



J. Dairy Sci. TBC

<https://doi.org/10.3168/jds.2026-28889>

© TBC, The Authors. Published by Elsevier Inc. on behalf of the American Dairy Science Association®.
This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Fusing Heterogeneous Sensor Data Using Artificial Intelligence Hybrid Models for Dairy Cattle Health Anomaly Detection

Arnas Nakrošis,^{1*} Mahammad Ismayilov,¹ Gurban Shukurov,¹ Adhithiyan Arul Mani,¹ Sanjaykrishna Raveli Srinivas,¹ Joris Vēžys,² Vytautas Ostaševičius,³ and Agnė Paulauskaitė⁴

¹Department of Applied Informatics, Faculty of Informatics, Kaunas University of Technology, 51368 Kaunas, Lithuania

²Department of Mechanical Engineering, Mechanical Engineering and Design Faculty, Kaunas University of Technology, 51424 Kaunas, Lithuania

³Institute of Mechatronics, Kaunas University of Technology, 51424 Kaunas, Lithuania

⁴Artificial Intelligence Excellence Centre, Kaunas University of Technology, 51423 Kaunas, Lithuania

ABSTRACT

This observational study evaluated a hybrid artificial intelligence framework for same day anomaly detection in dairy cows using multimodal sensor data collected under commercial farm conditions. The framework integrated thermal video recorded during routine milking, milking-system measurements, and rumen bolus sensor data from 88 cows. Thermal recordings were processed to extract region-level temperature statistics and deep visual-temporal features using residual convolutional networks and Transformer-based temporal modeling. These thermal features were temporally aligned with production and physiological variables, including somatic cell count, milk yield, electrical conductivity, blood indicators, milk flow, internal body temperature, activity, rumination, and bolus event flags. The fused feature sets were evaluated using several machine-learning classifiers under different modality configurations to assess the contribution of thermal, milking, and bolus data. Model performance depended strongly on modality selection and decision-threshold optimization. The best selected configuration achieved a receiver operating characteristic area under the curve of 0.934 and an F1-score of 0.890, indicating strong discrimination between normal and anomalous health states. Results suggest that production and physiological sensor variables provide strong predictive information, whereas thermal imaging adds complementary value when combined with structured farm data. The proposed framework demonstrates the potential of multimodal artificial intelligence systems to support same day health anomaly detection and decision-support tools for dairy cow health monitoring in real farm environments.

Key words: Heterogeneous Sensor Fusion, Dairy Cattle Anomaly Detection, Machine Learning, Thermal Imaging

INTRODUCTION

Advanced artificial intelligence (AI) systems that use multimodal sensing are increasingly applied in Precision Livestock Farming (PLF) to support animal health management, productivity, and welfare. Recent reviews show rapid growth in livestock applications involving deep learning (Rohan et al., 2024), wearable sensing (Lamanna et al., 2025), lameness detection (Siachos et al., 2024), and automated animal identification (Tassinari et al., 2021; Jiang et al., 2023). These developments reflect a transition from single sensor monitoring toward integrated frameworks that can support continuous decision making in commercial livestock environments.

Despite this progress, the reliability and generalization of AI based livestock monitoring systems remain important challenges (Perez et al., 2025; Wang et al., 2025). Models developed under controlled conditions may perform less consistently in commercial barns because of sensor noise, missing measurements, animal movement, visual occlusions, posture variation, and communication instability. These limitations are especially relevant for dairy cattle health monitoring, where disease related changes may appear through several physiological, behavioral, and production related signals rather than through a single measurable variable. Monitoring systems intended for farm deployment must therefore integrate heterogeneous sensor streams and remain robust when some signals are incomplete or degraded (Gómez et al., 2021; Stygar et al., 2021).

Multimodal sensing is a promising strategy for improving robustness in animal monitoring. By combining complementary information from multiple modalities, fusion based approaches can provide a more complete representation of animal health and behavior than single modality systems. Previous studies have shown that multimodal sensor fusion can improve livestock behavior and health classification, including cattle behavior recognition from accelerometry and global navigation satellite

Received May 2, 2026.

Accepted August 17, 2026.

*Corresponding author: arnas.nakrosis@ktu.lt

The list of standard abbreviations for JDS is available at adsa.org/jds-abbreviations-26. Nonstandard abbreviations are available in the Notes.

system data, abnormality detection from heterogeneous sensor representations, and explainable machine learning for farm decision support (Aguilar-Lazcano et al., 2023; Arablouei et al., 2023; Jadhav et al., 2024; Shi et al., 2024; Ma et al., 2026).

Wearable sensors, ingestible rumen boluses, computer vision systems, and Internet of Things platforms have been investigated for cattle behavior and health monitoring. Wearable and embedded systems can classify animal activity using inertial or physiological signals, whereas computer vision methods can capture visual indicators such as posture, gait, body condition, and movement patterns (Gutierrez-Galan et al., 2018; Mahmud et al., 2021; Mao et al., 2023; Ding et al., 2025; Sohan et al., 2025). In dairy cattle, farm management and sensor data have supported classification of healthy and disordered cows, with rumination, milk yield, and milk electrical conductivity identified as important predictors (Zhou et al., 2022). Similar approaches have been applied for clinical mastitis prediction, weakly supervised mastitis detection, lameness prediction, vision based mobility assessment, and digital dermatitis detection or short-term prediction (Anagnostopoulos et al., 2023; Magana et al., 2023; Cho et al., 2024; Dervić et al., 2024; Liu et al., 2026). These studies demonstrate the value of combining production, behavioral, physiological, and visual information, although energy consumption, communication bandwidth, limited labeled data sets, and imperfect annotation remain barriers to broad deployment (Akhigbe et al., 2021).

Thermal imaging is a relevant noncontact modality because it captures surface temperature distribution patterns associated with inflammation, infection, and metabolic stress. Infrared thermography has been investigated for mastitis detection, where localized temperature increases in the udder region may reflect inflammatory responses, and for hoof disease or lameness detection, where limb temperature asymmetry or localized heat may indicate pathological changes (Xudong et al., 2020; Wang et al., 2022; Kim et al., 2025; Paulauskaite-Taraseviciene et al., 2026). Recent studies combining thermal image analysis with deep learning have also demonstrated the potential to identify relevant anatomical regions and improve robustness in temperature based monitoring systems (Bobić et al., 2025; Kim et al., 2025). However, thermal imaging alone may be insufficient in operational environments because ambient conditions, camera placement, animal movement, partial visibility, and occlusions can reduce measurement consistency.

Although previous studies demonstrate the potential of individual sensing modalities and multimodal fusion strategies (Ferreira et al., 2024; Li et al., 2024), several limitations remain (Deng et al., 2025; Asogan et al., 2026). Many systems rely either on handcrafted features

or fully end to end deep learning models, whereas fewer approaches combine interpretable engineered features with learned visual representations. In addition, many studies focus on one primary condition, one modality, or controlled acquisition settings. Under commercial farm conditions, a practical monitoring system must integrate heterogeneous data streams, tolerate missing measurements, and remain interpretable enough to support decision making. These gaps motivate hybrid multimodal architectures that combine thermal video analysis, milking system records, and rumen bolus measurements for dairy cattle health anomaly detection.

The objective of this study was to develop and evaluate a multimodal artificial intelligence framework for cow level health anomaly detection in dairy cattle under commercial farm conditions by fusing thermal video features, milking-system records, and rumen bolus sensor measurements. We hypothesized that integrating complementary sensor modalities would improve anomaly detection performance compared with reduced modality configurations and that structured production and physiological variables would provide the strongest predictive signal, while thermal imaging would contribute additional discriminative information when combined with these data sources.

MATERIALS AND METHODS

Multimodal Framework for Cattle Health Monitoring

This study proposes a multimodal methodology for automated cattle health-state monitoring by integrating thermal video, milking-system records, and rumen bolus measurements into a unified anomaly-detection framework. The workflow consists of parallel modality-specific processing branches, in which thermal-video features and farm-management sensor features are extracted independently before being combined in a fusion-based classification stage. The study design was based on the premise that dairy cow health anomalies may be reflected in multiple complementary data streams. Thermal imaging captures external surface temperature patterns, milking system records describe production and udder health related changes, and rumen bolus sensors provide internal physiological and activity related information. Integrating these modalities was therefore expected to provide a more complete representation of cow health status than any single data source alone, particularly under commercial farm conditions where individual sensor streams may be incomplete or variable. Thermal video data are first decoded and sampled at fixed temporal intervals to balance computational cost and temporal coverage. For each cow-milking event, 30 thermal frames were sampled at fixed temporal intervals

from the period in which the cow was visible in the camera field of view. The sampling interval was determined by dividing the visible milking session duration by 30, ensuring that the selected frames were distributed across the available recording period. Thus, each cow milking event contributed 30 thermal frames to the temporal analysis module. Each selected frame was converted to a grayscale thermal-intensity map, and pixel values were normalized to a common intensity scale before spatial resizing. For CNN-based thermal feature extraction, detected regions of interest were resized to 128×128 pixels and normalized using the ImageNet mean and standard deviation values to ensure consistent input dimensions and intensity distributions across frames. After preprocessing, thermally relevant body regions were identified using adaptive threshold-based region extraction. Specifically, for each preprocessed frame an image-specific threshold was calculated from the mean and standard deviation of pixel intensities using a threshold multiplier of $\alpha = 0.2$. Pixels exceeding this threshold were classified as candidate thermal regions, whereas lower-intensity background areas were suppressed. The upper 8% of each frame was masked to remove the thermal camera temperature scale bar, and bounding boxes were then extracted from the resulting binary masks for connected regions exceeding the minimum area criterion. A region of interest was considered valid when it represented a connected high intensity thermal region corresponding to the visible cow body area, remained outside the masked temperature scale region, and exceeded the minimum area criterion used to exclude small background artifacts, thermal noise, or irrelevant scene elements. These valid regions of interest were retained for downstream thermal feature extraction. For each extracted region of interest, statistical descriptors of thermal intensity distribution are computed to characterize spatial heat patterns on the animal's surface, including mean intensity, standard deviation, maximum and minimum intensity values, intensity percentiles, heat ratio, and thermal uniformity. Within the thermal branch, the standardized 30 frame sequence was used to extract sequence level descriptors describing dynamic changes in heat distribution.

In parallel with thermal analysis, structured physiological and production-related variables were collected from 2 farm-management data sources: milking system records and rumen bolus sensor measurements. For the milking system data, each cow milking event contributed one session level data point for each recorded variable, including somatic cell count, milk yield, electrical conductivity, blood presence indicators, and milk flow descriptors. Thus, each milking variable was represented once per cow per milking event. Bolus measurements were not recorded specifically at each milking event. Instead, they were available at 10-min and 1-d intervals and

included internal body temperature and activity-related indicators. Therefore, for each cow milking event, the temporally closest available bolus measurement within the predefined synchronization window was assigned to the corresponding thermal and milking record. If no bolus measurement was available for that cow within the synchronization window, the bolus modality was treated as missing for that observation. This procedure resulted in cow specific multimodal records at the cow milking event level.

Figure 1 explains the methodology of the proposed multimodal structure. Thermal video is processed through region detection followed by 3 feature extraction branches: temporal CNN (ResNet-34), statistical features, and spatial CNN (ResNet-50). In parallel, milking and bolus data are transformed into structured features. All features are concatenated into a unified vector and fed into a multimodal fusion classifier (MLP) for anomaly prediction. Before classification, features are normalized using training set statistics.

