

Simulation-based detection of bovine health anomalies using machine learning and synthetic IoT sensor data

Detección de anomalías en la salud bovina mediante simulación, utilizando aprendizaje automático y datos sintéticos de sensores IoT

Grace Viteri-Guzmán^{1*} , Jorge García-Cevallos² , Sedolfo Carrasquero-Ferrer¹ , José Jaime-Carriel³ 

¹Universidad Tecnológica Empresarial de Guayaquil, Departamento de Innovación. Guayaquil, Ecuador.

²Universidad Tecnológica Empresarial de Guayaquil, Facultad de Ingeniería. Guayaquil, Ecuador.

³Universidad Politécnica Salesiana, Facultad de Computación. Guayaquil, Ecuador.

*Corresponding author: gviteri@uteg.edu.ec

ABSTRACT

The objective of this research was to evaluate a machine learning-based methodology for the early detection of health abnormalities in cattle by analyzing physiological and feeding behavior variables obtained from IoT sensors in real-world settings. For this research, a synthetic dataset of body temperature, heart rate, and feeding activity records from forty cattle was generated, incorporating normal physiological variations and controlled abnormalities. Preprocessing included the removing records (rows) containing missing values and normalizing of variables to ensure comparability. Three machine learning models were implemented: Random Forest as a supervised classifier to differentiate between normal and abnormal physiological states, K-Means as an unsupervised algorithm to identify feeding behavior patterns, and the Isolation Forest algorithm for the individual detection of abnormalities without requiring predefined clinical thresholds. The Random Forest model demonstrated adequate performance in classifying physiological states. The K-Means algorithm allowed the cattle to be grouped into two distinct clusters based on average feeding rate, revealing patterns of low and high feeding activity. For its part, Isolation Forest identified individuals with atypical values for physiological and behavioral variables, isolating unusual profiles. It is concluded that this approach provides a methodological basis for the development of intelligent monitoring systems in precision livestock farming, with potential application in real-world scenarios through the incorporation of IoT sensor data, thus contributing to improved animal welfare and optimized bovine productivity.

Key words: Artificial intelligence; Information technology; Simulation model; Cattle; Welfare

RESUMEN

El objetivo de esta investigación fue evaluar una metodología basada en aprendizaje automático para la detección temprana de anomalías en el estado de salud del ganado vacuno mediante el análisis de variables fisiológicas y de comportamiento alimentario que se obtienen de sensores IoT en la práctica real. Para esta investigación se generó un conjunto de datos sintético de registros de temperatura corporal, frecuencia cardíaca y actividad alimentaria de cuarenta bovinos, incorporando variaciones fisiológicas normales y anomalías controladas. El preprocesamiento incluyó la eliminación de registros que contienen valores perdidos y la normalización de variables para garantizar la comparabilidad. Se implementaron tres modelos de aprendizaje automático, el Random Forest como clasificador supervisado para diferenciar entre estados fisiológicos normales y anormales, el K-Means como algoritmo no supervisado para identificar patrones de conducta alimentaria, y el algoritmo Isolation Forest para la detección individual de anomalías sin requerir umbrales clínicos predefinidos. El resultado del modelo Random Forest demostró un rendimiento adecuado en la clasificación de estados fisiológicos. El resultado en el algoritmo K-Means permitió agrupar al ganado en dos conglomerados distintos según la tasa de alimentación promedio, mostrando patrones de actividad alimentaria baja y alta. Por su parte, Isolation Forest identificó individuos con valores atípicos de variables fisiológicas y de conducta, aislando perfiles inusuales. Se concluye que este enfoque proporciona una base metodológica para el desarrollo de sistemas de monitorización inteligentes en la ganadería de precisión, con potencial aplicación en escenarios reales mediante la incorporación de datos de sensores IoT, lo que contribuye a mejorar el bienestar animal y optimizar la productividad de los bovinos.

Palabras clave: Inteligencia artificial; Tecnología de la información; Modelo de simulación; Ganado vacuno; Bienestar

INTRODUCTION

Early detection of physiological changes in cattle is key to reducing production losses and improving animal welfare. In recent years, precision livestock farming has incorporated IoT technology such as sensors and systematic analysis for the continuous monitoring of physiological and behavioral variables [1, 2]. This allows for the efficient and sustained management of animal production, health, and welfare, reducing the workload and improving decision-making for producers [3, 4].

Physiological and feeding behavior variables are relevant indicators of bovine health status, and the use of machine learning techniques offers high potential for early detection of abnormalities in contexts with large volumes of data [2, 5].

In other studies, the classification of health status using supervised models depends on labeled data, which is scarce in production contexts [6]; therefore, physiological ranges reported in the literature are used to generate reference labels based on body temperature and heart rate in cattle.

On the other hand, variables related to eating behavior do not have defined clinical ranges and depend heavily on the type of sensor, environmental conditions and data processing methodologies [7]; therefore, they require unsupervised models to identify patterns and detect abnormal behaviors, which has been addressed in previous research through clustering and anomaly detection techniques [8].

In this context, this research aimed to develop and evaluate a machine learning (ML) based approach for the analysis of physiological and feeding behavior data in cattle (*Bos taurus*) in order to detect possible health alterations. To achieve this: a) develop a process to generate synthetic physiological and feeding behavior data for ML models; b) implement ML classification models. Random Forest (supervised learning), K-Means and Isolation Forest (unsupervised learning) to classify the health status of cattle; and c) evaluate the performance of the proposed ML models using classification metrics, clustering analysis, and anomaly detection.

