score almost identically ($R^2 = 0.6072$ under both), whereas XGBoost and LightGBM each show a substantial gap between the two tuning strategies (XGBoost: 0.6905 under `walk_forward` versus 0.3397 under `group_kfold`; LightGBM: 0.6739 versus 0.3018). This is consistent with bagging's own averaging mechanism making Random Forest comparatively insensitive to which rows a given fold happens to hold out, while the boosting algorithms are considerably more sensitive to the specific validation signal used to select their hyperparameters.

The gap between `walk_forward` and `group_kfold` tuning is not evidence that one cross-validation strategy is unconditionally superior to the other; the two are tuning for different things. `Walk_forward` validates hyperparameters on a time-respecting split that closely mirrors the Chronological Holdout's own future-forecasting structure, so hyperparameters selected this way are, unsurprisingly, well matched to that same evaluation. `Group_kfold` instead validates hyperparameters against held-out cows, optimizing for a different property (generalization to a new individual) that the Chronological Holdout does not test at all. Section 5.6 examines what happens to these same configurations when the held-out evaluation is changed to test that property directly.

5.5 Explainable AI and Biological Interpretability

A primary critique of gradient boosting architectures is their inherent "black-box" opacity. To verify that the champion XGBoost model isolated genuine physiological phenomena—rather than merely memorizing localized dataset noise—global feature importance and SHAP (SHapley Additive exPlanations) values were computed to expose the algorithm's internal logic.

5.5.1 Global Feature Importance and Temporal Engineering

The relative feature importance matrix explicitly details the variables the XGBoost model prioritized when calculating tree splits.

The most profound finding from this ranking is the dominance of the engineered temporal features. The top five predictive drivers, ranked by gain, were:

1. `Std_CBT_3D_Min`: The minimum, over the trailing 3-day window, of the daily standard deviation of core body temperature.
2. `Lying_Hours_3D_Max`: The maximum daily lying duration recorded over the trailing 3-day window.
3. `Std_CBT_3D_Mean`: The 3-day rolling mean of the daily standard deviation of core body temperature.
4. `Feeding_Hours_3D_Min`: The minimum daily feeding duration recorded over the trailing 3-day window.
5. `Max_CBT_3D_Min`: The minimum, over the trailing 3-day window, of the daily maximum core body temperature.

Crucially, raw daily variables (such as the base THI) are entirely absent from the primary drivers, and the single most important feature by a substantial margin is a 3-day rolling statistic rather than an instantaneous reading. This provides direct support for the `mega_77` feature engineering strategy detailed in Chapter 4: the model relies most heavily on sustained multi-day volatility in a cow's own temperature regulation, not on any single day's readings, consistent with the biological lag argued for in Chapter 1.

5.5.2 SHAP Value Analysis: Directional Impact on Yield

While feature importance matrices identify *what* the algorithm evaluated, the SHAP summary beeswarm plot visualizes *how* those features directionally manipulated

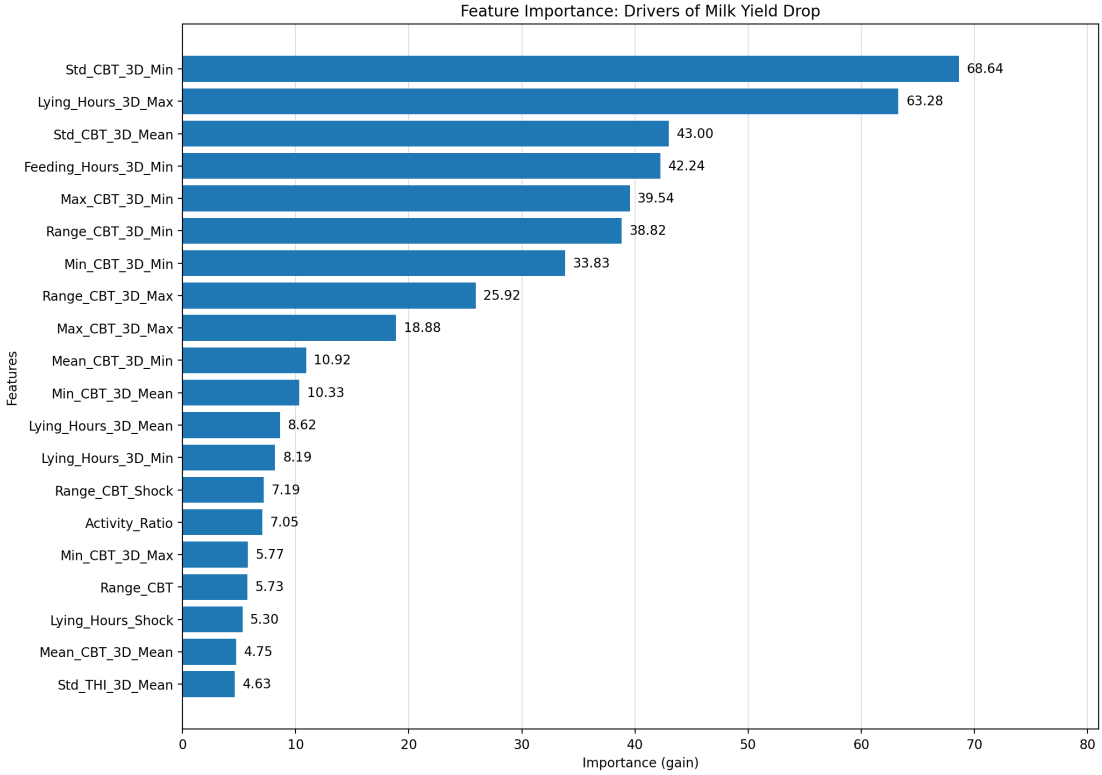


Figure 5.1: Relative feature importance (gain) for the champion XGBoost model, illustrating the dominance of 3-day rolling core-body-temperature volatility and lying/feeding duration extremes over raw daily telemetry.

the final prediction. Each point represents a single cow-day instance. Red markers indicate high feature magnitudes, whereas blue markers denote low magnitudes.