Data set Description

The experimental study relies on a multimodal data set collected from a commercial dairy farm in Lithuania during routine milking operations. The analyzed thermal recording campaign covered a 3 d study period from 2025-09-30 to 2025-10-02, during which thermal video data were collected and temporally matched with corresponding milking system records and available rumen bolus measurements. The study animals were Holstein-Friesian dairy cows raised under indoor commercial farm conditions in Kaunas, Lithuania. During the study period, cows were housed in a naturally ventilated open-sided barn and milked using DeLaval robotic milking system. The barn environment was not experimentally temperature controlled, reflecting routine commercial farm conditions. The data set consists of 3 primary modalities: thermal video recordings, milking system records, and physiological measurements obtained from rumen bolus sensors. These heterogeneous data sources are synchronized using timestamp-based alignment to associate each observed cow with its corresponding physiological measurements. The data set contains recordings from 88 cows monitored during routine milking sessions. Days in milk (DIM) were calculated as the number of days between each cow's most recent calving date and the thermal video acquisition start date. Among monitored cows, mean DIM was 180.0 ± 103.6 d, with a median of 189 d, an interquartile range of 100.3 to 242.0 d, and a range of 1 to 527 d. The model was trained to predict a binary same day cow health outcome, classified as normal or anomalous. During the observational period, mastitis and leg related disorders,

including lameness and leg infections, represented the predominant health conditions recorded in the veterinary records. These conditions were selected as the predefined target health categories because they were clinically relevant to the study objective and directly related to the available thermal field of view, which captured the udder and lower limb region. A small number of other health conditions, including metabolic or respiratory disorders, were occasionally recorded, but their occurrence was too infrequent to support reliable statistical analysis. Therefore, these cases were excluded from the outcome definition, and the anomalous class was restricted to veterinary recorded mastitis and leg related disorders. Bolus generated event indicators for non-target conditions were retained only as sensor derived predictor variables when available, not as diagnostic outcome labels. Observations without a veterinary recorded health event within the diagnostic window were labeled as normal. Cows were included when thermal recordings could be matched with cow identity and corresponding milking system records using timestamp-based alignment. In this context, farm records refer primarily to structured milking system records, including production and milk quality variables. Rumen bolus sensor measurements were linked as an additional physiological data source when available, but bolus data were not required for inclusion because not all cows were equipped with ingestible sensors during the study period. The use of real-world farm data introduces natural variability in animal posture, environmental conditions, and sensor availability. While such variability increases the complexity of the anomaly detection task,

it also provides a realistic representation of operational livestock monitoring environments. During the observational period bolus measurements were missing for 13 cows (15%) because not all animals were equipped with rumen bolus sensors during the study period.

Thermal Video Data. The thermal recordings were acquired using a camera (Model: Topdon TC001, Topdon Technology Co., Ltd., Shenzhen, China) installed in front of the milking robot at approximately 40 cm above ground level, oriented toward the udder and lower limb region of the cow. The camera field of view captured the left lateral caudal region of the cow, primarily including the udder, hind legs, and adjacent rear body surface, rather than the complete body surface. Because the camera was fixed on one side of the milking station, thermal information was obtained from one visible side of the animal, and simultaneous bilateral thermal imaging was not performed. Therefore, the extracted thermal features should be interpreted as region-specific thermal descriptors rather than whole body thermal measurements. Each video is stored using a timestamp based naming convention, which enables precise temporal alignment with other data sources. The thermal videos are captured during routine milking sessions and reflect natural variations in animal position, posture, and movement under real-world farm conditions. These recordings primarily provided information on heat distribution patterns across the visible animal body region. Thermal frames were displayed using a fixed temperature color scale, rather than a scale normalized separately to the minimum and maximum temperature of each frame. In addition to the fixed

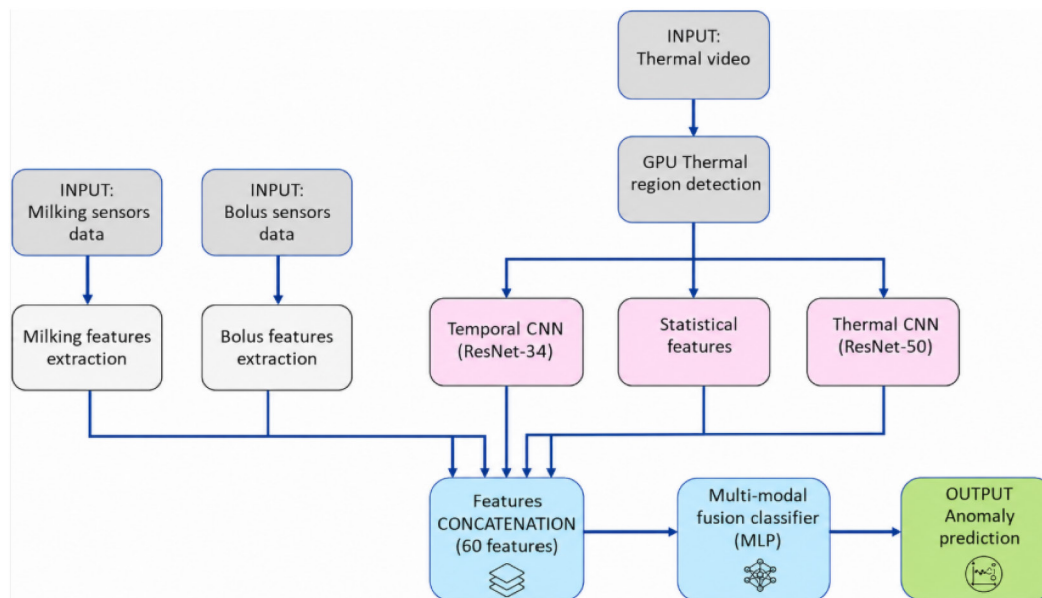


Figure 1. Workflow of the proposed multimodal anomaly-detection framework.

color scale, the camera interface displayed the minimum and maximum temperature measured within each individual frame. Thus, the color representation remained based on a fixed temperature scale, while the displayed minimum and maximum values described the temperature range observed within that specific frame. Because infrared thermal measurements can be influenced by ambient conditions inside the barn, the recordings were obtained from a fixed camera position within the same milking station and over a short acquisition period to minimize environmental variation. Ambient farm temperature was not experimentally controlled, reflecting real commercial farm conditions; therefore, frame-level normalization and region-based feature extraction were applied to reduce the influence of global intensity variation and to focus the analysis on relative thermal patterns associated with the animal body. Individual animals are not explicitly associated with identification labels in the recordings. Instead, the data consist of continuous thermal video sequences. Frames containing only background or irrelevant parts of the scene were excluded automatically during preprocessing. Specifically, each frame was processed using adaptive threshold-based region detection, and frames were retained only when at least one valid thermal region of interest corresponding to the animal body was detected. Frames without a detected region exceeding the predefined minimum area criterion were discarded to reduce background-related thermal noise. Feature extraction is carried out only on thermal features, specifically on characteristics related to temperature distribution and their variations over time in consecutive frames.

Figure 2 illustrates a representative thermal frame acquired during a milking session and provides an example

of the thermal intensity scale used in the recordings. During experiment period, 18 thermal videos, each one hour in duration, were recorded per day, resulting in 54 thermal videos included in the analysis. Each recording covered one to 6 milking sessions, depending on the frequency with which cows entered the milking station and the duration of individual milking sessions. Each milking session was identified using the corresponding milking system timestamp and cow identity information and was temporally matched with the thermal video recorded during the same period. This recording structure allowed multiple cow level milking observations to be extracted while maintaining relatively consistent recording conditions.

Milking Data. Milk was collected using a DeLaval automated milking robot (DeLaval Inc., Tumba, Sweden) installed on the commercial dairy farm. During each robotic milking session, the system automatically recorded production related and udder health associated measurements, including somatic cell count (SCC), milk yield, electrical conductivity, blood presence indicators, and milk flow rate. Somatic cell count is widely used as an indicator of udder health and mastitis risk, whereas variations in electrical conductivity can reflect changes in milk composition associated with inflammatory processes. These measurements provide complementary information to thermal imaging observations.

Bolus Sensor Data. SmaXtec Classic Bolus (Model: SCB-G5, SmaXtec Animal Care GmbH, Graz, Austria) sensors are ingestible devices placed inside the cow's rumen to monitor internal physiological conditions. The continuous bolus features included internal body temperature statistics, specifically mean, standard deviation, minimum, and maximum values; activity related

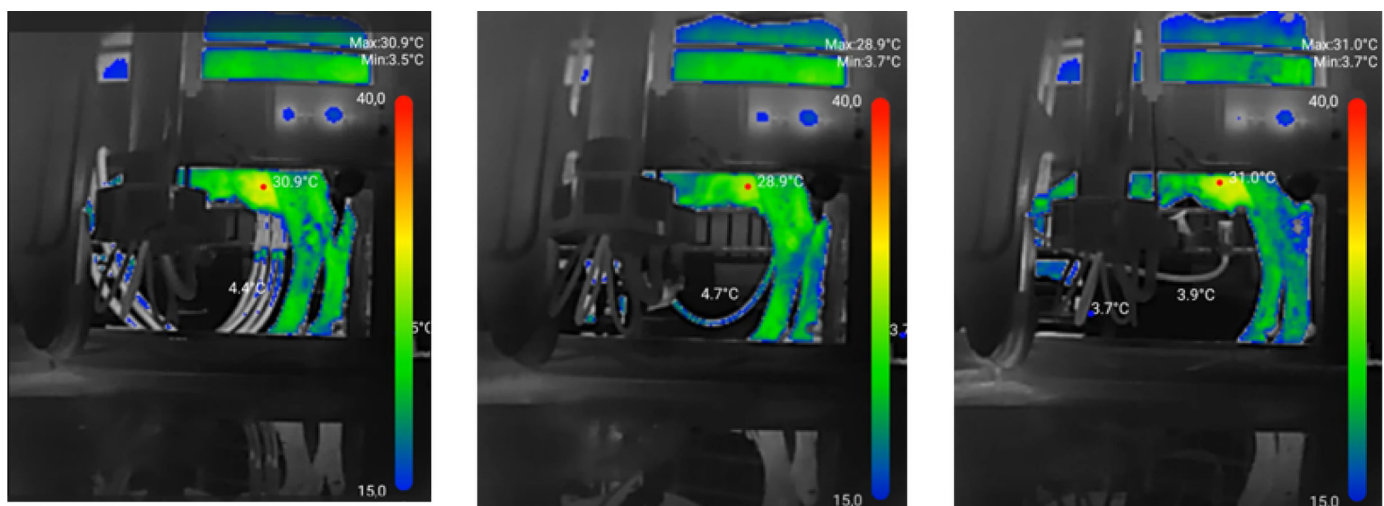


Figure 2. Sample thermal images showing temperature variations among milking cows.

summaries; temperature values excluding drink cycle effects; rumination summaries expressed as minutes per 24 h; and water intake. Bolus data provide physiological context that cannot be directly observed through external imaging. When combined with visual observations from thermal imaging and production indicators from milking systems, these measurements contribute to a multimodal representation of animal health.

To enable multimodal integration, bolus sensor measurements were temporally synchronized with the corresponding infrared (IR) video frames using a sliding window alignment strategy. Bolus data were available at 10-min and 1-d intervals, and each IR frame was matched with the temporally closest sensor measurement within a synchronization window of one day. When exact time-stamp matches were unavailable, the temporally closest sensor measurement within the defined window was selected to preserve temporal consistency. It is important to note that bolus data was only available for a subset of animals, as 75 of the 88 cows were equipped with ingestible sensors during the study period. Consequently, only a limited set of bolus features was incorporated, resulting in a heterogeneous yet realistic multimodal data set that reflects real-world farm deployment conditions.

Hybrid AI Model Architecture

The proposed hybrid AI architecture combines handcrafted thermal statistics with deep neural network representations and farm management sensor data. This design enables the system to capture both interpretable thermal patterns and complex spatiotemporal representations learned from thermal video.