MATERIALS AND METHODS

This study employed a quantitative and exploratory approach to analyze physiological and behavioral data from livestock using ML algorithms, with the aim of identifying anomalous patterns related to their health status. The methodology involved implementing three ML models in Python using the scikit-learn library. The procedure was structured into four main stages:

- » First, a simulated dataset was generated to emulate the readings from the cattle's IoT sensors, including physiological variables (temperature and heart rate) and behavioral variables (feeding behavior). Anomalous data were injected by modifying physiological ranges and feeding behavior.
- » Second, data preprocessing was limited to removing incomplete records or null values using pandas and scikit-learn (dropna). This procedure preserved continuous variables on their original scale, ensuring compatibility with ML algorithms.

- » Third, three machine learning models were implemented for anomaly detection: Isolation Forest and K-Means as unsupervised approaches, and Random Forest as a supervised classifier.
- » Fourth, the Random Forest, K-Means and Isolation Forest models were evaluated using metrics based on their classification, clustering anomaly detection capabilities.

Data generation

A synthetic dataset was generated using a Python script that employs a first-order autoregressive (AR(1)) model to ensure temporal consistency and biological realism in the simulated multivariate time series. Body temperature and heart rate were modeled with a temporal persistence coefficient of 0.85, ensuring that each continuous measurement is logically dependent on its immediately preceding state.

The synthetic data consists of physiological time series data collected every 5 minutes corresponding to 40 cattle, each associated with a unique IoT identifier (Things ID). This sampling frequency corresponds to 288 records per day; therefore, the dataset simulates approximately three days of continuous monitoring per cattle. A total of 864 records per animal were generated, resulting in 34,560 instances in the dataset. To ensure data structure and avoid bias during model training, the dataset architecture is stratified by animal identity (Animal Name) associated with a unique IoT identifier (Device ID), ensuring that each of the 40 cattle is represented fairly and proportionally in the time series.

Each record includes the following attributes: animal name, Things (IoT) identifier, measurement date, temperature, heart rate, and feeding (TABLE I).

The variable "Feeding" was generated as a binary variable using a stochastic Bernoulli process representing feeding activity events recorded by the sensor. To simulate natural feeding behavior in cattle, a value of 1 was assigned when the animal was feeding and

TABLE I
Simulated data outline for 40 cattle

Attributes	Description
Animal Name	Unique animal identifier
Things	IoT device identifier
Date	Measurement timestamp
Temperature (°C)	Simulated body temperature variable
Pulse Rate (bpm)	Simulated heart rate variable
Feeding	Binary variable indicating feeding activity (1 = feeding, 0 = not feeding)
Physio_Anomaly	Binary variable indicating physiological anomaly (1 = anomaly, 0 = no anomaly)
Behavior_Anomaly	Binary variable indicating feeding behavior anomaly (1 = anomaly, 0 = no anomaly)
Anomaly	Binary variable indicating normal (0) or abnormal (1) behavior determined from the detection of physiological (Physio_Anomaly) or feeding behavior (Behavior_Anomaly) anomalies

0 when it was not. To simulate variability in feeding behavior among cattle and over time, the probability of occurrence of this event was randomly generated within a range between 0.15 and 0.35.

Normal physiological ranges were defined based on [9], who state that the average temperature under conditions of well-being ranges from approximately 38.0°C to 39.3°C. This range reflects the average temperature of vital organs such as the heart, liver, and brain under normal conditions, although they note that it may vary slightly depending on factors such as age, stage of lactation, estrous cycle, and time of day, but generally remains within these limits when the cow is healthy and without heat stress. And 76–96 beats per minute (bpm) for heart rate in healthy animals under controlled management conditions. These reference data were used to define the “normal” data ranges in the synthetic dataset [10].

Small random variations through Gaussian white noise were injected into the dataset in order to simulate physiological conditions and abnormal behaviors to mimic how these physiological values naturally change in real animals and to include case closet o normal limits. This means that the data from healthy and abnormal animals are not perfectly separated, but are partially mixed, as occurs in real situations.

To identify atypical patterns using machine learning algorithms, explicit rules were defined for the generation of anomalous events within the synthetic dataset. Any record with a body temperature above 39.5°C or below 37.5°C, or a heart rate above 99 bpm or below 75 bpm, was considered a physiological anomaly.

Additionally, behavioral anomalies were defined when the proportion of feeding events deviated by more than one (1) standard deviation from average during a continuous period of two hours (24 consecutive records). This criterion is based on studies in which monitoring systems have observed deviations in behavioral patterns such as the relationship between feeding time and rumination, which can be detected in time for the early identification of abnormalities in bovine behavior [11, 12, 13].

These conditions allowed control variations to be introduced into the dataset to simulate possible physiological alterations or changes in the feeding behavior of cattle.

This synthetic dataset generation scheme (FIG. 1) allows controlled experimentation for anomaly detection algorithms, while avoiding the ethical and logistical limitations associated with collecting real data on animals. This approach is common in methodological studies of machine learning, where synthetic data allows the behavior of algorithms to be evaluated under controlled conditions before implementing them in real IoT-based animal monitoring systems.