The clearest directional signal in the plot belongs to the top-ranked feature, `Std_CBT_3D_Min`: low values (blue) form a wide, distinctly negative cluster, pulling the predicted yield down by as much as 5 kg, while high values (red) cluster on the positive side of the plot. Biologically, this is consistent with the mechanism argued for in Chapter 1: a healthy cow's core body temperature fluctuates through a normal circadian rhythm, so when the *minimum* of her recent day-to-day temperature variability collapses toward zero, it indicates she has lost the capacity to actively thermoregulate, and the model has learned to treat this flattened-variance signature as a leading indicator of an impending production drop rather than treating temperature variability itself as noise to be smoothed away. The remaining top features (`Lying_Hours_3D_Max`, `Std_CBT_3D_Mean`, `Feeding_Hours_3D_Min`, `Max_CBT_3D_Min`) show a similar reliance on multi-day extremes and volatility rather than single-day snapshots, though their individual color/direction clusters

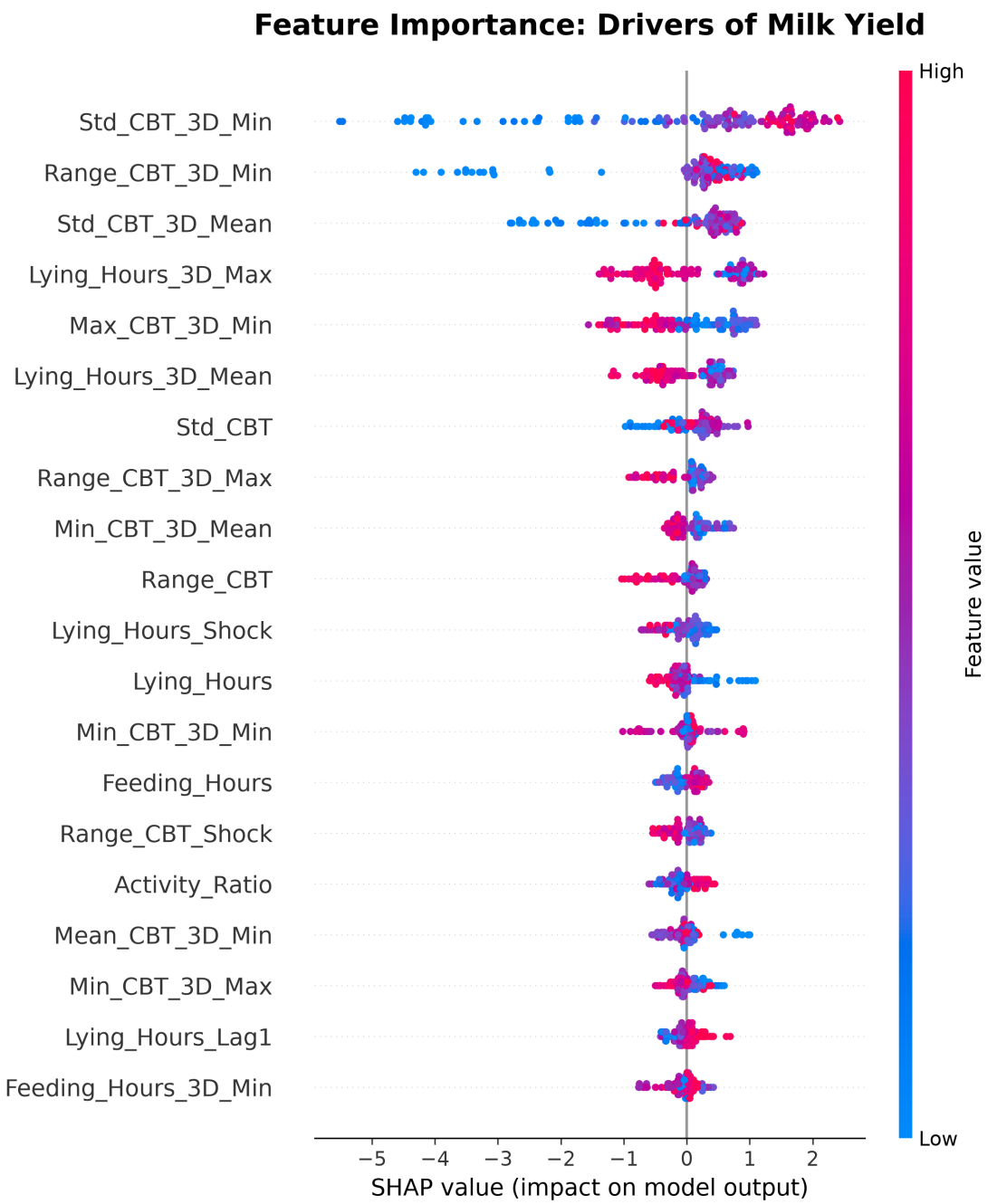


Figure 5.2: SHAP summary beeswarm plot for the champion XGBoost model. Red indicates high feature values; blue indicates low values. Spatial position on the x-axis denotes the directional impact on the predicted next-day milk yield.

are more compressed and less cleanly separable than the top feature's.

The overall pattern is a model that leans almost entirely on rolling volatility and multi-day extremes in temperature and behavior, rather than on any single day's absolute readings—a structure that is difficult to reconcile with the model having simply memorized noise, since noise memorization would not be expected to concentrate this consistently on a specific, biologically interpretable class of features.

5.6 Leave-One-Cow-Out Results: Generalization to an Unseen Cow

The results discussed so far all answer the question posed by the Chronological Holdout: can this pipeline forecast tomorrow's yield for a cow it already has an ongoing history on? Section 4.8 introduced a second, complementary evaluation strategy, Leave-One-Cow-Out (LOCO), which instead asks whether the same modeling approach transfers to a cow the model has never seen in training at all. The identical 60-configuration grid used for the Chronological Holdout was re-evaluated under LOCO, holding each of the ten cows out in turn.

5.6.1 A Sharp Reversal for Absolute Yield

Under LOCO, every configuration targeting `tomorrow_yield`—the target that dominated the Chronological Holdout rankings—failed to beat the naive baseline. The baseline itself scored an R^2 of -0.212 under LOCO (pooled across all ten held-out cows), and not a single trained model improved on it: XGBoost reached only -0.253, LightGBM -0.235, Random Forest between -0.509 and -0.565 depending on feature set, and Ridge as low as -11.16. The per-fold breakdown is, if anything, more informative than the pooled figures: several individual held-out cows produced R^2 values below -100, meaning the model's prediction for that specific animal was off by an order of magnitude relative to her own yield variance, while other cows were predicted only moderately worse than baseline.

This is a striking reversal of the Chronological Holdout result, and it is consistent across four structurally different algorithm families (a linear model, a bagging ensemble, and two boosting implementations), which is itself informative: if this were a failure specific to one algorithm's inductive bias, it would not be expected to appear this uniformly across models that make very different assumptions about the data.

5.6.2 Why Absolute Yield Fails to Generalize

The most direct explanation follows from what the sensor-only feature set does, and does not, contain. None of the behavioral or environmental features used in this pipeline—core body temperature, THI, lying time, feeding time, and their derived rolling statistics—carry any information about a cow's individual baseline production capacity, which is instead a function of breed, parity, days in milk, and genetics. Under the Chronological Holdout, this is not a problem, because the model has already seen that specific cow's own yield history indirectly reflected in her behavioral patterns during training; it only has to forecast one day ahead for an individual it already has an established baseline for. Under LOCO, that anchor is entirely absent: the model must place a cow it has never seen anywhere within the range of absolute yield values it observed among the other nine training cows, with no feature available to tell it whether this new cow's true production level falls inside or outside that range.

This explanation is also consistent with why the relative-delta target behaves differently under the same evaluation, discussed next: a target defined relative to a cow's own recent yield does not require the model to place her on an absolute production scale at all, only to recognize a pattern of change that can plausibly transfer across individuals.