The proposed architecture introduces several methodological contributions that distinguish it from existing cattle monitoring systems. Unlike conventional approaches that rely solely on either handcrafted features or end-to-end deep learning, our system employs a dual-pathway feature extraction design that combines interpretable statistical features with deep CNN representations in parallel. This enables both accurate predictions and interpretable intermediate features suitable for agricultural decision-support. While existing thermal imaging systems for livestock predominantly analyze individual frames independently (Hajnal et al., 2022; Wang et al., 2022), our architecture incorporates temporal modeling through a Transformer-based aggregation mechanism that processes 30-frame sequences to capture dynamic thermal patterns associated with disease progression. In addition, the fusion classifier integrates fundamentally heterogeneous data modalities, including continuous thermal video streams, discrete milking event records, and periodic bolus sensor measurements, thereby improv-

ing resilience to missing modalities in real-world farm deployments (Gómez et al., 2021; Jiang et al., 2023).

Thermal Region Detection Module. The first processing stage identifies thermal regions corresponding to the animal body. Video frames are converted to grayscale intensity maps and analyzed using adaptive thresholding. This procedure highlights regions with elevated temperature values while suppressing background elements. In the implemented system, adaptive thresholding uses image-specific statistics, including the mean and standard deviation of pixel intensities. Regions exceeding the defined threshold are classified as candidate thermal regions of interest.

$$T_i = \mu_i + \alpha \times \sigma_i, \quad (1)$$

where μ_i and σ_i represent the mean and standard deviation of frame i , respectively, and $\alpha = 0.2$ is the threshold multiplier.

The top 8% of each frame is masked to exclude the temperature scale bar present in thermal camera output. Bounding boxes are extracted from the resulting binary masks for regions exceeding a minimum area threshold.

Statistical Thermal Feature Extraction. For each detected thermal region, a set of statistical descriptors is computed to characterize the spatial distribution of temperature intensity. Let R denote the region of interest extracted from the grayscale intensity map. The following features are calculated:

Intensity Statistics:

- Mean intensity: $\mu_R = \left(\frac{1}{|R|} \right) \sum R(x, y)$;
- Standard deviation: σ_R ;
- Maximum and minimum intensity values
- Percentiles: P_{25} , P_{50} (*median*), P_{75} , P_{95} ;

Thermal Indicators:

$$\text{Heat ratio: } h = \frac{\mu_{red}}{(\mu_{blue} + \varepsilon)} ; \text{ Uniformity: } u = \frac{1}{(\sigma_R + 0.01)},$$

$$\text{where } \varepsilon = 10^{-6}$$

μ_{red} is the mean intensity of the red channel within the detected region of interest, μ_{blue} is the mean intensity of the blue channel within the same region, and ε is a small positive constant added to avoid division by zero. In the thermal color representation, higher red channel intensity corresponds to warmer regions, whereas higher blue channel intensity corresponds to cooler regions. There-

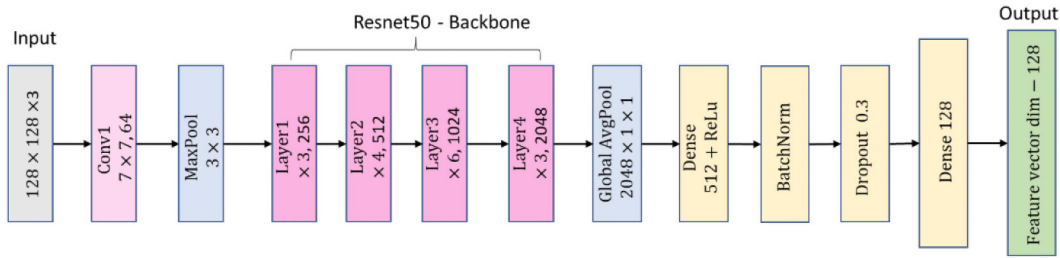


Figure 3. Thermal Feature Extractor CNN architecture

fore, larger h values indicate a greater relative dominance of warm thermal areas within the region of interest.

The P_{95} percentile specifically captures hotspot regions that may indicate inflammation or infection. The heat ratio leverages the color mapping of thermal cameras where warmer regions appear red and cooler regions appear blue.

CNN-Based Visual Feature Extraction

The first CNN component, designated the Thermal Feature Extractor, employs a 50-layer Residual Network (ResNet-50) backbone pretrained on ImageNet for transfer learning. The selection of ResNet-50 as the backbone architecture was motivated by several key considerations. First, ResNet's residual connections effectively address the vanishing gradient problem, enabling training of deeper networks that can learn more complex hierarchical feature representations (He et al., 2016). Second, ResNet-50 provides an optimal balance between model capacity and computational efficiency; deeper variants such as ResNet-101 or ResNet-152 offer marginal accuracy improvements but substantially increase inference time, which is critical for real-time video processing applications. ResNet architectures have demonstrated strong generalization capabilities across diverse visual domains, including medical imaging and thermal analysis tasks, making them well-suited for transfer learning to our thermal cow imagery domain where labeled training data is limited (Pan and Yang, 2010). Therefore, the extensive availability of pretrained ImageNet weights for ResNet models facilitates effective knowledge transfer, allowing the network to leverage low-level feature detectors (edges, textures, shapes) learned from millions of natural images.

The final fully-connected classification layer and average pooling layer of ResNet-50 are removed, retaining only the convolutional feature extraction layers. A custom feature head is appended (see Figure 3).

The early layers of the backbone are frozen to preserve low-level representations learned from ImageNet while allowing higher-level layers to be fine-tuned for

the target domain. Input Region of Interests (ROI) are resized to 128×128 pixels and normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). This module outputs a 128-dimensional feature vector for each detected thermal region, capturing abstract visual patterns that may correlate with health anomalies.

Temporal Analysis CNN

Figure 4 illustrates the temporal thermal analysis module used in the proposed system. The module is designed to capture spatiotemporal patterns in thermal video sequences to detect subtle physiological changes associated with potential health anomalies.

The input to the module consists of a sequence of thermal frames represented as a tensor of shape (N, T, H, W) , where N denotes the batch size, T represents the number of frames in the temporal sequence, and H and W correspond to the spatial resolution of each frame. The frames are first reshaped into a format suitable for convolutional processing, enabling per-frame feature extraction using a deep convolutional backbone.

Feature extraction is performed using a ResNet-34 frame encoder, which processes each frame independently while sharing weights across the sequence. The encoder begins with a convolutional layer using a 7×7 kernel with 64 filters, followed by a series of residual blocks organized into 4 stages with progressively increasing channel depth. Specifically, the residual layers contain 64, 128, 256, and 512 feature channels, respectively. Each stage captures increasingly abstract visual representations of thermal patterns such as edges, texture distributions, and localized temperature variations. After the final convolutional stage, the feature maps are transformed through global average pooling, producing a compact 512-dimensional feature vector for each frame in the sequence.

The extracted frame-level representations are then reshaped into a temporal feature matrix of size $(N, T, 512)$. To preserve information about frame ordering, positional encodings are added to the feature vectors. These encod-

ings allow the model to distinguish between different temporal positions within the sequence and enable effective temporal reasoning during sequence processing.

Temporal relationships between frames are modeled using a Transformer encoder composed of 4 stacked layers. Each Transformer layer applies multi-head self-attention mechanisms that allow the model to learn dependencies between frames regardless of their relative distance in the sequence. This mechanism enables the network to capture gradual temperature changes, motion patterns, and other temporal dynamics that may reflect developing health conditions.

Following temporal aggregation, the sequence features are mean-pooled across the temporal dimension, producing a single 512-dimensional representation that summarizes the observed thermal dynamics of the sequence.

The pooled representation is then processed by 3 specialized prediction heads that perform complementary tasks within the monitoring system:

- (1) *Health Score Head* – estimates a continuous health score using a fully connected network composed of dense layers with ReLU activation followed by a sigmoid output.
- (2) *Anomaly Detection Head* – produces binary classification logits indicating the probability of an anomalous health state.
- (3) *Temperature Trend Head* – predicts statistical descriptors of thermal behavior, including mean temperature, standard deviation, and temperature trend.

Each prediction head consists of fully connected layers combined with dropout regularization to improve generalization and reduce overfitting. The outputs of these heads provide complementary information about the animal's health condition, allowing the downstream fusion classifier to incorporate both anomaly predictions and interpretable thermal indicators.

The Temporal Analysis CNN was trained using supervised learning within the same cow level cross validation framework used for the downstream classifiers. The anomaly detection head was trained using the binary anomaly label as the target outcome, whereas the health score head was trained as an auxiliary supervised output derived from the same health status annotation. To avoid data leakage, the temporal model was trained separately within each training fold, and cows assigned to the validation fold were not used during CNN training, model selection, or feature normalization. CNN derived outputs used in the multimodal fusion classifier, including *health_score* and *anomaly_probability*, were therefore generated for validation animals only by temporal models trained on the corresponding training-fold animals. These outputs were treated as out of fold thermal features and were not generated by a model trained on the same validation observations.

Overall, the architecture enables the system to analyze both spatial thermal patterns and their temporal evolution, providing a robust mechanism for identifying early deviations in thermal behavior associated with potential livestock health anomalies.

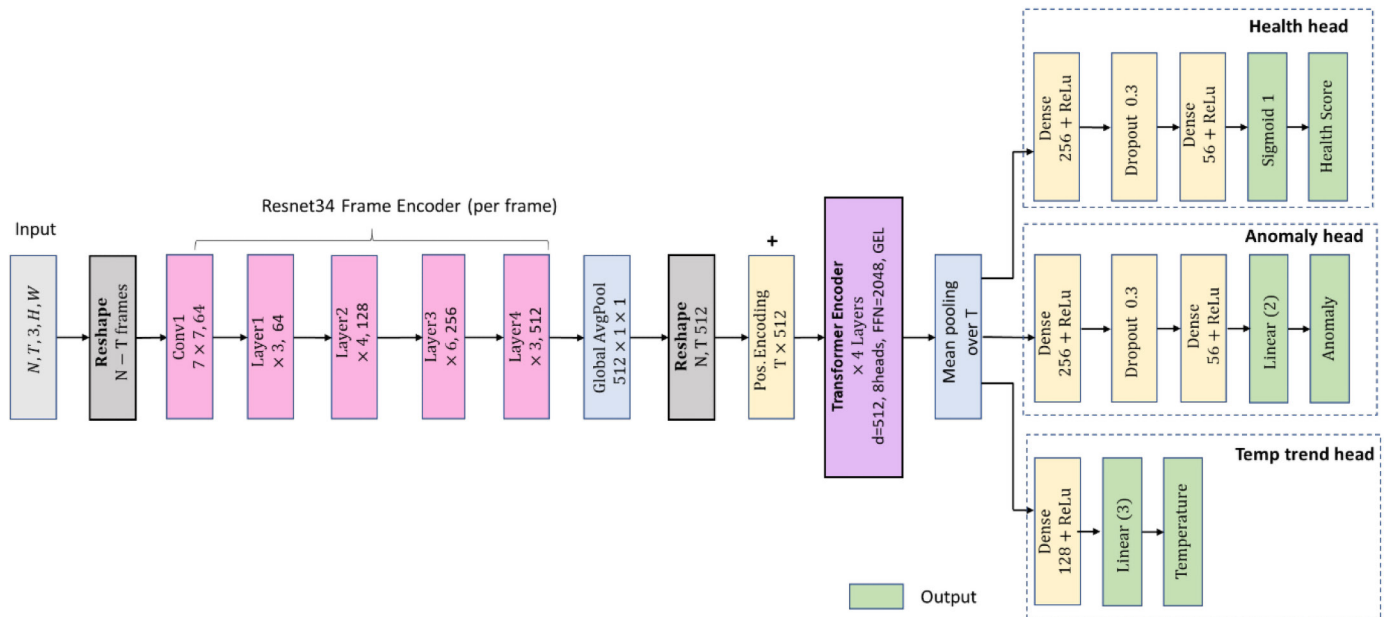


Figure 4. Transformer Encoder architecture

Multimodal Fusion Classifier

The multimodal fusion classifier represents the core integration mechanism of the proposed hybrid architecture. Depending on the modality configuration being evaluated, it synthesizes thermal statistical features, CNN derived thermal features, milking system features, and available rumen bolus sensor features into a unified anomaly prediction framework (see Figure 5).