Data preprocessing

Preprocessing was limited to removing incomplete records (rows) or null values using the pandas and scikit-learn libraries, while maintaining physiological variables at their original scale due to the scale invariance of the tree-based algorithms evaluated in this research. To convert the sequential time-series data into

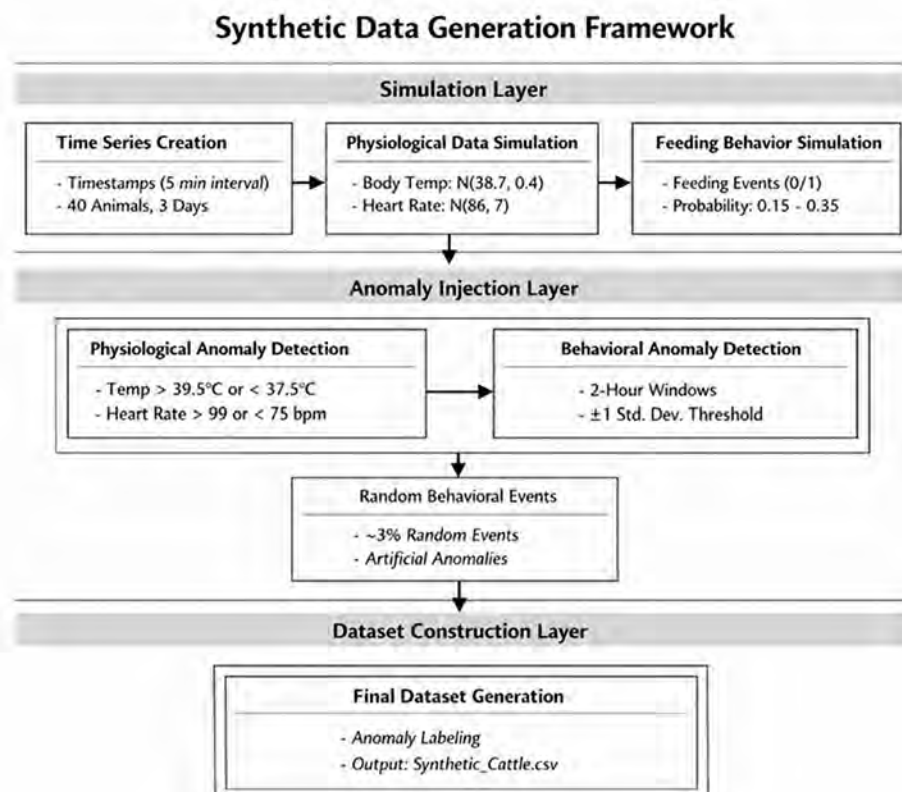


FIGURE 1. Diagram of the synthetic data generation process

column variables suitable for machine learning, the information was organized into a tabular feature vector format. In this structure, each row represents a specific 5-minute time interval, chronologically aligned by the animal's identity (Animal Name), while individual vital signs (Temperature, Pulse Rate) and behavioral states (Feeding) were arranged as separate columns. (FIG. 2).

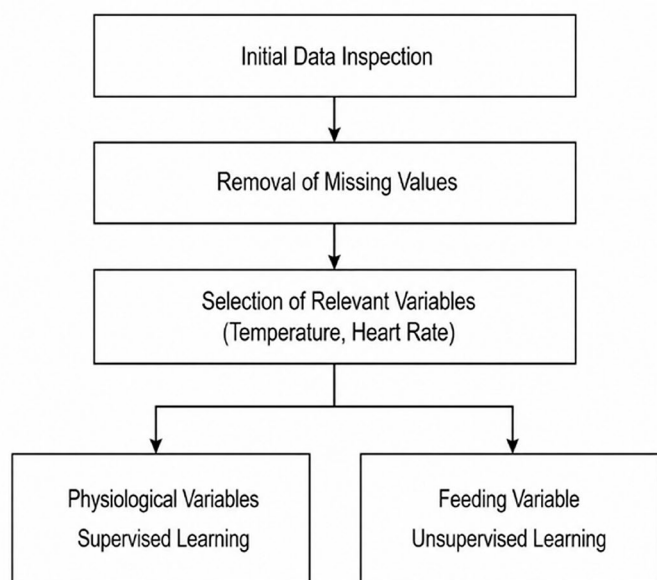


FIGURE 2. Data pre-processing scheme

Implementation of machine learning models

In order to comprehensively analyze bovine health status and detect anomalies, a hybrid approach was implemented that combines supervised and unsupervised machine learning techniques, depending on the nature of the variables considered.

Due to the availability of reference labels, physiological and behavioral variables were analyzed using supervised learning to train models for prediction and classification [14]. For this purpose, the classification models were constructed using the generated body temperature, heart rate, and feeding data in their original scale.

Since the dataset used in this research was synthetic, in other words, the reference labels did not come from actual clinical diagnoses but from physiological ranges reported in scientific literature. In this context, the ML model was not used as a definitive clinical diagnostic system, but rather as a mechanism for validating the behavior of the simulated dataset, verifying whether the patterns generated in the synthetic data can be correctly identified using supervised learning algorithms.

This approach is common in methodological studies based on synthetic data, where supervised models allow the statistical consistency of the dataset to be validated before its application in real scenarios with data from IoT sensors.

Additionally, the feeding behavior variable was integrated alongside the physiological data. This approach allows classification models to simultaneously analyze vital signs and behavioral patterns to detect complex abnormal events [14, 15].

To classify the physiological state of the animals, the Random Forest algorithm was used, using body temperature and heart rate as input variables. This algorithm constructs multiple decision trees on different subsets of data and combines their results to improve the accuracy and stability of the classifier [16]. To apply this algorithm, 70% of the data set was used for training and 30% for evaluating the model (FIG. 3). The model was configured with 100 decision trees ($n_estimators = 100$) and a value of $random_state = 42$ was used to ensure the reproducibility of the algorithm.