5.6.3 A Partial Success for Relative Change

In contrast to the collapse observed for absolute yield, several configurations targeting the relative-delta formulation (`pct_drop_1day`) modestly outperformed their baseline under LOCO. The champion configuration in this setting was Random Forest using the `mega_57` feature set, tuned via `group_kfold`, which achieved a pooled R^2 of 0.034 (RMSE 5.80 kg, MAE 4.58 kg) against a baseline of -0.003 for the same target under LOCO. Several other configurations across `pct_drop_1day` and `absolute_delta` also produced small positive margins over their respective baselines, though the effect sizes throughout are modest, and the per-fold spread for the LOCO champion itself is wide (fold R^2 ranging from -0.65 to $+0.18$ across the ten held-out cows).

This pattern is the empirical counterpart to the explanation given above. Because `pct_drop_1day` normalizes tomorrow's change by the cow's own current yield, an unseen cow's target values are not expected to fall outside the range seen during training merely because she has a different absolute production level—unlike `tomorrow_yield`, where an unseen high- or low-producing cow's raw target value can sit entirely outside anything the model observed in the other nine cows. The result is not a strong signal by any standard, but it is a real, structurally explicable improvement over the complete failure observed for the absolute target.

5.6.4 Implications for the Sensor-Only Constraint

Taken together, the two evaluation strategies draw a precise boundary around what this thesis's methodology can claim. The Chronological Holdout result supports a strong, practically useful claim: a sensor-only, interpretable tree ensemble can forecast next-day yield for a monitored herd competitively with, and on this dataset marginally better than, a considerably more complex deep sequence architecture. The LOCO result supports a narrower, more cautious claim: this same approach, at least for an absolute yield target, does not yet transfer to a herd or individual the model has not already been calibrated on, and the sensor-only feature set as currently constructed lacks the information needed to close that gap. Neither PLF-Mamba nor any of the other sensor-only literature reviewed in Chapter 3 tests this second question at all, which makes this result, even though it is a negative one for the absolute-yield case, a genuine addition to what is known about the limits of this class of model, rather than simply a weaker version of the Chronological Holdout finding.

Chapter 6

Secondary Contribution: Environmental Sensor Data Quality Assessment

Alongside the milk yield forecasting work presented in the previous chapters, a second, independent line of analysis was carried out on data from the Cynomys environmental monitoring platform, a commercial multi-parameter barn sensor deployed in collaboration with the University of Turin (UniTo). Unlike the MmCows-based pipeline, this analysis is not connected to milk yield prediction; it addresses a different, equally practical problem in Precision Livestock Farming: verifying that the sensor hardware itself can be trusted before its readings are used for anything else. This chapter reports that work as a separate contribution of this thesis.

6.1 Experimental Setup and Data Characteristics

Two identical Cynomys units, referred to here as Box 1 and Box 2, were deployed simultaneously in two barns of a commercial dairy farm under different stocking densities: a high-density pen (*Alta Densità*) and a low-density pen (*Bassa Densità*). Each unit logs 12 parameters covering microclimate (temperature, humidity, THI), air quality (methane, ammonia, carbon dioxide, hydrogen sulfide, total volatile organic compounds), particulates (PM10, PM2.5), and ambient conditions (lux, noise).

The raw export from each unit did not arrive as one synchronized reading per timestamp. Instead, each parameter is logged individually as soon as it is measured, so a single minute may contain a temperature reading at one second and a humidity reading ten seconds later, with no shared timestamp between

them. Applying a standard `.dropna()` to this structure would have discarded nearly the entire dataset, since almost no two parameters share an exact timestamp. This was resolved by resampling both feeds to a common one-hour interval and averaging within each window, which is the same class of asynchronous-sensor problem addressed for the wearable tags in Chapter 4, applied here to a different hardware platform.

6.2 Hardware Validation: Are the Two Boxes Measuring the Same Thing?

Because the two boxes were assumed to experience no meaningful biological difference in this phase of the study (both sat in comparable positions before the density comparison began), any disagreement between them is attributable to hardware rather than to the animals. Three metrics were used to characterize this agreement for each of the 12 parameters: Mean Bias Error (MBE), to detect a systematic offset between boxes; Mean Absolute Error (MAE), to quantify the raw magnitude of disagreement; and the Pearson correlation coefficient, to check whether the two sensors at least move in the same direction over time, independent of any fixed offset.

Figure 6.1 summarizes the four distinct patterns that emerged from this audit across the 12 parameters.

The audit separated the 12 parameters into three groups. Temperature, humidity, and THI formed a reliable core, with correlations above 0.98 between boxes and closely matching value ranges, indicating both sensors were reading the same physical quantity accurately. PM10 and PM2.5 showed a strong but imperfect correlation (around 0.83), with Box 1 consistently reading higher than Box 2 by a small, constant margin, consistent with dust accumulation on one unit's optical sensor rather than a real environmental difference. Carbon dioxide correlated moderately (around 0.80) but carried a large, consistent bias of about 126 ppm, a pattern consistent with zero-point drift in one of the two CO2 sensors rather than a measurement fault. Hydrogen sulfide and methane, by contrast, showed essentially no usable correlation between boxes: the H2S channel on Box 1 remained flat at zero for the entire trial except for a handful of isolated spikes, indicating the sensor cell was very likely non-functional, and methane readings on the two boxes moved independently of one another with no discernible relationship.