The classifier receives a concatenated feature vector comprising 60 features organized into 4 distinct groups.

Thermal Statistical Features (10) *intensity_mean*, *intensity_std*, *intensity_max*, *intensity_min*, *intensity_p25*, *intensity_p50*, *intensity_p75*, *intensity_p95*, *heat_ratio*, *uniformity*. Basic statistics (mean, standard deviation, min, max) capture central tendency and spread of thermal intensity within detected regions, while the 4 percentiles (P25, P50, P75, P95) characterize the full distribution shape without assuming normality. The P5 percentile was not included because the lower tail of the thermal intensity distribution was expected to be more strongly affected by residual background pixels, boundary effects, shadowed areas, or cooler non informative regions within the detected region of interest. In contrast, the minimum value and P25 already provided information on the lower intensity range. The P95 percentile specifically targets hotspot detection indicative of localized inflammation, as mastitis-affected tissue produces elevated temperatures in the upper tail of the distribution while potentially leaving mean values within normal ranges. The heat ratio exploits the thermal camera's color mapping to provide an additional temperature indicator, while uniformity captures spatial homogeneity-healthy tissue typically exhibits more uniform temperature distributions than inflamed regions.

CNN-Extracted Features (2) *health_score*, *anomaly_probability*. These features represent high-level abstractions learned by the ResNet-34 backbone and Transformer encoder from 30-frame temporal sequences, capturing dynamic thermal patterns and subtle visual anomalies that may escape statistical characterization. To prevent data leakage, these CNN derived features were generated using the fold specific Temporal Analysis CNN: for each validation fold, the temporal model was trained only on

training fold animals and then applied to validation fold animals to obtain out of fold *health_score* and *anomaly_probability* values for the multimodal fusion classifier. Unlike the handcrafted thermal features, these CNN outputs encode nonlinear relationships learned directly from the data, potentially identifying complex spatiotemporal signatures associated with disease onset. The selection of these 2 specific outputs rather than the full 512-dimensional Transformer representation reflects the architecture of the Temporal Analysis CNN, which compresses temporal features through task-specific output heads: a regression head producing *health_score* $\in [0,1]$ and a binary classification head producing *anomaly_probability*. This design ensures that the features passed to the fusion classifier are already aligned with the health assessment objective, reducing the risk of overfitting that would arise from incorporating high-dimensional intermediate representations given the limited labeled training data available.

Milking Sensor Features (22) The session level milking features were: *scc_mean*, *scc_max*, *scc_elevated* (binary), *milk_yield_mean*, *blood_detected* (binary), *blood_max_ppm*, *conductivity_mean*, *conductivity_max*, *flow_mean*, *flow_max*. The quarter level features were electrical conductivity, maximum milk flow, and milk yield for each udder quarter: *conductivity_dp*, *conductivity_du*, *conductivity_kp*, *conductivity_ku*, *flow_max_dp*, *flow_max_du*, *flow_max_kp*, *flow_max_ku*, *milk_yield_dp*, *milk_yield_du*, *milk_yield_kp*, and *milk_yield_ku*. These features represent established clinical biomarkers for bovine health monitoring. The *scc_elevated* indicator uses the internationally recognized 200,000 cells/mL threshold for subclinical mastitis detection (Ramuada et al., 2024). Blood detection and conductivity measurements provide complementary indicators of udder health, as conductivity changes reflect alterations in milk ion concentration associated with inflammation. Milk yield and flow rate capture production anomalies that often accompany health issues.

Bolus Sensor Features (26) The bolus modality contributes 17 continuous physiological summaries and 9 binary event indicators or event count variables. The continuous bolus features consisted of 4 physiological variable groups: body temperature, activity, tempera-

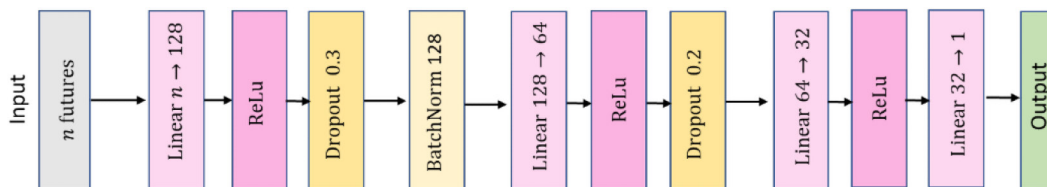


Figure 5. Multi Modal Fusion Classifier architecture

ture excluding drink cycle effects, and rumination, each represented by mean, standard deviation, minimum, and maximum values. Water intake was included as an additional continuous bolus feature. The event-based bolus features were: *bolus_heat_event*, *bolus_mastitis_event*, *bolus_metabolic_disorder_event*, *bolus_respiratory_issue_event*, *bolus_lameness_event*, *bolus_rumination_abnormality_event*, *bolus_calving_event*, *bolus_fever_event*, and *bolus_total_event_count*. Near constant or overly sparse fields were excluded to avoid noise and noninformative predictors.

Feature normalization is applied using statistics computed from the training set to ensure that all input features operate on comparable scales. This z-score normalization is important for stable optimization because the raw feature magnitudes span substantially different ranges, such as somatic cell counts in the hundreds of thousands and normalized intensity values between 0 and 1. To address class imbalance inherent in health monitoring data sets where healthy observations substantially outnumber anomaly cases, the training objective employs weighted Binary Cross Entropy loss:

$$L = -\frac{1}{N} \sum_{i=1}^N \left[w_{pos} \cdot y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \right], (2)$$

where $w_{pos} = n_{negative}/n_{positive}$ inversely weights the positive class contribution to compensate for underrepresentation in the data set. When class imbalance is severe (anomaly ratio below 35%), focal loss is applied instead, which introduces an additional modulating factor $(1 - p_i)^\gamma$ to further down-weight easily classified samples and concentrate training on hard examples near the decision boundary. The Adam optimizer with learning rate 0.001 and weight decay 10^{-4} is employed, with ReduceLROnPlateau scheduling (patience = 15, factor = 0.5) to reduce the learning rate when validation loss plateaus. For classical machine learning models, an equivalent balanced class weighting scheme is used during training to achieve the same effect.

The multi-modal fusion strategy offers several advantages over single-modality approaches. For instance, the fusion strategy provides resilience to missing or noisy data: if thermal imaging is compromised by environmental conditions, milking sensor features can still contribute to the prediction. The fusion network can learn complementary relationships, such as correlations between elevated thermal intensity and increased somatic cell count that individually may fall below detection thresholds. Moreover, the inclusion of both handcrafted statistical features and learned CNN representations combines the interpretability of traditional feature engineering with the representational power of deep learning.

Classification Back-End Models

To identify the most effective classification approach for the multimodal tabular feature space, the proposed framework evaluates a diverse set of classification models spanning both neural network and classical machine learning families. All models receive the same fused feature vector as input and are trained under identical data-splitting and augmentation conditions, allowing direct comparison of model suitability.

Four neural network architectures were evaluated, each designed as a fully connected classifier operating on the fused tabular input:

- (1) Basic MLP (MLP) - a 3-layer multilayer perceptron with batch normalization and dropout regularization, serving as a lightweight neural baseline for the fusion task.
- (2) Deep MLP - an extended multilayer perceptron with 4 hidden layers and batch normalization at each level, intended to test whether additional depth improves learning on the available data.
- (3) ResNet-Style MLP - a fully connected classifier with residual connections that allow gradient flow through skip connections, reducing the risk of degradation in deeper layers.
- (4) Ensemble MLP - a collection of 5 lightweight subnetworks whose predictions were averaged to reduce variance and improve robustness. The 5 subnetworks used first hidden-layer widths of 64, 80, 96, 112, and 128 neurons, respectively, whereas the second hidden layer was fixed at 32 neurons in all subnetworks.

All neural network models are trained using the Adam optimizer with a learning rate of 0.001 and weight decay regularization. For class-imbalanced configurations, Focal Loss is applied to down-weight easily classified samples and focus training on hard examples near the decision boundary. In addition, 3 classical machine learning classifiers were evaluated as alternative classification back-ends:

- (1) Logistic Regression (LR) - a linear classifier with L2 regularization.
- (2) Random Forest (RF) - an ensemble of 100 decision trees with balanced class weights
- (3) Gradient Boosting (GB) - a sequential ensemble of 100 weak learners.

The classical models use standard scaling (z-score normalization) fitted on the training set and applied consistently to validation data. Class weighting is used across

all classical models to compensate for the imbalanced distribution of anomaly labels in the data set.

The inclusion of classical machine learning models alongside neural networks was motivated by the characteristics of the data set. With a relatively small number of labeled samples and a high-dimensional but structured tabular feature space, tree-based and linear models were expected to offer competitive or superior performance compared with deeper architectures, as they are less prone to overfitting on limited data. Given the relatively small number of labeled samples and the structured tabular feature space, tree-based and linear models were included to evaluate whether less parameterized classifiers could provide robust alternatives to deeper neural network architectures.

Feature importance was calculated after model training to support interpretation of the selected classifiers. For tree-based models, including RF and GB, feature importance was estimated using the impurity-based importance score, calculated as the normalized total reduction in node impurity attributable to each feature across all trees in the model. For LR, feature importance was assessed using the absolute values of the standardized model coefficients, allowing comparison of predictor contributions after z-score normalization. Feature importance values were used only for model interpretation and were not used for feature selection or model training.

Experimental Design and Evaluation Strategy

The experimental evaluation was designed to systematically assess the performance of the proposed multimodal anomaly detection framework under multiple sensing-modality configurations. The primary goal of the evaluation was to determine how different combinations of sensing modalities contribute to anomaly detection performance and to analyze whether the integration of thermal video information improves predictive capability when combined with physiological and production-related sensor data.

The experimental data set contains synchronized thermal video recordings, milking system measurements, and rumen bolus sensor data collected from 88 cows during routine milking sessions. The final analytic unit was a cow milking event level observation. Each observation represented one cow specific record in which thermal video features, milking system variables, and available rumen bolus measurements were aligned using cow identity and timestamp proximity. Thus, the 838 observations do not represent raw thermal frames or fixed cow level labels, but repeated cow milking event level records. Across the 88 monitored cows, each cow contributed on average 9.5 observations to the final analytic data set, corresponding to approximately 3.2 observations per

cow per day during the 3 d thermal recording period. The number of observations per cow varied according to cow entry frequency into the robotic milking station, video availability, and successful temporal matching with milking records. These heterogeneous data sources provide complementary information about the animal's physiological state and behavior. Thermal imaging captures spatial temperature distribution patterns on the animal's body surface, while milking system data provide indicators of udder health and milk production dynamics. Bolus sensors, on the other hand, measure internal physiological signals such as body temperature and activity-related metrics. As described in the Data set Description section, the binary outcome was same day cow health status, classified as normal or anomalous according to veterinary diagnosis records recorded on the same calendar day as the cow milking observation. All cows in the herd were checked daily as part of routine farm health management. Veterinary examination and health condition recording were performed independently of the sensor outputs used in this study. Outputs from the milking robot or rumen bolus system did not determine which cows were examined and were not used to assign the outcome labels. Therefore, the predictor variables derived from milking system records and bolus sensors did not influence the veterinary diagnosis process, reducing the risk of circularity between sensor-based predictors and the outcome label. Veterinary health records were updated during the disease period rather than being limited only to the first day of diagnosis. Cows diagnosed with one of the predefined health conditions were checked again on subsequent days, and if the condition persisted, the corresponding cow milking observations on those days were also labeled as anomalous. Therefore, the anomalous label represented an active veterinary recorded mastitis or leg related disorder within the same calendar day as the sensor observation, including both initial diagnosis and continued disease status during treatment or recovery. During the observational period, 30 cows produced at least one mastitis record, and 27 cows produced at least one leg condition record, corresponding to 34.1% and 30.7% of the 88 monitored cows, respectively. Because health status was assigned at the observation level rather than as a fixed cow level label, individual cows could contribute both normal and anomalous observations during the study period. Therefore, class distribution was summarized primarily at the observation level. In the final analytic data set, 316 of 838 observations (38%) were labeled as anomalous and 522 of 838 observations (62%) were labeled as normal. At the cow level, 31 cows contributed only normal observations, whereas 57 cows contributed at least one anomalous observation during the predefined diagnostic window. Because the veterinary diagnosis records were matched to sensor ob-

servations at the same calendar day level, the temporal order between the sensor measurement and the veterinary assessment could not always be established within that day. Therefore, the analysis should be interpreted as same day anomaly detection rather than confirmed prospective early prediction. Because predictor variables were measured at the cow milking event level, whereas veterinary health status was assigned at the same calendar day level, multiple observations from the same cow on the same day could share the same health label while having different sensor feature values. This structure was used to retain within day variation in milk production, thermal distribution, and physiological measurements, because these parameters can change between milking events even when the cow's active disease status remains the same. However, this difference in temporal resolution should be considered when interpreting the results. Event level observations from the same cow and day are not fully independent, and cows with more eligible milking events may contribute more observations to the data set. To reduce the risk of overly optimistic evaluation caused by repeated measurements from the same animal, model validation was performed using cow level cross validation, ensuring that observations from the same cow were not included in both training and validation sets.