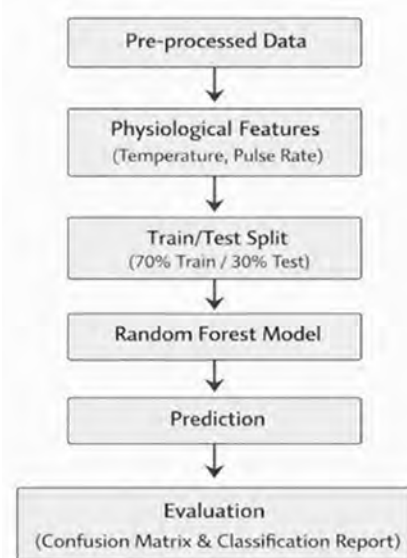


FIGURE 3. Random Forest algorithm scheme

To detect feeding behavior patterns without using prior labels, the K-Means algorithm was applied, a clustering method that divides a data set into homogeneous groups based on the distance between observations [8]. Feeding behavior was aggregated at the individual level by calculating the proportion of feeding events per cow. This metric naturally yields a bounded value between 0 and 1, representing the density of feeding activity within the analyzed time window without requiring any external scaling transformations. In this K-Means algorithm, the complete preprocessed dataset was used and the variable k = number of groups was configured to identify two groups of animals with different feeding activity patterns, allowing cows with normal and abnormal inactivity patterns to be detected. (FIG. 4).

In addition, Isolation Forest, an anomaly detection algorithm that identifies outliers by isolating points in a high-dimensional feature space [5], was used. The complete preprocessed dataset was used to implement this algorithm, with the aim of identifying atypical animal profiles within the herd. Body temperature, heart rate, and feeding behavior data were used for each individual cow,

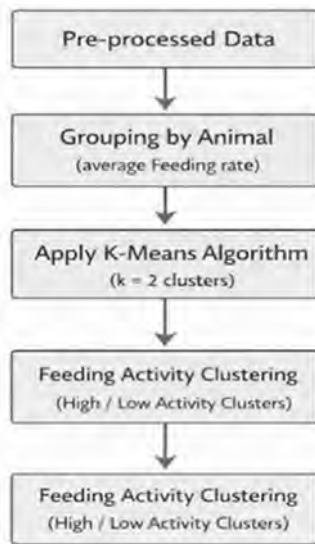


FIGURE 4. K-Means algorithm scheme

and descriptive statistics (mean and variance) were calculated. The contamination parameter was set to auto, allowing the algorithm to automatically determine the anomaly detection threshold from the distribution of anomaly scores generated by the data, enabling the identification of atypical animal profiles according to the inherent structure of the dataset. (FIG. 5).

Evaluation of the implemented algorithms

The performance of the supervised model based on Random Forest was evaluated using classification metrics (Confusion matrix), including accuracy, precision, recall, and F1-score, based on a comparison between the actual labels and the model's predictions. In the case of the unsupervised approach, the K-Means algorithm was validated using the silhouette coefficient to quantify the degree of separation and cohesion of the two clusters associated with eating behavior. Anomaly detection at the individual level was performed using the Isolation Forest algorithm, which was evaluated by analyzing the proportion of individuals

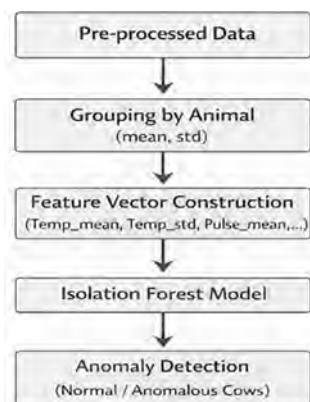


FIGURE 5: Schematic of the Isolation Forest algorithm

identified as atypical within the dataset, in accordance with the contamination parameter set at 'auto'.

Finally, the results generated by the supervised and unsupervised models were integrated to provide a joint assessment of the physiological status and feeding behavior pattern of the cattle, contributing to a monitoring scheme aimed at the early detection of abnormal conditions.

FIGURE 6, presents the general architecture of the proposed approach, in which physiological variables are analyzed using supervised learning, while behavioral variables are processed using an unsupervised method, and anomaly detection is performed at the individual level using an isolation model.

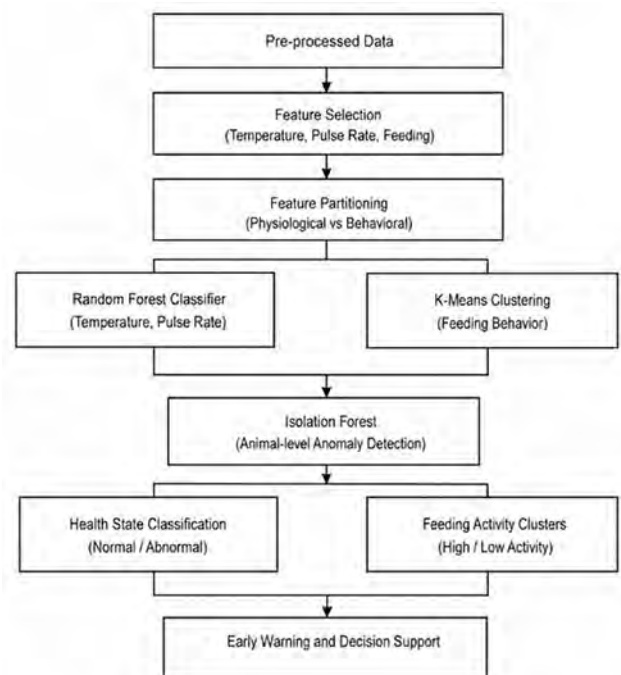


FIGURE 6. Global architecture of the implemented algorithms

Reproducibility of the methodology

The experiment was implemented in Google Colab using Python 3.10 with the pandas and scikit-learn libraries. To ensure reproducibility, the following was established:

- » Random_state = 42 in the Random Forest, K-Means, and Isolation Forest models, as well as in the partitioning of the dataset.
- » The Random Forest model was configured with 100 trees (n_estimators = 100) and test_size = 0.3 (test data).
- » K-Means was run with n_clusters = 2, n_init = 10.
- » Isolation Forest was configured with n_estimators = 100 and contamination='auto'.