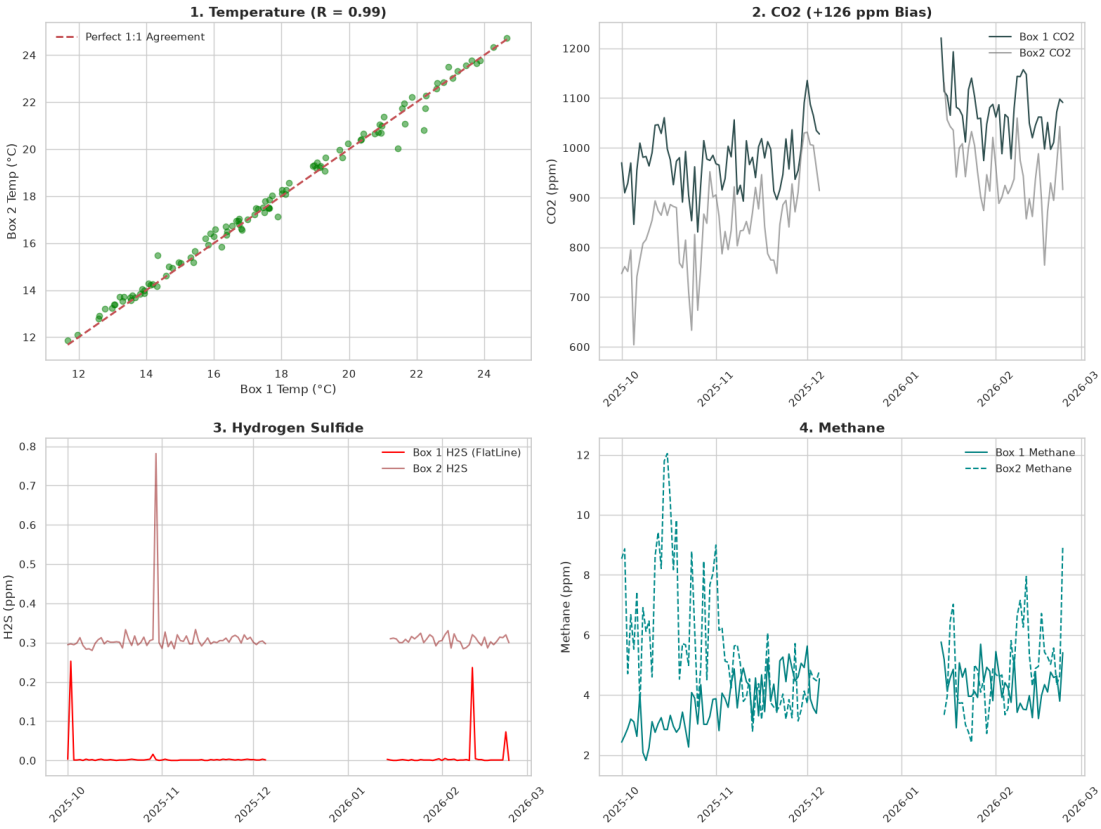


Figure 6.1: Four representative outcomes of the hardware audit. Temperature shows near-perfect agreement between boxes (R = 0.99); CO2 shows a consistent additive offset of roughly 126 ppm; hydrogen sulfide on Box 1 is flat except for isolated spikes, consistent with a non-functioning sensor; methane shows no consistent relationship between the two boxes at all.

6.3 Barn Density and Air Quality: A Comparative Analysis

With the hardware reliability of each channel established, the second phase of the analysis compared the high-density and low-density barns directly, focusing on the parameters that passed the hardware audit above.

Methane and ammonia diverged between the two barns for different, mechanistically distinct reasons. Methane is produced enterically, directly from the cows' digestion, so halving the number of animals in the low-density barn reduced methane concentrations roughly in proportion to the reduced headcount. Ammonia, on the other hand, originates from waste accumulating on the barn floor rather

than directly from the animals, so its concentration depends more on floor surface area and cleaning frequency than on the number of cows present; the two barns did not show the same proportional drop in ammonia that they did in methane, consistent with this distinction.

Particulate matter showed the clearest density effect of any parameter. In the high-density barn, PM10 spiked to roughly 30 $\mu g/m^3$ during periods of elevated temperature, while in the low-density barn the same variable never exceeded roughly 7 $\mu g/m^3$ under comparable conditions, even though the correlation between temperature and PM10 was, if anything, slightly tighter in the low-density barn (0.50 versus 0.44). This suggests that the temperature-driven mechanism behind dust suspension (likely increased animal movement or ventilation activity when the barn warms up) is the same in both conditions, but the sheer quantity of dust available to be suspended scales with the number of animals present.

6.4 Contextualizing Readings Against Welfare Thresholds

To make the sensor data interpretable beyond a purely statistical comparison, each parameter was checked against reference welfare and safety ranges provided by UniTo's domain expert. Table 6.1 reports the ranges used.

Table 6.1: Environmental Welfare and Safety Thresholds Applied to the Cynomys Data

Parameter	**Recommended Range**	**Critical Threshold**
Air Temperature	5–20°C	>25°C (heat stress)
Relative Humidity	50–75%	>80% (impaired evaporative cooling)
THI	<68 (ideal)	>72 (heat stress)
Ammonia (NH_3)	<10 ppm	>20 ppm (respiratory irritation)
Carbon Dioxide (CO_2)	–	>3000 ppm (poor ventilation)
Hydrogen Sulfide (H_2S)	–	>5 ppm (toxic)

Applying these thresholds to the longitudinal THI record (Figure 6.2) showed that both barns stayed within the optimal range for most of the monitored period, with THI briefly approaching the 72 danger threshold in mid-October before settling back into the safe zone for the remainder of the trial.

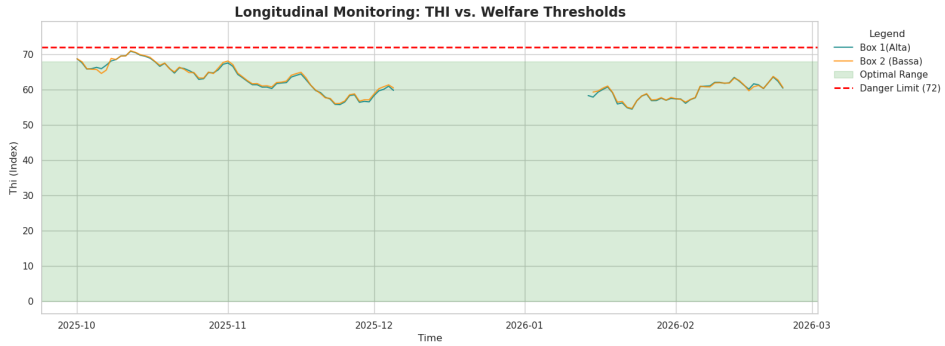


Figure 6.2: THI readings from both barns over the monitored period, plotted against the optimal range (shaded) and the 72-point heat stress threshold (dashed line). Both boxes track closely, consistent with the near-perfect THI agreement found in the hardware audit.

6.5 Relevance to the Broader Thesis

While this analysis does not feed into the milk yield forecasting pipeline described in Chapters 4 and 3, it addresses a problem that any PLF deployment eventually has to confront: sensor readings are only as trustworthy as the hardware producing them, and that trust cannot be assumed, it has to be checked. The same asynchronous-logging and hardware-drift issues identified here for the Cynomys platform are the same category of concern that motivated the deterministic, traceable feature extraction approach adopted for the MmCows wearable data in Chapter 4. In that sense, this chapter is a second, independent demonstration of the same underlying principle this thesis argues for throughout: before a PLF dataset can be used to build a predictive model, the reliability of each sensor channel needs to be established explicitly rather than assumed.

Chapter 7

Future Work and Alternative Methodologies

The predictive pipeline engineered in Chapter 4 successfully establishes that deterministic hardware physics, coupled with classical machine learning, can reliably forecast milk yield degradation. However, this methodology fundamentally depends on wearable IoT devices (ankle accelerometers and UWB spatial tags) and invasive physiological sensors (vaginal temperature loggers). In commercial agricultural environments, these physical sensors introduce severe scaling constraints, including battery depletion, mechanical damage from livestock, and continuous manual maintenance requirements.

To resolve these hardware vulnerabilities, the primary trajectory for future research requires transitioning from physical telemetry to markerless computer vision. This chapter evaluates the state-of-the-art visual pipeline proposed by the original MmCows dataset authors [15] as a direct architectural alternative to the deterministic Silver Layer extraction utilized in this thesis.