To analyze the contribution of each sensing modality, 4 modality configurations were evaluated (see Table 1)

These modality configurations were selected to simulate different deployment conditions in real farm environments. In practice, certain sensing modalities may occasionally become unavailable due to sensor failures, communication interruptions, or maintenance operations. Evaluating multiple modality configurations therefore provides insight into the robustness of the proposed monitoring system when operating with partial sensor availability.

A set of machine learning classifiers was evaluated for anomaly detection, including RF, LR, GB, and a basic MLP. These models were selected because they represent different families of machine learning algorithms with varying levels of model complexity and interpretability. Model training was performed using supervised learning, where each sample was labeled as healthy or anomalous based on veterinary diagnosis records. To avoid data

leakage between training and evaluation stages, the data set was split at the animal level rather than at the individual sample level. This ensures that observations from the same animal do not appear in both the training and validation sets.

Performance evaluation was conducted using several measures commonly used in anomaly detection tasks. These include accuracy, precision, recall, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and precision-recall area under the curve (PR-AUC). Because the data set was class-imbalanced, with 316 of 838 observations classified as anomalous and 522 of 838 observations classified as normal, accuracy was reported only as a supplementary metric. Model comparison and interpretation emphasized F1-score, recall, ROC-AUC, and PR-AUC, which are more informative than accuracy when the minority class is of primary clinical interest. Moreover, model performance was evaluated using 5-fold cross-validation. To prevent data leakage between training and validation sets, the data were split at the cow level using a cross-validation strategy (StratifiedGroupKFold) so that samples from the same cow could not appear in both subsets simultaneously.

RESULTS AND DISCUSSION

Benchmark Comparison Across Modality Configurations

The benchmark comparison across the evaluated modality configurations is presented in Table 2. This table summarizes the best cross validation performance achieved within each modality configuration for accuracy, F1-score, and ROC-AUC values. Accuracy and F1 score were calculated using the default decision threshold of 0.5, whereas ROC AUC was calculated from predicted probabilities and is therefore threshold independent.

The findings indicate that the choice of sensing modalities had a substantial impact on anomaly detection performance. Among the evaluated configurations, C3 (milking system data + bolus sensors) achieved the strongest threshold dependent classification performance at the default decision threshold. This configuration yielded the highest F1-scores, with RF reaching an average F1-score of 0.838, indicating a good balance between precision and recall. Within the C3 configuration, GB achieved the highest ROC AUC among the C3 classifiers, with a cross validation ROC AUC of 0.915. This indicates strong discriminative capability, meaning that the model was able to rank anomalous observations higher than normal observations over the full range of possible probability cutoff values used to convert predicted probabilities into binary classifications.

Table 1. Sensor-modality combinations evaluated in the anomaly detection experiments

No.	Modality configurations
C1	Thermal video + milking system data
C2	Thermal video + milking system data + bolus sensor data (full multimodal fusion)
C3	Milking system data + bolus sensor data
C4	Thermal video + bolus sensor data

Table 2. Benchmark summary across the four modality configurations

Configuration	Cross-validation accuracy	Cross-validation F1-score	Cross-validation ROC-AUC
C1	RF, 0.913 ± 0.078	RF, 0.396 ± 0.452	LR, 0.922 ± 0.063
C2	RF, 0.903 ± 0.076	LR, 0.712 ± 0.116	LR, 0.927 ± 0.011
C3	RF, 0.888 ± 0.108	RF, 0.838 ± 0.147	GB, 0.915 ± 0.101
C4	RF, 0.881 ± 0.067	RF, 0.374 ± 0.402	RF, 0.891 ± 0.034

Note. Values represent mean $\pm$ standard deviation across cross validation folds.

C1 = thermal video and milking system data;.

C2 = thermal video, milking system data, and rumen bolus sensor data;.

C3 = milking system data and rumen bolus sensor data;.

C4 = thermal video and rumen bolus sensor data;.

LR = Logistic Regression;.

RF = Random Forest;.

GB = Gradient Boosting;.

MLP = multilayer perceptron;.

ROC AUC = receiver operating characteristic area under the curve.

In contrast, C2 (full multimodal configuration) showed the best ranking performance overall, with LR achieving the highest ROC-AUC (0.927). This suggests that thermal imaging provides complementary information that improves class separability, even if it does not directly translate into higher F1-scores.

C1 (thermal video + milking system data) achieved high accuracy. However, its relatively low F1-scores indicate limitations in detecting minority anomaly cases. Finally, C4 (thermal video + bolus data) produced the weakest results overall, highlighting the importance of milking system features for reliable anomaly detection.

Based on the benchmark comparison across all 4 modality configurations, the subsequent detailed analysis focused on C2 and C3 because these 2 configurations provided the most informative performance profiles. In this context, most informative means that the configurations either achieved the strongest overall ranking performance or the strongest threshold dependent classification performance. C2, representing the full multimodal configuration, showed the strongest ranking performance, as reflected by the highest ROC AUC values. C3, combining milking system data and rumen bolus sensor data, achieved the strongest threshold dependent classification performance at the default decision threshold, as reflected by the highest F1 scores. Therefore, C2 and C3 were selected for more detailed evaluation of ROC curves, classification measures, optimized thresholds, and confusion matrices.

Figure 6 compares the detailed cross validation performance of these 2 selected modality configurations using 5-fold cross-validation. Subfigures (a) and (b) show the ROC curves for C2 (thermal + milking + bolus) and C3 (milking + bolus), respectively, while subfigures (c) and (d) summarize cross-validation metrics with standard deviation.

In Figure 6(a), all models exhibit strong discrimination performance, with ROC-AUC values above 0.88. The LR achieves the highest ROC-AUC (0.927), followed by Random Forest (0.916) and Gradient Boosting (0.904), while MLP reaches 0.883. The curves indicate consistent separation between classes across thresholds, confirming that the combination of all modalities supports effective ranking performance.

In Figure 6(b), performance under C3 shows greater variability. GB achieves the highest ROC-AUC (0.915), while RF (0.896) and LR (0.786) show reduced performance. The aggregated results in Figure 6(c) and (d) highlight differences in threshold-dependent behavior. In this analysis, the default decision threshold was 0.5, meaning that observations with a predicted anomaly probability of 0.5 or higher were classified as anomalous, whereas observations with predicted probabilities below 0.5 were classified as normal. For C2, models achieve high ROC-AUC and accuracy but comparatively lower F1-scores and recall, indicating conservative predictions at the default threshold. In contrast, C3 produces more balanced F1-scores, particularly for RF and GB, suggesting stronger direct classification capability. Overall, the figure shows that C2 provides superior ranking performance, while C3 yields better threshold-based classification results, with tree-based models consistently demonstrating the most stable performance across both configurations.

In addition to analyzing sensing modality configurations, the performance of different classifier architectures was evaluated. Figure 7 presents a comparison of classification measures across models for the 2 most informative modality settings (C2 and C3).

The results show that classifier performance varied across modality configurations and evaluation measures. In C2 (Figure 7a), LR and RF achieved the strongest overall performance among the evaluated models. RF

showed the highest accuracy and recall, whereas LR achieved a comparable accuracy and slightly higher F1 score, indicating that both models were well suited to the full multimodal feature set. Neural network-based approaches, including Deep MLP and Basic MLP, also achieved competitive accuracy, but their lower F1 scores indicate weaker detection of minority anomaly cases. In C3 (Figure 7b), GB achieved the highest precision and F1 score, indicating the strongest threshold-based classification performance for the milking plus bolus configuration. However, LR remained comparable to RF, although slightly lower across most measures.

Overall, these results indicate that tree-based models were competitive, particularly GB in C3, but LR also provided strong performance, especially in the full multimodal C2 configuration. In contrast, neural network models appear more sensitive to class imbalance and require additional tuning to achieve comparable performance.

Table 3 presents the validation performance of the selected models under configurations C2 and C3, highlighting the effects of threshold optimization. For C2, both Logistic Regression and Random Forest achieved strong ranking performance (ROC-AUC up to 0.934 and PR-AUC up to 0.924). Their best F1-scores are obtained at very low thresholds (0.030 and 0.175), respectively. This indicates that the default decision threshold of 0.5 was too conservative for these models, because lower thresholds were required to improve recall and maximize F1 score. For C3, both Random Forest (RF) and Gradient Boosting (GB) showed competitive performance, although their behavior differed. RF achieved a best F1-score of 0.746 at a threshold of 0.150, together with a ROC-AUC of 0.876 and a PR-AUC of 0.719, indicating stable overall discrimination. In contrast, GB achieved a higher best F1-score of 0.874 at a threshold of 0.370, while maintaining a comparable ROC-AUC of 0.874 and a stronger PR-AUC of 0.815. These results suggest that,