Real-world implementation considerations

In a real-world environment, the analyzed variables (temperature, heart rate, and feeding behavior) can be acquired through an IoT architecture based on the Arduino WiFi ESP8266 microcontroller, which integrates and transmits data to cloud platforms for processing. Body temperature can be effectively measured using the DS18B20 sensor, while heart rate is obtained through the AD8232, and feeding behavior is inferred using the MPU6050 inertial sensor. This infrastructure (TABLE II) enables the continuous acquisition of physiological and feeding behavior data, facilitating their analysis in precision livestock farming systems, as reported in [17, 18].

System Function	Device/Sensor Type	Brand	Model	Country of Manufacture
Temperature	Temperature sensor	Maxim Integrated	DS18B20	United States
Pulse Rate	ECG sensor	Analog Devices	AD8232	United States
Feeding Activity	Inertial measurement unit (IMU)	ITDK InvenSense	MPU6050	Taiwan
Data Transmission	WiFi microcontroller	Espressif Systems	ESP8266	China
Data Management	Cloud platform	Arduino	Arduino Cloud	Italy

RESULTS AND DISCUSSION

The results obtained support the potential of both supervised and unsupervised machine learning approaches used for identifying the physiological and behavioral status of cattle based on synthetic data.

Results of the random forest model

TABLE III presents the confusion matrix obtained from the model, which indicates that, of 10,368 cases that correspond to the test set, it correctly classified 7,491 instances of normal physiological conditions (True Negatives) and 1,308 anomalous instances (True Positives). Conversely, the effects of simulated noise and overlapping boundaries led to misclassifications, registering 284 false positives where normal states were flagged as abnormal and 1,285 false negatives, representing anomalous events that went undetected by the algorithm.

Actual class / Predicted class	Normal (0)	Abnormal (1)
Normal (0)	7.491	284
Abnormal (1)	1.285	1.308
Total	8.776	1.592

The Random Forest model (TABLE IV) showed good overall performance in classifying bovine physiological states, achieving an overall accuracy of 85%. For the normal class (0), it obtained a precision of 0.85, a recall sensitivity of 0.96, and an F1 score of 0.91, demonstrating a high capacity to correctly identify normal physiological conditions.

On the other hand, for the abnormal class (1), the model achieved a precision of 0.82 but a sensitivity of 0.50, indicating that it only detected half of the abnormal cases present in the test group. The macro average (F1 score = 0.77) and weighted average (F1 score = 0.84) reflect generally stable performance.

	Precision	Recall	F1-score	Support
Normal (0)	0.85	0.96	0.91	7,775
Abnormal (1)	0.82	0.50	0.63	2,593
Accuracy			0.85	10,368
Macro avg	0.84	0.73	0.77	10,368
Weighted avg	0.85	0.85	0.84	10,368

As shown in Figure 7, the Random Forest model achieved an average accuracy of 85% and an AUC-ROC value of 0.74, demonstrating good ability to differentiate between normal and abnormal states in cattle. Although the model more readily identified normal cases than abnormalities, the results obtained demonstrate consistent and reliable performance. Furthermore, the 10-fold cross-validation showed minimal variation (± 0.01), confirming stable and consistent results, which supports its usefulness for animal health monitoring and reduces the risk of overfitting.

In particular, the performance of the Random Forest model was consistent with previous studies that have demonstrated the effectiveness of algorithms for classification tasks related

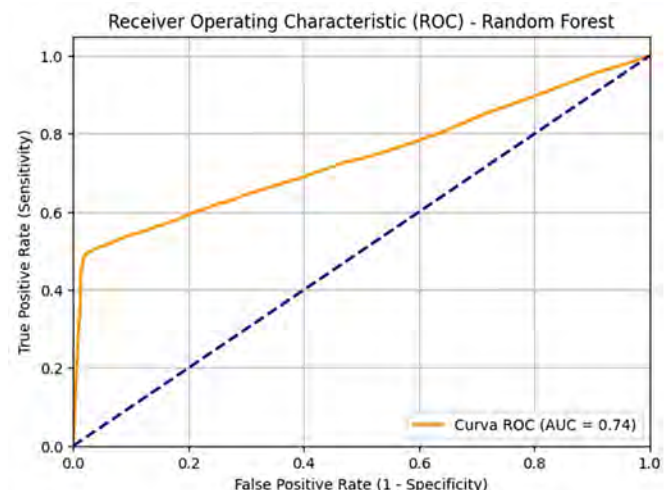


FIGURE 7. Receiver Operating Characteristic curve for the Random Forest model

to animal health and welfare, as indicated by the researchers [19] who applied this model to distinguish with good accuracy the health states in cattle using data derived from automatic sensors, surpassing other traditional methods for its ability to handle complexity and variability of the data.

Results of the K-means model

The results of the K-Means algorithm allowed the cows to be grouped according to their average feeding rate (Feeding_rate), based on the two fixed groups (K=2) that correspond to different patterns of normal and reduced feeding activity (TABLE V).

To evaluate the quality of the clustering, the silhouette coefficient was calculated, yielding a value of 0.59, which indicates a substantial and reliable cluster structure. Cluster 0 grouped 20 animals characterized by a higher feeding frequency, with average Feeding_rate values ranging approximately between 0.261574 and 0.366898. And cluster 1 also grouped 20 animals with lower feeding activity, with average Feeding_rate values between 0.149306 and 0.254630.

From a diagnostic perspective, these results indicated that Cluster 0 presented a higher proportion of feeding events, whereas Cluster 1 showed a lower frequency of feeding activity. This result agrees with the findings reported by researchers [20], who indicate that cattle exhibit individual variations in the frequency, duration, and temporal distribution of feeding events, thus identifying diverse feeding behavior patterns in cattle.