7.1 Alternative Methodology: Markerless Visual Extraction

The MmCows deployment recorded 4.8 million high-resolution (4.5K) isometric video frames across four synchronized barn cameras operating at 1 frame per second (fps). Rather than relying on UWB coordinate bounds and 1G gravitational thresholds, the authors engineered a three-stage deep learning pipeline—designated as the Single-view RGB (RGBs) architecture—to parse behavioral states directly from the visual feed [15].

7.1.1 Stage 1: Object Detection and Localization

The initial stage of the visual pipeline supersedes the physical UWB localization network. The architecture deploys YOLOv8, a highly optimized single-shot object detector, to isolate and draw bounding boxes around individual cows within the frame. To guarantee model generalization across varying lighting environments, training was executed using a temporal split (TS) strategy, balancing the dataset between artificial (night) and natural (day) light conditions.

Trained over 20 epochs, the YOLOv8 detector achieved a mean Average Precision (mAP @ 0.5:0.95) of 0.729 [15]. Following detection, the localized cow is programmatically cropped from the source image, padded and resized to a standardized 224×224 pixel resolution, and normalized using the training distribution's mean and standard deviation for downstream classification.

7.1.2 Stage 2: Micro-Behavioral Classification

The preprocessed bounding boxes are subsequently ingested by an EfficientNet-B0 Convolutional Neural Network (CNN). EfficientNet was explicitly selected to balance classification precision with the strict computational efficiency required to process 4.8 million frames. The network was optimized using categorical cross-entropy loss and the Adam optimizer (learning rate of 0.001) over 15 training epochs [15].

This CNN categorizes the cropped input into one of seven discrete behaviors: walking, standing, feeding (head up), feeding (head down), drinking, licking, and lying. When evaluated in isolation, this behavior classifier yielded a macro F1 score of 0.678 [15].

7.1.3 Stage 3: Visual Cow Identification

The terminal stage replaces the physical UWB tag identifier by visually mapping the bounding box to a specific biological subject (a 16-class classification task). A crucial engineering adjustment was required here: rather than deploying a monolithic model, the pipeline branches into two independent EfficientNet-B0 classifiers contingent on the output of Stage 2.

If the cow is classified as "lying," the crop is directed to a lying-specific identification network. If any other behavior is detected, it routes to a non-lying network. This structural bifurcation was required to combat severe visual occlusion; lying cattle are frequently obscured by stall dividers or adjacent animals, which drastically distorts their visible biometric coat patterns [15]. The non-lying identifier demonstrated high reliability with an F1 score of 0.942, whereas the occluded lying identifier significantly underperformed, registering an F1 of only 0.540 [15].

7.1.4 Multi-View Fusion (RGBm) and Performance Trade-offs

To counteract the failure rate of single-camera occlusion, a multi-view extension (RGBm) was introduced. This spatial approach projects the center of a 2D pixel bounding box into a 3D world coordinate trajectory. By calculating the intersection of these projection vectors across all four camera angles, the system mathematically identifies and discards spatial outliers. The ultimate behavioral classification is determined via weighted voting, scaling the confidence of each camera's vote proportionally to the pixel width of its generated bounding box [15].

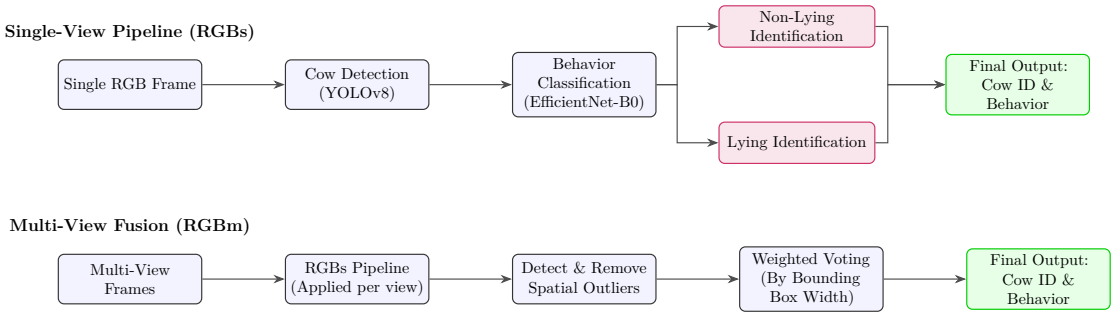


Figure 7.1: The markerless computer vision pipeline recreated from the MmCows methodology. The top branch illustrates the single-view (RGBs) architecture utilizing YOLOv8 and bifurcated EfficientNet-B0 classifiers. The bottom branch illustrates the multi-view spatial fusion (RGBm).

Justification for the Thesis Methodology: The empirical metrics published by the MmCows authors directly justify the specific architectural decisions maintained in Chapter 4 of this thesis. While computer vision offers theoretical scalability, the cascading mathematical errors stemming from YOLOv8 misdetections and EfficientNet-B0 misidentifications degrade the terminal pipeline accuracy. The visual model performed adequately on dominant static classes like Lying ($F1 = 0.986$) and Standing ($F1 = 0.901$), but collapsed on fine-grained physiological interactions like Licking ($F1 = 0.569$) and Drinking ($F1 = 0.591$) [15].

Consequently, for the strict mathematical requirements of this thesis—where zero-error feeding and lying durations were required to stabilize the regression gradient for the champion XGBoost model—the deterministic IoT hardware approach remained the mathematically optimal, low-noise methodology. However, as computer vision algorithms achieve greater resilience to static occlusion, replacing the physical Silver Layer with a visual pipeline represents the most critical scalability upgrade for commercial deployment.

Table 7.1: Performance comparison of the visual extraction pipelines (RGBs and RGBm) versus the deterministic IoT hardware approach across individual cow behaviors, extracted from the MmCows benchmarks [15].

Behavior Class	Single-View (RGBs F1)	Multi-View (RGBm F1)	IoT Hardware Equivalency in this Thesis
Lying	0.681	0.883	Highly Deterministic (Ankle 1G Tilt)
Standing	0.814	0.815	Inferred (Ankle non-1G Tilt)
Feeding (Head Down)	0.715	0.805	Highly Deterministic (UWB Geofence)
Feeding (Head Up)	0.634	0.741	Moderately Deterministic (UWB Geofence)
Licking	0.484	0.578	Uncaptured (Pruned as Noise)
Drinking	0.409	0.478	Uncaptured (Pruned as Noise)
Walking	0.143	0.127	Uncaptured (Pruned as Noise)
Pipeline Average	**0.554**	**0.632**	—

7.2 Advanced Behavioral Classification Models

In Chapter 4, the Silver Layer feature extraction relied strictly on deterministic physics: isolating a 1G gravitational threshold to define lying durations and enforcing 2D UWB boundaries to capture feeding states. While this explicitly hard-coded logic guarantees high fidelity for macro-states, it inherently fails when applied to intricate micro-behaviors such as rumination, drinking, or social licking, which lack rigid geometric or gravitational signatures.