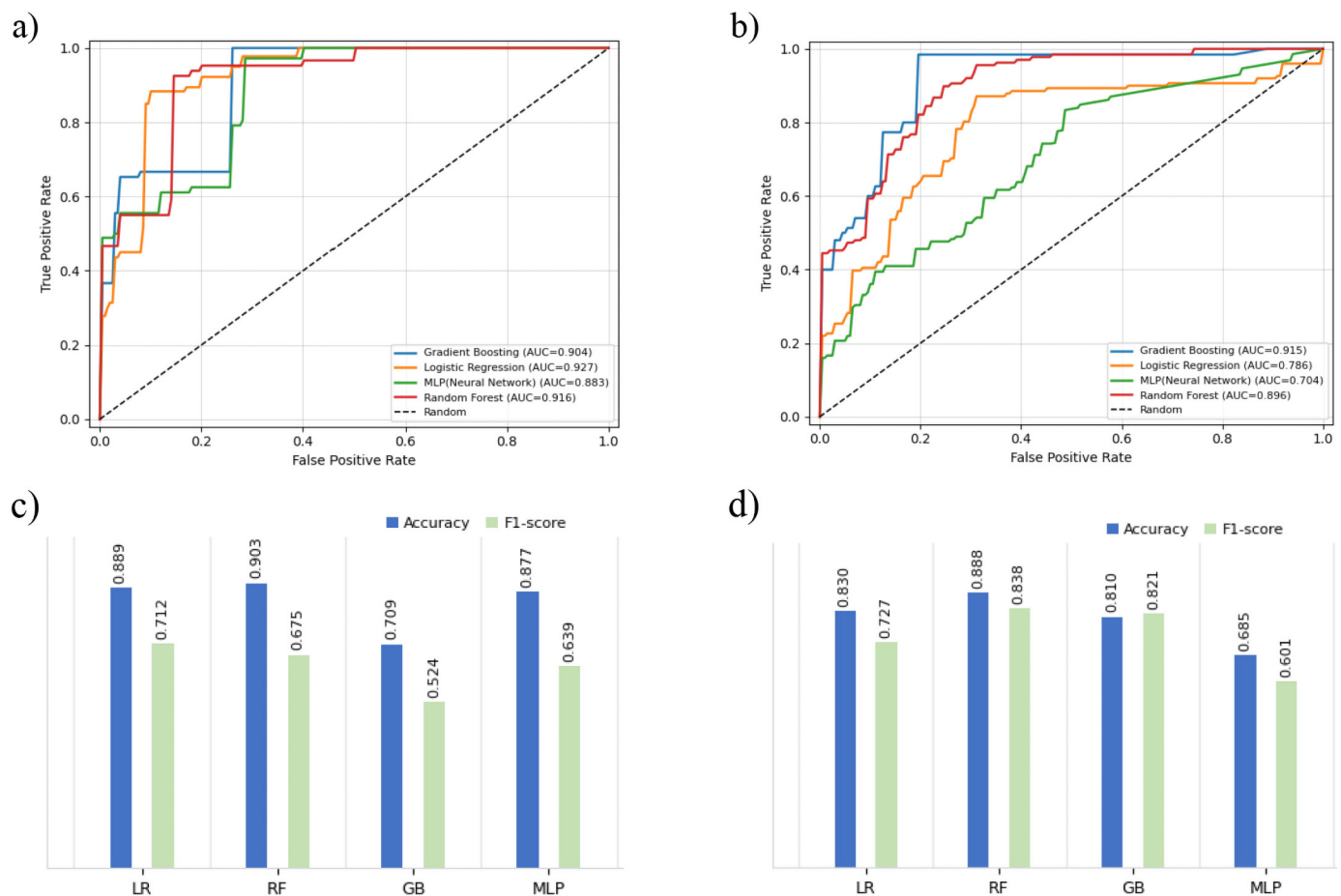
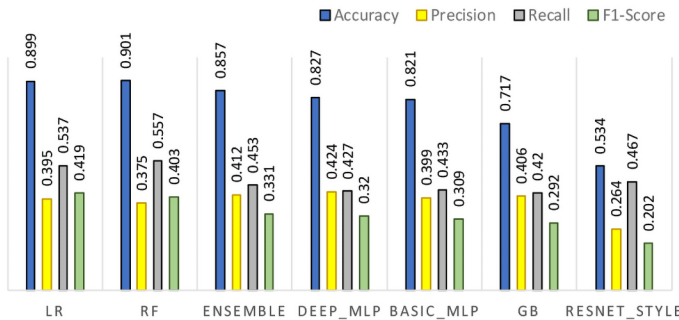


Figure 6. Cross-validation benchmark comparison for 2 most informative modality settings. ROC curve comparison for a) C2 and b) C3; Cross-validation performance for c) C2 and d) C3. **Note.** C2 = thermal video, milking system data, and rumen bolus sensor data; C3 = milking system data and rumen bolus sensor data; LR = Logistic Regression; RF = Random Forest; GB = Gradient Boosting; MLP = multilayer perceptron; ROC AUC = receiver operating characteristic area under the curve.

a) performance results with C2



b) performance results with C3

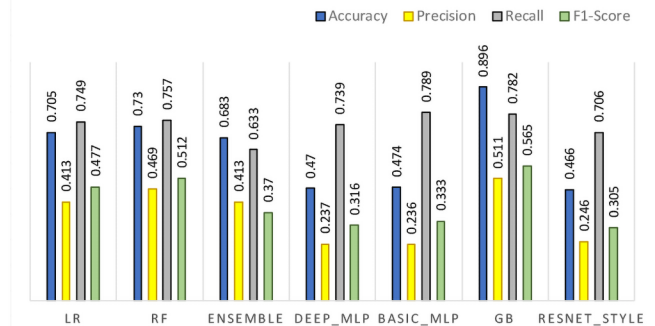


Figure 7. Architecture-level comparison for the 2 most informative modality settings. **Note.** C2 = thermal video, milking system data, and rumen bolus sensor data; C3 = milking system data and rumen bolus sensor data; LR = Logistic Regression; RF = Random Forest; GB = Gradient Boosting; MLP = multilayer perceptron.

under configuration C3, GB provided stronger threshold-based classification performance, whereas RF remained competitive in terms of overall ranking ability. Because C2 provided the strongest ranking performance among the evaluated modality configurations, the remaining detailed validation analysis focused on this configuration. This focus allowed a more detailed assessment of discrimination ability, threshold behavior, and classification error patterns for the best performing multimodal setting. Results for the other modality configurations are reported in the benchmark comparison, whereas the following analysis examine C2 in greater detail.

Figure 8 shows the confusion matrices of the Logistic Regression (LR) and Random Forest (RF) models at their optimal thresholds, calculated on the held-out validation subset. This validation subset contained 170 observations, corresponding to approximately 20.3% of the full analytic data set of 838 observations. The LR model (threshold 0.03) correctly identifies 49 anomalies but misses 19, while maintaining 94 correct healthy predictions and 8 false positives. In contrast, the RF model (threshold 0.18) improves anomaly detection, correctly identifying 61 cases and reducing false negatives to 7, with the same number of correctly classified healthy instances (94) and false positives (8). Overall, RF achieves a better balance by significantly improving anomaly recall.

Figure 9 illustrates the relationship between classification threshold and performance measures for the LR and RF models. In both cases, the default threshold (0.5) leads to low F1-scores due to reduced recall, indicating that the models fail to identify a sufficient number of anomaly cases.

Feature Importance Analysis

Figure 10 presents the feature importance analysis for the C2 configuration using Random Forest (RF) and Logistic Regression (LR). The results show that both models are primarily driven by milking and behavior derived variables, while thermal features play a secondary role. For the RF model, the most influential predictors are conductivity and bolus activity-related features, with *conductivity_max*, and *conductivity_dp*, *bolus_activity_max* having the highest importance scores. These variables are well-established indicators of udder health and inflammation, explaining their dominant contribution to anomaly detection.

A similar pattern is observed for the LR model, where *conductivity_max* is by far the most influential feature, followed by *bolus_Temp without Drinkcycles_std* and *bolus_rumination min/24h_std*. In contrast to RF, LR assigns stronger weight to conductivity-related changes and rumination dynamics, but still relies primarily on milking system variables.

Table 3. Detailed validation summary for the selected C2 and C3 models

No	Model	Best F1	Best threshold	ROC-AUC	PR-AUC
C2	LR	0.780	0.030	0.906	0.861
C2	RF	0.890	0.175	0.934	0.924
C3	RF	0.746	0.150	0.876	0.719
C3	GB	0.874	0.370	0.874	0.815

Note. Values were calculated after model specific decision threshold optimization.

C2 = thermal video, milking system data, and rumen bolus sensor data;.

C3 = milking system data and rumen bolus sensor data;.

LR = Logistic Regression;.

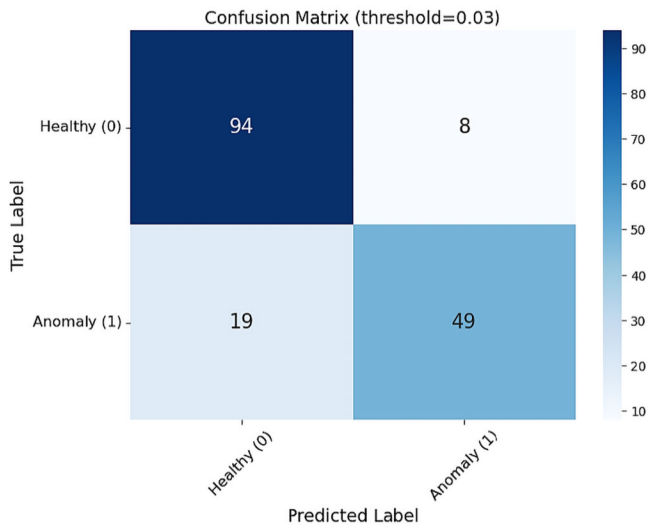
RF = Random Forest;.

GB = Gradient Boosting;.

ROC AUC = receiver operating characteristic area under the curve;.

PR AUC = precision recall area under the curve.

a) Confusion matrix of LR model



b) Confusion matrix of RF model

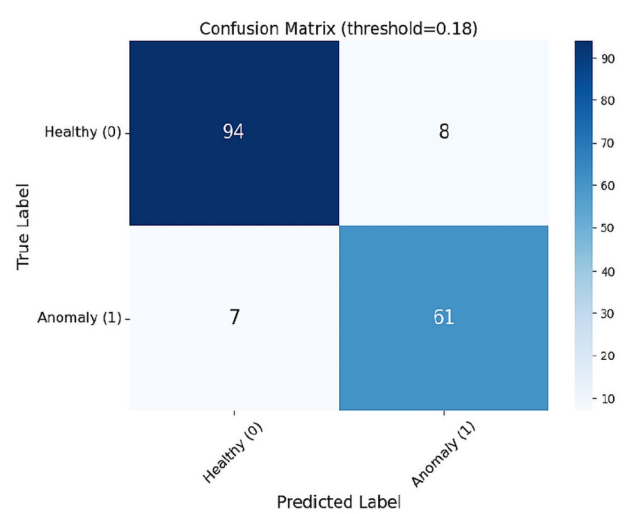


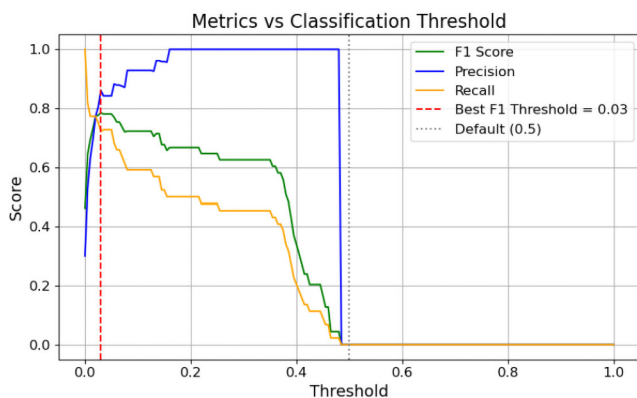
Figure 8. Confusion matrices of C2 logistic regression and random forest models. **Note.** C2 = thermal video, milking system data, and rumen bolus sensor data; LR = Logistic Regression; RF = Random Forest.

Thermal features (e.g., intensity-based measures) appear lower in the ranking for both models, indicating that they contribute additional, complementary information rather than serving as primary predictors. Overall, the results confirm that while thermal imaging enhances the model, the dominant predictive signal in C2 originates from milking data.

Feature importance analysis reveals how C3 models achieve these results (see Figure 11). C3 RF distributed importance across several milking and bolus variables, with `conductivity_max` (0.250), `scc_mean` (0.130), `scc_max` (0.119), `flow_max_dp` (0.075), `flow_mean` (0.062), and `bolus_activity_std` (0.062) among the most

influential predictors. This pattern suggests that the RF model exploited a genuinely multimodal signal. In contrast, the C3 GB model relied overwhelmingly on a single dominant feature, `conductivity_max` (0.842), with substantially smaller contributions from `flow_max_du` (0.075), `blood_max_ppm` (0.045), and `blood_detected` (0.039). Therefore, both models performed well in C3, but they did so in different ways: RF exploited broader multimodal complementarity, whereas GB extracted most of its predictive power from a narrower but very strong core of milking-related indicators.

a) Threshold analysis for LR model



b) Threshold analysis for RF model

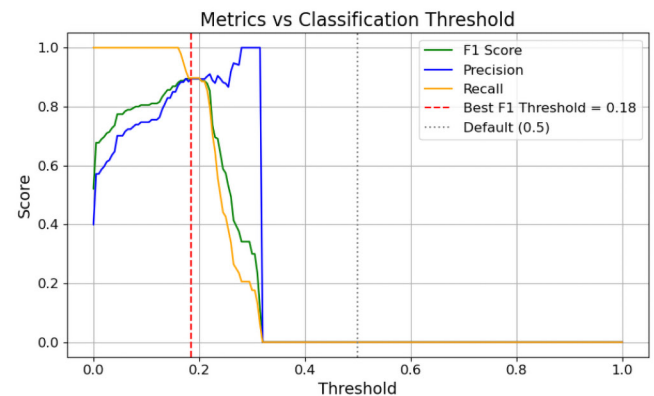


Figure 9. Threshold analysis for best performing C2 models: a) LR model and b) RF model. **Note.** C2 = thermal video, milking system data, and rumen bolus sensor data; LR = Logistic Regression; RF = Random Forest.