Similarly, the silhouette coefficient (0.59) indicates a substantial and robust separation between the groups, which means that the model has the ability to continuously discriminate between patterns of behavior. Consequently, the application of the K-Means algorithm has allowed for the identification of these patterns, revealing variations in feeding behavior that could be valuable for monitoring animal welfare in real-world precision livestock farming systems.

Results of Isolation Forest model

The results of the Isolation Forest algorithm (TABLE VI) indicated that twelve cows were identified as anomalous, while the remaining twenty-eight were classified as normal, according to the contamination parameter set at 'auto'. The detection was performed based on the joint analysis of the mean and standard deviation of body temperature, heart rate, and feeding behavior for each animal.

In this context, Isolation Forest identified 12 cows as abnormal: Cow-02, Cow-03, Cow-05, Cow-14, Cow-15, Cow-18, Cow-20, Cow-23, Cow-28, Cow-30, Cow-31 and Cow-32. In contrast, K-Means grouped 20 cows (50%) into cluster 1, characterized primarily by lower feeding rates. Both methods agreed in identifying cows Cow-18, Cow-20, Cow-23, Cow-28, Cow-30, Cow-31 and Cow-32, suggesting that these animals exhibit behavioral patterns clearly differentiated from the rest. However, K-Means also included other cows with relatively low feeding levels that were not considered abnormal by Isolation Forest. This indicates that Isolation Forest performed a more selective and specific detection of abnormalities, considering both physiological and behavioral variables. In contrast, K-Means has generated groupings based on similarities in feeding behavior, so both approaches are complementary for the analysis of livestock health and behavior.

TABLE V
Average and feed classification

Animal name	Feeding rate	Cluster	Animal name	Feeding rate	Cluster
Cow-1	0.299769	0	Cow-21	0.327546	0
Cow-2	0.317130	0	Cow-22	0.321759	0
Cow-3	0.270833	0	Cow-23	0.231481	1
Cow-4	0.290509	0	Cow-24	0.312500	0
Cow-5	0.364583	0	Cow-25	0.237269	1
Cow-6	0.247685	1	Cow-26	0.273148	0
Cow-7	0.265046	0	Cow-27	0.173611	1
Cow-8	0.254630	1	Cow-28	0.163194	1
Cow-9	0.309028	0	Cow-29	0.245370	1
Cow-10	0.216435	1	Cow-30	0.252315	1
Cow-11	0.172454	1	Cow-31	0.189815	1
Cow-12	0.310185	0	Cow-32	0.180556	1
Cow-13	0.261574	0	Cow-33	0.317130	0
Cow-14	0.366898	0	Cow-34	0.226852	1
Cow-15	0.299769	0	Cow-35	0.307870	0
Cow-16	0.318287	0	Cow-36	0.194444	1
Cow-17	0.340278	0	Cow-37	0.248843	1
Cow-18	0.149306	1	Cow-38	0.228009	1
Cow-19	0.325231	0	Cow-39	0.214120	1
Cow-20	0.165509	1	Cow-40	0.196759	1

From an analytical perspective, these findings demonstrated the capability of the Isolation Forest algorithm to identify atypical patterns in physiological and behavioral data, which could be associated with potential alterations in animal health or welfare status. The detection does not result from an arbitrary assignment, but rather from the internal structure of the data and the average path lengths generated by the model. From this point of view, some studies have highlighted the importance of unsupervised models for identifying abnormalities in the absence of clinical labels, particularly in contexts characterized by high biological variation.

Furthermore, although this initial analysis addressed the detection of global anomalies per bovine through statistical summaries, the architecture of the Isolation Forests algorithm also allows the identification of specific anomalies in time series at their original scale, opening the door to the early detection of isolated critical events such as fever spikes or abrupt drops in intake at a specific time in a given animal. As shown in FIG. 8, the temperature of Cow-02 fluctuates mainly between 38°C and 40°C throughout the time series. The red dots represent anomalies identified by the algorithm, associated with both elevated temperature spikes and atypical multivariable patterns detected through the combined analysis of physiological and behavioral variables.

All the results of this study not only validated the use of classification and clustering algorithms for the management of physiological and feeding behavior data, but also highlighted the importance of integrating multiple ML techniques for a more robust