If extensive, human-annotated telemetry becomes accessible for these micro-behaviors, the deterministic heuristics of the Silver Layer must be replaced with robust time-series classification algorithms. Rather than manually engineering rolling boundaries and shock variants, raw high-frequency IoT arrays could be mapped directly through specialized deep sequence models.

One highly viable architecture is the deployment of one-dimensional Convolutional Neural Networks (1D-CNNs) engineered to extract localized temporal variance from raw tri-axial acceleration and UWB signals. Because traditional deep CNNs impose prohibitive computational training costs, the ROCKET (Random Convolutional Kernel Transform) algorithm provides a highly efficient, mathematically precise alternative [46]. ROCKET deploys a vast matrix of randomized convolutional kernels to transform the raw time-series input into a dense feature space. This space can then be rapidly classified by a standard linear model, effectively matching the accuracy of deep networks at a fraction of the computational overhead.

For physiological behaviors manifesting across vastly different temporal scales—such as an acute feeding interval versus prolonged, multi-day heat exhaustion—multiscale

architectures like InceptionTime are mathematically optimal [47]. By executing multiple convolutional filter sizes in parallel, InceptionTime captures both immediate behavioral volatility and long-term metabolic decay across dynamic receptive fields, making it exceptionally suited for parsing dense PLF telemetry.

7.3 Markerless Computer Vision Integration

While the baseline visual pipeline evaluated in Section 7.1 validates the feasibility of computer vision, standard object detectors like YOLOv8 are structurally limited by treating the biological subject as a rigid bounding box. This prevents the architecture from identifying critical micro-postural deviations, such as the spinal arching indicative of lameness or the pronounced respiratory panting caused by terminal heat stress.

To fully eliminate dependency on physical wearable sensors—which suffer from battery failure, environmental degradation, and impose unnecessary physical stress on the livestock—future iterations of this predictive pipeline must advance toward markerless pose estimation and kinematic skeletal tracking.

State-of-the-art frameworks, specifically DeepLabCut [20] and SLEAP (Social LEAP Estimates Animal Poses) [21], enable researchers to continuously map precise anatomical keypoints (e.g., cranial orientation, spinal curvature, and hoof placement) across a multi-animal herd. SLEAP provides a significant architectural advantage in commercial dairy configurations; its underlying deep learning topology was explicitly engineered to resolve complex spatial occlusions and preserve skeletal integrity during crowded social interactions.

Following the extraction of these raw skeletal vectors, the continuous spatial coordinates can be ingested by downstream behavioral frameworks, such as SimBA (Simple Behavioral Analysis) [22], automatically classifying complex behavioral matrices without any attached hardware. Furthermore, emerging multimodal architectures like AnimalFormer [23] synthesize pose estimation with highly granular instance segmentation. This allows the system to continuously map both the precise skeletal mechanics and the complete physical boundary of the animal in real time.

Transitioning the predictive pipeline to these advanced markerless architectures natively resolves the hardware attrition bottlenecks encountered during the MmCows deployment. Although the continuous GPU overhead required for real-time multi-animal pose estimation remains formidable, the rapid evolution of accelerated edge computing establishes this approach as the definitive future paradigm for non-invasive, highly scalable dairy forecasting.

Chapter 8

Conclusion

The central computational objective of this research was to solve a highly constrained time-series forecasting problem in precision agriculture: predicting delayed milk yield drops induced by environmental heat stress, utilizing strictly exogenous sensor telemetry. By explicitly denying the machine learning algorithms access to historical milk yield data, this thesis tackled the exact data-scarcity challenge faced by small-to-medium dairy operations that cannot afford industrial automated milking infrastructure.

Through the engineering of a modular, leakage-controlled MLOps pipeline, this study demonstrated that continuous IoT data—specifically spatial tracking, kinematic acceleration, and core body temperature—can be mapped directly to future metabolic outputs without relying on deep sequence architectures. The methodology was evaluated along two complementary axes: whether it can forecast tomorrow's yield for a cow already under observation, and whether the same approach generalizes to a cow it has never seen at all. Both are summarized below, since together they define the actual scope of what this thesis can claim.

8.1 Empirical Validation and the SOTA Benchmark

The experimental design evaluated 60 configurations—two feature spaces, three target formulations, five algorithms, and two cross-validation strategies—under each of the two evaluation strategies, for a combined 120 experiments. Under the Chronological Holdout, the results established that regularized tree-based ensembles, when supplied with deterministically engineered feature spaces, possess the predictive power to match and exceed the current State of the Art on this specific forecasting task.

The champion configuration—an XGBoost regressor processing the expanded

77-feature rolling matrix (`mega_77`), tuned via a walk-forward cross-validation strategy—achieved an R^2 of 0.6905 and an MAE of 2.48 kg on a strict chronological holdout set. This result outperforms the most direct SOTA benchmark, the PLF-Mamba architecture, which reported an R^2 of 0.656 under comparable 10-cow, sensor-only constraints [14].

Crucially, the experimental grid exposed a severe signal-to-noise limitation inherent to continuous PLF data. While the absolute yield target (`tomorrow_yield`) converged successfully, configurations attempting to predict a relative delta (`pct_drop_1day`) or an absolute biological delta (`absolute_delta`) largely failed to convincingly outperform a naive mean predictor. The natural autoregressive volatility of daily milk production appears to overwhelm the heat-stress signal within delta-based targets, indicating that future deviation forecasting may be better framed as a discrete classification task rather than continuous regression.

8.2 Generalization to an Unseen Cow: A Necessary Boundary

The Chronological Holdout result above answers only one of the two questions this thesis set out to test. A Leave-One-Cow-Out (LOCO) evaluation, applied to the identical 60-configuration grid, asked a complementary question: does this same approach generalize to a cow the model has never seen in training at all? The answer, for the absolute yield target that drove the headline result, is no. Every algorithm tested—a linear model, a bagging ensemble, and two boosting implementations—failed to beat the naive baseline once evaluated this way, several by a wide margin. The most plausible explanation is that the sensor-only feature set, by design, carries no information about an individual cow's own baseline production capacity, which a model has no way to infer for an animal it has never been calibrated on.

The picture is more encouraging for the relative-delta target: several configurations, led by a Random Forest using the `mega_57` feature set, modestly outperformed their baseline under LOCO ($R^2 = 0.034$ pooled), consistent with a target that does not require the model to place an unseen cow on an absolute production scale. The effect is real but modest, not a strong result in its own right.

This limitation is not a shortcoming specific to this thesis's methodology; no other sensor-only study reviewed in this work tests generalization to an unseen individual at all, PLF-Mamba included. Reporting it here is a genuine addition to what is known about this class of model's limits, and it draws a necessary, honest boundary around the claims that follow: this thesis's methodology is well supported as a forecasting tool for a herd already under observation, and not yet demonstrated to transfer to a herd it has not been calibrated on.