Interpretation of Hybrid Model Performance

The results demonstrate that the effectiveness of the proposed multimodal anomaly detection framework is highly dependent on both the selected sensing modalities and the classification models. A clear distinction emerges between configurations that optimize ranking performance and those that provide stronger threshold-based classification, highlighting the importance of evaluation perspective in anomaly detection tasks.

The best overall classification performance is achieved by C3 (milking system data + bolus sensors), indicating

that structured physiological and production-related variables contain the most direct and reliable indicators of animal health status. Features such as somatic cell count and milk conductivity are strongly associated with pathological conditions and consistently dominate the predictive process. This aligns with established domain knowledge and confirms that these modalities provide high signal-to-noise ratios for anomaly detection. In contrast, C2 (full multimodality) achieves the highest ROC-AUC values, demonstrating superior ranking capability. This suggests that thermal imaging contributes additional discriminative information that improves class



Figure 10. Feature importance analysis for C2 models. **Note.** C2 = thermal video, milking system data, and rumen bolus sensor data.

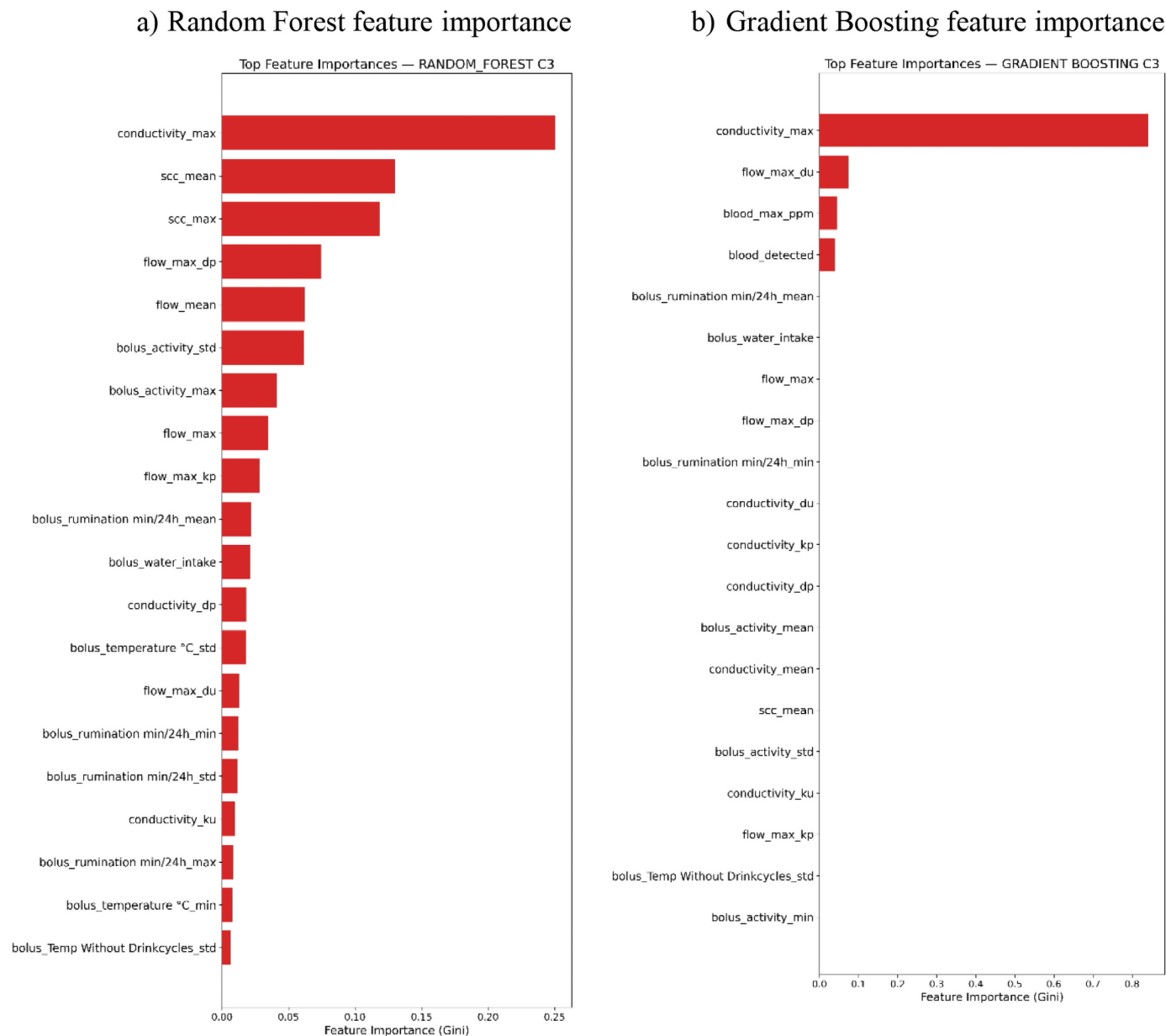


Figure 11. Feature importance analysis for C3 models. **Note.** C2 = thermal video, milking system data, and rumen bolus sensor data.

separability across thresholds. However, its impact on threshold-based measures such as F1-score is more limited, indicating that thermal features primarily enhance relative risk estimation rather than direct classification without calibration.

Across all configurations, tree-based models consistently outperform neural network approaches. These models show strong robustness to heterogeneous feature spaces and effectively capture nonlinear relationships between variables. In contrast, the MLP exhibits higher variability and reduced performance, particularly in configurations relying on tabular features, suggesting lim-

ited suitability for this task without further architectural adaptation.

A critical finding is the strong dependence of model performance on decision threshold selection and probability calibration. At the default threshold of 0.5, most models exhibit conservative behavior, resulting in low recall and under-detection of anomalies. After threshold optimization, performance improves substantially, indicating that raw classification outputs do not fully reflect model capability. This highlights the importance of threshold tuning in practical deployment scenarios, where the cost of false negatives is often high. Feature

importance analysis further confirms that milking system variables dominate prediction, while thermal features provide complementary support rather than primary predictive power. This suggests that thermal imaging enhances robustness and contextual understanding but is not sufficient as a standalone modality.

From an application perspective, the framework demonstrates robustness to partial modality availability, which is essential for real-world deployment. Strong performance using only milking and bolus data indicates that the system remains functional even in the absence of thermal imaging, while the inclusion of thermal data provides additional value when available.

Study Limitations

Despite these strengths, the study is limited by data set size and the relatively small number of anomaly cases, which may affect generalizability. Additionally, variability in sample sizes across configurations introduces challenges in direct performance comparison. A further limitation is that thermal recordings were obtained during the autumn study period and therefore do not fully represent seasonal variation in barn temperature, humidity, ventilation, or animal surface-temperature contrast. Although adaptive thresholding and feature normalization reduce dependence on absolute thermal intensity values, model performance under winter, spring, and summer conditions remains to be validated. Seasonal recalibration or incorporation of ambient environmental measurements may be required before broader deployment.

CONCLUSIONS

This study presents a multimodal framework for automated anomaly detection in dairy cattle by integrating thermal video, milking-system data, and rumen bolus measurements in a machine-learning pipeline designed for real farm conditions. The approach combines statistical thermal descriptors, deep visual-temporal features, and structured sensor variables to exploit complementary information across modalities. Results show that performance depends strongly on modality selection and decision-threshold optimization. Among the evaluated selected models, the C2 configuration achieved the strongest overall performance, with RF reaching the highest F1-score of 0.890 at a threshold of 0.175, together with the highest ROC-AUC of 0.934 and PR-AUC of 0.924. In C3, GB achieved the best threshold-based result, with an F1-score of 0.874 at a threshold of 0.370. These findings indicate that production and physiological variables provide strong predictive signals, while thermal information can further improve model discrimination. Overall, tree-based models showed the most robust performance.

NOTES

Funding This research was co-funded by Research Council of Lithuania, programme “Information technologies for the development of science and knowledge society” agreement No. S-ITP-24-5, “Machine learning algorithms for cow health analysis and prediction (MALACA).” The funder had no role in study design, data analysis, interpretation, or manuscript preparation.

Institutional Review Board Statement All procedures performed in studies involving animals were in accordance with the ethical standards of the institution at which the studies were conducted (Lithuanian University of Health Sciences (Kaunas, Lithuania) Bioethics Committee protocol No 2024-BEC3-T-006, 12-03-2024).

Authorship contribution statement: Arnas Nakrošis Data Curation, Methodology, Supervision, Validation, Writing – original draft, Writing – review and editing. **Mahammad Ismayilov:** Conceptualization, Software, Writing – original draft. **Gurban Shukurov:** Formal Analysis, Methodology, Software, Visualization, Writing – original draft. **Adhithiyan Arul Mani:** Investigation, Software. **Sanjaykrishna Ravali Srinivas:** Methodology, Software. **Joris Věžys:** Data Curation, Validation. **Vytautas Ostaševičius:** Funding Acquisition, Resources, Supervision. **Agnė Paulauskaitė:** Conceptualization, Funding Acquisition, Project Administration, Visualization, Writing – review and editing.

Conflict of interest The authors have not stated any conflicts of interest.

Nonstandard abbreviations: AI = artificial intelligence; AUC = area under the curve; CNN = convolutional neural network; GB = gradient boosting; GNSS = global navigation satellite system; IMU = inertial measurement unit; IoT = Internet of Things; IR = infrared; LR = logistic regression; MLP = multilayer perceptron; PLF = precision livestock farming; PR-AUC = area under the precision–recall curve; ReLU = rectified linear unit; ResNet = residual network; RF = random forest; ROC-AUC = area under the receiver operating characteristic curve; ROI = region of interest.

REFERENCES

- Aguilar-Lazcano, C. A., I. E. Espinosa-Curiel, J. A. Ríos-Martínez, F. A. Madera-Ramírez, and H. Pérez-Espinosa. 2023. Machine Learning-Based Sensor Data Fusion for Animal Monitoring: Scoping Review. *Sensors (Basel)* 23:5732. <https://doi.org/10.3390/s23125732>.
- Akhigbe, B. I., K. Munir, O. Akinade, L. Akanbi, and L. O. Oyedele. 2021. IoT Technologies for Livestock Management: A Review of Present Status, Opportunities, and Future Trends. *Big Data and Cognitive Computing* 5:10. <https://doi.org/10.3390/bdcc5010010>.
- Anagnostopoulos, A., B. E. Griffiths, N. Siachos, J. Neary, R. F. Smith, and G. Oikonomou. 2023. Initial validation of an intelligent video surveillance system for automatic detection of dairy cattle lameness. *Front. Vet. Sci.* 10:1111057. <https://doi.org/10.3389/fvets.2023.1111057>.