TABLE VI
Descriptive statistics and anomaly detection per bovine

Animal name	Temp mean	Temp var	Pulse mean	Pulse var	Feeding mean	var	Anomaly
Cow-1	38.707	0.209	84.238	60.268	0.299	0.210	0
Cow-2	38.807	0.189	86.206	51.194	0.317	0.216	1
Cow-3	38.699	0.146	85.889	41.862	0.270	0.197	1
Cow-4	38.713	0.210	84.900	55.166	0.290	0.206	0
Cow-5	38.650	0.187	86.684	52.875	0.364	0.231	1
Cow-6	38.711	0.169	86.356	53.378	0.247	0.186	0
Cow-7	38.655	0.184	87.215	56.801	0.265	0.195	0
Cow-8	38.650	0.167	86.061	45.595	0.254	0.190	0
Cow-9	38.731	0.205	86.360	56.690	0.309	0.213	0
Cow-10	38.735	0.185	85.946	58.598	0.216	0.169	0
Cow-11	38.716	0.189	86.351	55.299	0.172	0.142	0
Cow-12	38.671	0.203	86.355	59.655	0.310	0.214	0
Cow-13	38.663	0.180	86.098	51.602	0.261	0.193	0
Cow-14	38.703	0.209	85.852	55.073	0.366	0.232	1
Cow-15	38.694	0.177	84.129	57.234	0.299	0.210	1
Cow-16	38.762	0.196	85.601	61.797	0.318	0.217	0
Cow-17	38.725	0.197	86.364	51.707	0.340	0.224	0
Cow-18	38.715	0.197	86.289	58.112	0.149	0.127	1
Cow-19	38.691	0.222	86.034	57.305	0.325	0.219	0
Cow-20	38.738	0.178	85.997	52.241	0.165	0.138	1
Cow-21	38.713	0.199	86.214	49.252	0.327	0.220	0
Cow-22	38.680	0.177	86.099	55.981	0.321	0.218	0
Cow-23	38.751	0.165	85.639	46.871	0.231	0.178	1
Cow-24	38.640	0.170	85.850	55.724	0.312	0.215	0
Cow-25	38.695	0.174	85.333	54.443	0.237	0.181	0
Cow-26	38.661	0.202	85.453	51.804	0.273	0.198	0
Cow-27	38.682	0.189	85.957	54.079	0.173	0.143	0
Cow-28	38.745	0.237	86.396	62.218	0.163	0.136	1
Cow-29	38.631	0.200	85.606	56.800	0.245	0.185	0
Cow-30	38.743	0.200	87.334	65.342	0.252	0.188	1
Cow-31	38.769	0.230	86.108	64.184	0.189	0.153	1
Cow-32	38.665	0.219	86.180	60.958	0.180	0.148	1
Cow-33	38.670	0.158	86.398	50.879	0.317	0.216	0
Cow-34	38.672	0.172	86.164	51.059	0.226	0.175	0
Cow-35	38.704	0.202	86.886	51.126	0.307	0.213	0
Cow-36	38.703	0.201	85.725	48.419	0.194	0.156	0
Cow-37	38.657	0.214	86.429	60.866	0.248	0.187	0
Cow-38	38.646	0.216	86.280	54.641	0.228	0.176	0
Cow-39	38.721	0.202	85.918	49.521	0.214	0.168	0
Cow-40	38.667	0.195	86.485	47.798	0.196	0.158	0

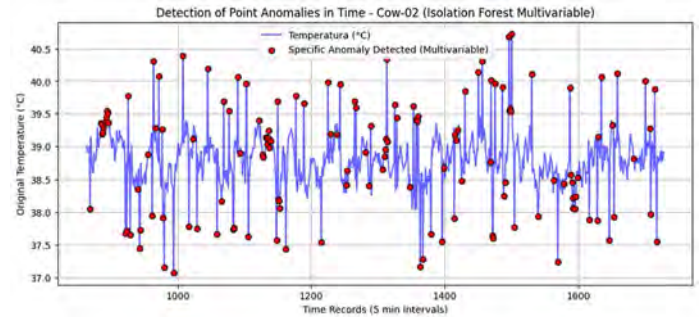


FIGURE 8. Detection of temperature anomalies of the COW 02

and complementary assessment of animal status. This integration is particularly relevant in real production environments where the availability of clinical labels is limited and data exhibit variability inherent to the biological behavior of animals. However, it is important to recognize the use of simulated data as a methodological limitation; under real production scenarios, the performance and classification accuracy would experience a decrease due to unforeseen environmental factors, noise in the transmission of IoT sensors and the profound biological complexity of free-range cattle.

CONCLUSIONS

This research enabled the implementation and evaluation of a machine learning-based approach for diagnosing physiological and feeding behavior data in cattle, demonstrating the significant potential of these AI models for the early detection of animal health issues. The implementation of machine learning models enabled the effective classification of cattle health status using a hybrid approach based on machine learning for intelligent monitoring of bovine health and behavior.

The importance of the supervised Random Forest model lies in its high overall robustness in discriminating physiological states across more than ten thousand records evaluated, achieving an overall accuracy of 85%. Its main strength lies in confirming the stability of healthy animals, supported by a sensitivity of 96% in the normal class. However, the moderate sensitivity of 50% in classifying abnormal conditions highlights that simulated noise and overlapping variables impose limitations that must be addressed with combined approaches, in order to avoid relying on a single algorithm to mitigate false negatives in early warnings.

Meanwhile, the unsupervised approaches provided a complementary and highly valuable diagnostic tool by analyzing 100% of the dataset of forty cows at the individual level. The implementation of K-means demonstrated that blind clustering is capable of reliably isolating variations in feeding behavior, achieving a structure that evenly divided the group into twenty animals with high feeding rates and twenty with reduced activity. This macro-segmentation was optimally refined by the Isolation Forest algorithm, which operated in a more restrictive and multidimensional manner to specifically identify twelve cows as abnormal and classify the remaining twenty – eight as normal.

The most valuable finding of this technical integration was the exact agreement between the two unsupervised methods in identifying

the same seven cows with the most critical profiles in the dataset. While K-means grouped generally based on similarities in feed intake, Isolation Forest combined thermal and heart rate deviations with feeding behavior to perform a much more selective and specific filtering. These results confirm that the use of machine learning is a suitable alternative for intelligent monitoring in livestock production systems, as it contributes to informed decision-making and the optimization of livestock management in this economic sector.

Conflict of interest

The authors declare no potential conflicts of interest.

Ethical considerations

The research was based exclusively on synthetic data generated by computer simulation, without involving direct experimentation with animals or the collection of real biological data. For this reason, approval from an ethics committee on the use of animals was not necessary. However, the physiological ranges used for the simulation were based on values reported in scientific literature, ensuring the biological consistency of the data generated.