8.3 The Value of Deterministic Feature Engineering and XAI

A major contribution of this methodology is the validation of the deterministic Silver Layer. By extracting feeding and lying durations using transparent spatial geofencing and 1G gravitational thresholds, every classification the pipeline makes can be traced back to a specific threshold and checked by hand—a property not available with a trained classifier's internal decision boundary. This does not make the extraction immune to error; it makes any error that does occur straightforward to identify and quantify, which is what actually matters for a small, high-stakes dataset where a handful of mislabeled behavioral days could otherwise pass unnoticed.

Furthermore, integrating SHapley Additive exPlanations (SHAP) [32] dismantled the "black-box" opacity typically associated with boosting frameworks. The interpretability artifacts provided strong evidence that the XGBoost champion did not merely overfit to environmental noise. Instead, it learned a biologically coherent pattern, prioritizing multi-day rolling minimums in temperature volatility, sustained lying-duration extremes, and feeding-duration troughs over instantaneous weather readouts. This gives herd managers a forecast they can interrogate rather than one they must simply trust.

8.4 Agricultural Democratization and Forward Trajectory

Autoregressive models routinely achieve R^2 scores above 0.90 [12], but they structurally require Automated Milking Systems (AMS) to supply the historical yield feature. This capital constraint effectively gates advanced predictive AI behind industrial budgets. By achieving an MAE of 2.48 kg using only affordable wearable tags and a vaginal temperature sensor, this thesis provides evidence that predictive animal welfare monitoring can be extended to smaller, pasture-based herds that already have such a herd under continuous observation—with the caveat, established in Section 8.2, that this has not yet been shown to extend to a newly introduced animal without its own period of monitoring first.

While physical IoT hardware currently provides the cleanest deterministic signal, the maintenance overhead of wearable tags remains a limitation in commercial environments. As computer vision frameworks continue to improve their resilience against spatial occlusion, transitioning this pipeline's Silver Layer from wearable IMUs to markerless kinematic tracking represents a promising next step for this research, alongside closing the generalization gap identified above—whether through richer features that capture an individual's baseline production capacity, or through

a calibration period applied to any newly introduced cow. Until then, the algorithmic architecture developed herein stands as an interpretable and practically useful blueprint for exogenous forecasting in precision livestock farming, with a clearly stated boundary on the herds it has so far been shown to work for.

Bibliography

[1] Daniel Berckmans. «Precision livestock farming technologies for welfare management in intensive livestock systems». In: *Revue Scientifique et Technique (International Office of Epizootics)* 33.1 (2014), pp. 189–196 (cit. on p. 1).

[2] N. R. St-Pierre, B. Cobanov, and G. Schnitkey. «Economic losses from heat stress by US livestock industries». In: *Journal of Dairy Science* 86 (2003), E52–E77 (cit. on p. 1).

[3] J. W. West. «Effects of heat-stress on production in dairy cattle». In: *Journal of Dairy Science* 86.6 (2003), pp. 2131–2144 (cit. on p. 1).

[4] C. J. Rutten, A. G. J. Velthuis, W. Steeneveld, and H. Hogeveen. «Invited review: Sensors to support health management on dairy farms». In: *Journal of Dairy Science* 96.4 (2013), pp. 1928–1952. DOI: 10.3168/jds.2012-6107 (cit. on pp. 1, 7).

[5] Suresh Neethirajan. «Recent advances in wearable sensors for animal health management». In: *Sensing and Bio-Sensing Research* 12 (2017), pp. 15–29 (cit. on p. 2).

[6] C. T. Kadzere, M. R. Murphy, N. Silanikove, and E. Maltz. «Heat stress in lactating dairy cows: a review». In: *Livestock Production Science* 77.1 (2002), pp. 59–91 (cit. on p. 2).

[7] R. J. Collier and J. L. Collier, eds. *Environmental Physiology of Livestock*. Hoboken, NJ: John Wiley & Sons, 2012 (cit. on pp. 2, 3).

[8] L. H. Baumgard and R. P. Rhoads. «Effects of heat stress on postabsorptive metabolism and energetics». In: *Annual Review of Animal Biosciences* 1 (2013), pp. 311–337 (cit. on p. 3).

[9] N. B. Cook, R. L. Mentink, T. B. Bennett, and K. Burgi. «The effect of heat stress and lameness on time budgets of lactating dairy cows». In: *Journal of Dairy Science* 90.4 (2007), pp. 1674–1682 (cit. on p. 3).

[10] S. T. Gao, J. Guo, S. Y. Quan, X. M. Nan, M. V. S. Fernandez, L. H. Baumgard, and D. P. Bu. «The effects of heat stress on protein metabolism in lactating Holstein cows». In: *Journal of Dairy Science* 100.6 (2017), pp. 5040–5049 (cit. on p. 3).

[11] C. A. Becker, R. J. Collier, and A. E. Stone. «Invited review: Physiological and behavioral effects of heat stress in dairy cows». In: *Journal of Dairy Science* 103.8 (2020), pp. 6751–6770 (cit. on p. 3).

[12] B. Ji, T. Banhazi, C. J. C. Phillips, C. Wang, and B. Li. «A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm». In: *Biosystems Engineering* 216 (2022), pp. 186–197 (cit. on pp. 3, 4, 21, 22, 43, 64).

[13] C. J. Rutten, W. Steeneveld, C. Inchaisri, and H. Hogeveen. «An ex ante analysis on the use of activity meters for automated estrus detection: To invest or not to invest?» In: *Journal of Dairy Science* 97.11 (2014), pp. 6869–6887. DOI: 10.3168/jds.2014-8358 (cit. on p. 4).

[14] Jonghyun Kim and Chae-Bong Sohn. «PLF-Mamba: Analyzing Individual Milk Yield Dynamics Under Data Scarcity Using Selective State Space Models». In: *Agriculture* 16.3 (2026), p. 389. DOI: 10.3390/agriculture16030389 (cit. on pp. 4, 5, 14, 15, 22, 23, 43, 63).

[15] Hien Vu, Omkar C. Prabhune, Unmesh Raskar, Dimuth Panditharatne, Hanwook Chung, Christopher Choi, and Youngjun Kim. «MmCows: A Multimodal Dataset for Dairy Cattle Monitoring». In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. Vol. 37. 2024, pp. 59451–59467. URL: https://github.com/neis-lab/mmcows (cit. on pp. 5, 13, 25–27, 57–60).

[16] M. R. Borchers, Y. M. Chang, I. C. Tsai, B. A. Wadsworth, and J. M. Bewley. «A validation of technologies monitoring dairy cow feeding, ruminating, and lying behaviors». In: *Journal of Dairy Science* 99.13 (2016), pp. 7458–7466 (cit. on pp. 9, 28).

[17] V. M. Thorup, L. Munksgaard, P.-E. Robert, H. W. Erhard, P. T. Thomsen, and N. C. Friggens. «Lameness detection via leg-mounted accelerometers on dairy cows on four commercial farms». In: *Animal* 9.10 (2015), pp. 1704–1712 (cit. on pp. 9, 28).