- Arablouei, R., Z. Wang, G. J. Bishop-Hurley, and J. Liu. 2023. Multimodal sensor data fusion for in situ classification of animal behavior using accelerometry and GNSS data. *Smart Agricultural Technology* 4:100163. <https://doi.org/10.1016/j.atech.2022.100163>.
- Asogan, A., N. Szali, A. S. Veerendra, L. Samylingam, N. Aslfattahi, C. K. Kok, and K. Kadirgama. 2026. A review on the impact of AI-enabled thermal imaging and IoT sensor fusion on early detection of mastitis in dairy cattle. *Biosens. Bioelectron.* X 28:100735. <https://doi.org/10.1016/j.biosx.2025.100735>.
- Bobić, T., N. Raguž, M. Oroz, M. Oroz, M. Gregić, P. Mijić, D. Kranjac, and B. Lukić. 2025. Application of Infrared Thermography in the Detection of Hoof Disease and Lameness in Cattle. *Animals (Basel)* 15:1086. <https://doi.org/10.3390/ani15081086>.
- Cho, S.-H., M. Lee, W.-H. Lee, S. Seo, and D.-H. Lee. 2024. Mastitis Classification in Dairy Cows Using Weakly Supervised Representation Learning. *Agriculture* 14:2084. <https://doi.org/10.3390/agriculture14112084>.
- Deng, Y., H. Qu, A. Leng, X. Tang, and S. Zhai. 2025. Methods and challenges in computer vision-based livestock anomaly detection, a systematic review. *Biosyst. Eng.* 253:104135. <https://doi.org/10.1016/j.biosystemseng.2025.104135>.
- Dervić, E., C. Matzhold, C. Egger-Danner, F. Steininger, and P. Klimek. 2024. Improving lameness detection in cows: A machine learning algorithm application. *J. Dairy Sci.* 107:11550–11562. <https://doi.org/10.3168/jds.2024-24730>.
- Ding, L., C. Zhang, Y. Yue, C. Yao, Z. Li, Y. Hu, B. Yang, W. Ma, L. Yu, R. Gao, and Q. Li. 2025. Wearable Sensors-Based Intelligent Sensing and Application of Animal Behaviors: A Comprehensive Review. *Sensors (Basel)* 25:4515. <https://doi.org/10.3390/s25144515>.
- Ferreira, R. E. P., M. Angels de Luis Balaguer, T. Bresolin, R. Chandra, G. J. M. Rosa, H. M. White, and J. R. R. Dórea. 2024. Multi-modal machine learning for the early detection of metabolic disorder in dairy cows using a cloud computing framework. *Comput. Electron. Agric.* 227:109563. <https://doi.org/10.1016/j.compag.2024.109563>.
- Gómez, Y., A. H. Stygar, I. J. M. M. Boumans, E. A. M. Bokkers, L. J. Pedersen, J. K. Niemi, M. Pastell, X. Manteca, and P. Llonch. 2021. A Systematic Review on Validated Precision Livestock Farming Technologies for Pig Production and Its Potential to Assess Animal Welfare. *Front. Vet. Sci.* 8:660565. <https://doi.org/10.3389/fvets.2021.660565>.
- Gutiérrez-Galan, D., J. P. Dominguez-Morales, E. Cerezuela-Escudero, A. Rios-Navarro, R. Tapiador-Morales, M. Rivas-Perez, M. Dominguez-Morales, A. Jimenez-Fernandez, and A. Linares-Barranco. 2018. Embedded neural network for real-time animal behavior classification. *Neurocomputing* 272:17–26. <https://doi.org/10.1016/j.neucom.2017.03.090>.
- Hajnal, É., L. Kovács, and G. Vakulya. 2022. Dairy Cattle Rumen Bolus Developments with Special Regard to the Applicable Artificial Intelligence (AI). *Sensors (Basel)* 22:6812. <https://doi.org/10.3390/s22186812>.
- He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. Pages 770–778 in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Jadhav, S. P., A. Srinivas, P. Dipak Raghunath, M. Ramkumar Prabhu, J. Suryawanshi, and A. Haldorai. 2024. Deep learning approaches for multi-modal sensor data analysis and abnormality detection. *Measurement. Sensors (Basel)* 33:101157. <https://doi.org/10.1016/j.measen.2024.101157>.
- Jiang, B., W. Tang, L. Cui, and X. Deng. 2023. Precision Livestock Farming Research: A Global Scientometric Review. *Animals (Basel)* 13:2096. <https://doi.org/10.3390/ani13132096>.
- Kim, S., N. Yamagishi, S. Ishikawa, and S. Tsuchiaka. 2025. Unique temperature change patterns in calves eyes and muzzles: a non-invasive approach using infrared thermography and object detection. *Front. Vet. Sci.* 12:1548906. <https://doi.org/10.3389/fvets.2025.1548906>.
- Lamanna, M., M. Bovo, and D. Cavallini. 2025. Wearable Collar Technologies for Dairy Cows: A Systematized Review of the Current Applications and Future Innovations in Precision Livestock Farming. *Animals (Basel)* 15:458. <https://doi.org/10.3390/ani15030458>.
- Li, D., B. Li, Q. Li, Y. Wang, M. Yang, and M. Han. 2024. Cattle identification based on multiple feature decision layer fusion. *Sci. Rep.* 14:26631. <https://doi.org/10.1038/s41598-024-76718-x>.
- Liu, C., M. Cui, D. Zhang, Y. Lu, X. Yan, Y. Li, Z. Wang, Y. Zhang, M. Li, G. Zhao, and X. Liu. 2026. Machine learning-based prediction of clinical mastitis in dairy cows: A comparative analysis of 9 algorithms using production and management data. *J. Dairy Sci.* 109:5260–5275. <https://doi.org/10.3168/jds.2025-27483>.
- Ma, J., X. Yang, Y. Liu, P. Xin, Q. Tong, C. Liang, L. Yu, A. Liu, and C. Wang. 2026. Toward smart health monitoring: multimodal sensing and intelligent disease diagnosis in poultry and livestock. *Anim. Front.* <https://doi.org/10.1093/af/vfaf060>.
- Magana, J., D. Gavojdian, Y. Menahem, T. Lazebnik, A. Zamansky, and A. Adams-Progar. 2023. Machine learning approaches to predict and detect early-onset of digital dermatitis in dairy cows using sensor data. *Front. Vet. Sci.* 10:1295430. <https://doi.org/10.3389/fvets.2023.1295430>.
- Mahmud, M. S., A. Zahid, A. K. Das, M. Muzammil, and M. U. Khan. 2021. A systematic literature review on deep learning applications for precision cattle farming. *Comput. Electron. Agric.* 187:106313. <https://doi.org/10.1016/j.compag.2021.106313>.
- Mao, A., E. Huang, X. Wang, and K. Liu. 2023. Deep learning-based animal activity recognition with wearable sensors: Overview, challenges, and future directions. *Comput. Electron. Agric.* 211:108043. <https://doi.org/10.1016/j.compag.2023.108043>.
- Pan, S. J., and Q. Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22:1345–1359. <https://doi.org/10.1109/TKDE.2009.191>.
- Paulauskaite-Taraseviciene, A., A. Nakrosis, J. Zymantiene, V. Jurenas, J. Vezys, A. Sederevicius, R. Gruzauskas, V. Obarauskas, R. Japertiene, A. Bubulis, L. Kizauskiene, I. Silinskas, J. Zemaitis, and V. Ostasevicius. 2026. AI-Driven Multimodal Sensing for Early Detection of Health Disorders in Dairy Cows. *Animals (Basel)* 16:411. <https://doi.org/10.3390/ani16030411>.
- Perez, M. M., E. Cabrera, C. Rial, Y. You, Y. Wang, K. Weinberger, D. V. Nydam, and J. O. Giordano. 2025. Screening and selection of a machine learning algorithm for development of a model to select cows for clinical examination using data from automated health monitoring technologies and other predictors of cow health. *J. Dairy Sci.* 108:13720–13738. <https://doi.org/10.3168/jds.2025-26511>.
- Ramuada, M., T. L. Tyasi, L. Gumede, and T. Chitura. 2024. A practical guide to diagnosing bovine mastitis: a review. *Front. Anim. Sci.* 5:1504873. <https://doi.org/10.3389/fanim.2024.1504873>.
- Rohan, A., M. S. Rafiq, M. J. Hasan, F. Asghar, A. K. Bashir, and T. Dotorini. 2024. Application of deep learning for livestock behaviour recognition: A systematic literature review. *Comput. Electron. Agric.* 224:109115. <https://doi.org/10.1016/j.compag.2024.109115>.
- Shi, Z., F. Chang, Y. Jia, J. Li, Y. Qiu, J. Miao, W. Jiang, X. Guo, X. Han, and W. Tang. 2024. Classifying and Understanding of Dairy Cattle Health Using Wearable Inertial Sensors With Random Forest and Explainable Artificial Intelligence. *IEEE Sens. Lett.* 8:1–4. <https://doi.org/10.1109/LSENS.2024.3358619>.
- Siachos, N., J. M. Neary, R. F. Smith, and G. Oikonomou. 2024. Automated dairy cattle lameness detection utilizing the power of artificial intelligence; current status quo and future research opportunities. *Vet. J.* 304:106091. <https://doi.org/10.1016/j.tvjl.2024.106091>.
- Sohan, M. F., R. Alzubi, H. Alzoubi, E. Albalawi, and A. H. A. Hafez. 2025. Direct video-based spatiotemporal deep learning for cattle lameness detection. *Sci. Rep.* 16:142. <https://doi.org/10.1038/s41598-025-29118-8>.
- Stygar, A. H., Y. Gómez, G. V. Berteselli, E. Dalla Costa, E. Canali, J. K. Niemi, P. Llonch, and M. Pastell. 2021. A Systematic Review on Commercially Available and Validated Sensor Technologies for Welfare Assessment of Dairy Cattle. *Front. Vet. Sci.* 8:634338. <https://doi.org/10.3389/fvets.2021.634338>.
- Tassinari, P., M. Bovo, S. Benni, S. Franzoni, M. Poggi, L. M. E. Mammi, S. Mattoccia, L. Di Stefano, F. Bonora, A. Barbaresi, E. Santolini, and D. Torreggiani. 2021. A computer vision approach based on deep learning for the detection of dairy cows in free stall barn. *Comput. Electron. Agric.* 182:106030. <https://doi.org/10.1016/j.compag.2021.106030>.

- Wang, J., Z. Yu, R. C. Chebel, and H. Yu. 2025. Impact of cross-validation designs on cattle behavior prediction using machine learning and deep learning models with tri-axial accelerometer data. *Smart Agricultural Technology* 12:101483. <https://doi.org/10.1016/j.atech.2025.101483>.
- Wang, Y., X. Kang, Z. He, Y. Feng, and G. Liu. 2022. Accurate detection of dairy cow mastitis with deep learning technology: a new and comprehensive detection method based on infrared thermal images. *animal* 16:100646. <https://doi.org/10.1016/j.animal.2022.100646>.
- Xudong, Z., K. Xi, F. Ningning, and L. Gang. 2020. Automatic recognition of dairy cow mastitis from thermal images by a deep learning detector. *Comput. Electron. Agric.* 178:105754. <https://doi.org/10.1016/j.compag.2020.105754>.
- Zhou, X., C. Xu, H. Wang, W. Xu, Z. Zhao, M. Chen, B. Jia, and B. Huang. 2022. The Early Prediction of Common Disorders in Dairy Cows Monitored by Automatic Systems with Machine Learning Algorithms. *Animals (Basel)* 12:1251. <https://doi.org/10.3390/ani12101251>.

ORCIDS

- Arnas Nakrošis,  <https://orcid.org/0000-0002-2168-9981>
- Mahammad Ismayilov,  <https://orcid.org/0009-0000-0212-7628>
- Gurban Shukurov,  <https://orcid.org/0000-0002-5140-4588>
- Adhithyan Arul Mani,  <https://orcid.org/0009-0006-9318-0419>
- Sanjaykrishna Ravali Srinivas,  <https://orcid.org/0009-0009-7294-9737>
- Joris Věžys,  <https://orcid.org/0009-0000-3731-0588>
- Vytautas Ostaševičius,  <https://orcid.org/0000-0002-4234-6996>
- Agne Paulauskaitė  <https://orcid.org/0000-0002-8787-3343>