BIBLIOGRAPHIC REFERENCES

- [1] Morrone S, Dimauro C, Gambella F, Cappai MG. Industry 4.0 and precision livestock farming (PLF): an up-to-date overview across animal productions. *Sensors* [Internet]. 2022; 22(12):4319. doi: <https://doi.org/rf74>
- [2] Grzesiak W, Zaborski D, Pluciński M, Jędrzejczak-Silicka M, Pilarczyk R, Sablik P. The use of selected machine learning methods in dairy cattle farming: a review. *Animals* [Internet]. 2025; 15(14):2033. doi: <https://doi.org/rf8d>
- [3] Prieto-Luna JC, Alarcón-Sucasaca A, Fernández-Romero V, Turpo-Galeano YH, Delgado-Berrocal YR, Holgado-Apaza LA. Automated monitoring system for estrus signs in cattle using precision livestock farming with IoT technology in the Peruvian Amazon. *Rev. Cient. Sist. Inform.* [Internet]. 2025; 5(1):e837. doi: <https://doi.org/rf8m>
- [4] Kraft M, Bernhardt H, Brunsch R, Büscher W, Colangelo E, Graf H, Marquering J, Tapken H, Toppel K, Westerkamp C, Ziron M. Can livestock farming benefit from Industry 4.0 technology? Evidence from recent study. *Appl. Sci.* [Internet]. 2022; 12(24):12844. doi: <https://doi.org/g9qdfp>
- [5] Higaki S, Noronha de Andrade-Freitas E, Negreiro A, Dórea JRR, Cabrera VE. Leveraging unsupervised machine learning techniques for detecting outliers in the daily milk yield data of dairy cows. *J. Dairy Sci.* [Internet]. 2025; 108(9):9696–9711. doi: <https://doi.org/rf8r>
- [6] Hossain ME, Kabir MA, Zheng L, Swain DL, McGrath S, Medway J. A systematic review of machine learning techniques for cattle identification: datasets, methods and future directions. *Artif. Intell. Agric.* [Internet]. 2022; 6:138–155. doi: <https://doi.org/grg35z>
- [7] Chelotti JO, Martínez-Rau L, Ferrero M, Vignolo L, Galli J, Planisich A, Rufiner HL, Giovanini L. Livestock feeding behaviour: A review on automated systems for ruminant monitoring. *Biosystems Eng.* [Internet]. 2024; 246:150–177. doi: <https://doi.org/gt6x7v>
- [8] Michelena Á, Díaz-Longueira A, Novais P, Simić D, Fontenla-Romero Ó, Calvo-Rolle J. Comparative analysis of unsupervised anomaly detection techniques for heat detection in dairy cattle. *Neurocomputing* [Internet]. 2024; 618:129088. doi: <https://doi.org/rf8w>
- [9] Idris M, Uddin J, Sullivan M, McNeill DM, Phillips CJC. Non-invasive physiological indicators of heat stress in cattle. *Animals* [Internet]. 2021; 11(1):71. doi: <https://doi.org/g95qcd>
- [10] Indarjulianto S, Nururrozi A, Datrianto DS, Fen TY, Priyo Jr TW, Setyawan EMN. Physiology value of breath, pulse and body temperature of cattle. *BIO Web Conf.* [Internet]. 2022; 49(2):01007. doi: <https://doi.org/rf8x>
- [11] King MTM, Dancy KM, LeBlanc SJ, Pajor EA, DeVries TJ. Deviations in behavior and productivity data before diagnosis of health disorders in cows milked with an automated system. *J. Dairy Sci.* [Internet]. 2017; 100(10):8358–8371. doi: <https://doi.org/gbz7t9>
- [12] Beauchemin KA. Invited review: current perspectives on eating and rumination activity in dairy cows. *J. Dairy Sci.* [Internet]. 2018; 101(6):4762–4784. doi: <https://doi.org/gdmrbs>
- [13] Codl R, Ducháček J, Stádník L, Gašparík M. Behavioral indicators of mastitis: feeding-to-rumination ratio as a predictive tool in Holstein–Czech Fleckvieh crossbreeds. *Cogent Food Agric.* [Internet]. 2026; 12(1):2629620. doi: <https://doi.org/rgc5>
- [14] Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. *Behav. Ther.* [Internet]. 2020; 51(5):675–687. doi: <https://doi.org/gm2xrd>
- [15] Kuo CT, Xu D, Friesen R. A brief review of unsupervised machine learning algorithms in astronomy: dimensionality reduction and clustering. *Universe* [Internet]. 2025; 11(12):412. doi: <https://doi.org/rgc6>
- [16] Breiman L. Random forests. *Mach. Learn.* [Internet]. 2001; 45(1):5–32. doi: <https://doi.org/d8zjwq>
- [17] Viteri-Guzmán G, Molineros-Meza R, Oyarzún-Soares E. Impact on cattle welfare with IoT technology and statistical data analytics. *E3S Web of Conferences* [Internet]. 2025; 658:03005. doi: <https://doi.org/rgc7>
- [18] Sun D, Webb L, van der Tol PPJ, van Reenen K. A systematic review of automatic health monitoring in calves: glimpsing the future from current practice. *Front. Vet. Sci.* [Internet]. 2021; 8:761468. doi: <https://doi.org/rgc8>
- [19] Girdauskaitė A, Grigė S, Džermeikaitė K, Krištolaitytė J, Malašauskienė D, Televičius M, Šertvytytė G, Lembovičiūtė G, Antanaitis R. Supervised machine learning approaches for early detection of metabolic and udder health disorders in dairy cows using sensor-derived data. *Front. Vet. Sci.* [Internet]. 2025; 12:1726719. doi: <https://doi.org/rgc9>
- [20] Brown W, Cavani L, Peñagaricano F, Weigel K, White HM. Feeding behavior parameters and temporal patterns in mid-lactation Holstein cows across a range of residual feed intake values. *J. Dairy Sci.* [Internet]. 2022; 105(10):8130–8142. doi: <https://doi.org/rgdb>