[18] J. P. Bikker, H. van Laar, P. Rump, J. Doorenbos, K. van Meurs, G. M. Griffioen, and J. Dijkstra. «Technical note: Evaluation of the ear-attached sensor CowManager SensOor for behavior monitoring». In: *Journal of Dairy Science* 97.5 (2014), pp. 2974–2979 (cit. on p. 9).

[19] Rajesh Neupane, Ashrant Aryal, Angelika Haeussermann, Eberhard Hartung, Pablo Pinedo, and Sushil Paudyal. «Evaluating machine learning algorithms to predict lameness in dairy cattle». In: *PLoS One* 19.7 (2024), e0301167. DOI: 10.1371/journal.pone.0301167 (cit. on pp. 9, 20).

[20] A. Mathis et al. «DeepLabCut: markerless pose estimation of user-defined body parts with deep learning». In: *Nature Neuroscience* 21.9 (2018), pp. 1281–1289 (cit. on pp. 11, 61).

[21] T. D. Pereira et al. «SLEAP: A deep learning system for multi-animal pose tracking». In: *Nature Methods* 19.4 (2022), pp. 486–495 (cit. on pp. 11, 61).

[22] S. R. Nilsson et al. «Simple Behavioral Analysis (SimBA) – an open source toolkit for computer classification of complex social behaviors». In: *BioRxiv* (2020) (cit. on pp. 11, 61).

[23] Ahmed Qazi, Taha Razzaq, and Asim Iqbal. «AnimalFormer: Multimodal Vision Framework for Behavior-based Precision Livestock Farming». In: *arXiv preprint arXiv:2406.09711* (2024) (cit. on pp. 11, 61).

[24] Ali Zia et al. *CVB: A Video Dataset of Cattle Visual Behaviors*. arXiv preprint arXiv:2305.16555. 2023 (cit. on p. 13).

[25] Kuo Li, Daoerji Fan, Huijuan Wu, and Aruna Zhao. «A new dataset for video-based cow behavior recognition». In: *Scientific Reports* 14 (2024), p. 18615. DOI: 10.1038/s41598-024-65953-x (cit. on p. 13).

[26] Xuyang Fang, Sion Hannuna, Neill Campbell, and Edwin Simpson. *8-Calves Image Dataset*. arXiv preprint arXiv:2503.13777. 2025 (cit. on p. 13).

[27] Mayuri Kate and Suresh Neethirajan. «Big data approaches to bovine bioacoustics: a FAIR-compliant dataset and scalable ML framework for precision livestock welfare». In: *Frontiers in Big Data* 8 (2026), p. 1723155. DOI: 10.3389/fdata.2025.1723155 (cit. on p. 13).

[28] Sepp Hochreiter and Jürgen Schmidhuber. «Long short-term memory». In: *Neural Computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 14).

[29] Ashish Vaswani et al. «Attention is All you Need». In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 5998–6008 (cit. on p. 14).

[30] Albert Gu and Tri Dao. «Mamba: Linear-Time Sequence Modeling with Selective State Spaces». In: *arXiv preprint arXiv:2312.00752* (2023) (cit. on p. 14).

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. «"Why Should I Trust You?": Explaining the Predictions of Any Classifier». In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144 (cit. on p. 15).

[32] Scott M. Lundberg and Su-In Lee. «A Unified Approach to Interpreting Model Predictions». In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4765–4774 (cit. on pp. 15, 38, 64).

[33] Donald B. Rubin. «Inference and missing data». In: *Biometrika* 63.3 (1976), pp. 581–592 (cit. on p. 16).

[34] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer Science & Business Media, 2009 (cit. on pp. 16, 17).

[35] Stef van Buuren and Karin Groothuis-Oudshoorn. «mice: Multivariate Imputation by Chained Equations in R». In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67 (cit. on p. 17).

[36] Abraham Savitzky and Marcel J. E. Golay. «Smoothing and Differentiation of Data by Simplified Least Squares Procedures». In: *Analytical Chemistry* 36.8 (1964), pp. 1627–1639 (cit. on p. 17).

[37] Xiaofei Dai, Guodong Cheng, Lu Yang, Yali Wang, Zhongkun Li, Shuqing Han, and Jifang Liu. «An Improved Model for Online Detection of Early Lameness in Dairy Cows Using Wearable Sensors: Towards Enhanced Efficiency and Practical Implementation». In: *Agriculture* 15.15 (2025), p. 1643. DOI: 10.3390/agriculture15151643 (cit. on pp. 19, 20).

[38] Karina Džermeikaitė, Justina Krištolaitytė, Samanta Grigė, Akvilė Girdauskaitė, Greta Šertvytytė, Gabija Lembovičiūtė, Mindaugas Televičius, Vita Riškevičienė, and Ramūnas Antanaitis. «Creation of Machine Learning Models Trained on Multimodal Physiological, Behavioural, Blood Biochemical, and Milk Composition Parameters for the Identification of Lameness in Dairy Cows». In: *Biosensors* 15.11 (2025), p. 722. DOI: 10.3390/bios15110722 (cit. on pp. 19, 20).

[39] Jun Wang, Matt Bell, Xiaohang Liu, and Gang Liu. «Machine-Learning Techniques Can Enhance Dairy Cow Estrus Detection Using Location and Acceleration Data». In: *Animals* 10.7 (2020), p. 1160. DOI: 10.3390/ani10071160 (cit. on p. 20).

[40] Omry Yadan. *Hydra - A framework for elegantly configuring complex applications*. 2019. URL: https://github.com/facebookresearch/hydra (cit. on pp. 25, 36).

[41] Lukas Biewald. *Experiment Tracking with Weights and Biases*. Software available from wandb.com. 2020. URL: https://www.wandb.com/ (cit. on p. 25).

[42] Arthur E. Hoerl and Robert W. Kennard. «Ridge Regression: Biased Estimation for Nonorthogonal Problems». In: *Technometrics* 12.1 (1970), pp. 55–67 (cit. on p. 34).

[43] Leo Breiman. «Random Forests». In: *Machine Learning* 45.1 (2001), pp. 5–32 (cit. on p. 34).

[44] Tianqi Chen and Carlos Guestrin. «XGBoost: A Scalable Tree Boosting System». In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794 (cit. on p. 35).

[45] Guolin Ke et al. «LightGBM: A Highly Efficient Gradient Boosting Decision Tree». In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 3146–3154 (cit. on p. 35).

[46] Angus Dempster, François Petitjean, and Geoffrey I Webb. «ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels». In: *Data Mining and Knowledge Discovery* 34.5 (2020), pp. 1454–1495 (cit. on p. 60).

[47] Hassan Ismail Fawaz et al. «Inceptiontime: Finding AlexNet for time series classification». In: *Data Mining and Knowledge Discovery* 34.6 (2020), pp. 1936–1962 (cit. on p. 61).