

Performance of multiple models under different test strategies in predicting the ingestive behavior of grazing cattle

Lázaro Henrique da Silva¹ , Caio Matheus Leite da Silva² , Erick Galani Maziero² , Daniel Rume Casagrande¹ , Marina de Arruda Camargo Danes^{1*} 

¹ Universidade Federal de Lavras, Departamento de Zootecnia, Lavras, MG, Brasil.

² Universidade Federal de Lavras, Departamento de Automática, Lavras, MG, Brasil.

*Corresponding author:
marina.danes@ufla.br

Received: September 29, 2025

Accepted: March 11, 2026

How to cite: Silva, L. H.; Silva, C. M. L.; Maziero, E. G.; Casagrande, D. R. and Danes, M. A. C. 2026. Performance of multiple models under different test strategies in predicting the ingestive behavior of grazing cattle. *Revista Brasileira de Zootecnia* 55:e20250201.
<https://doi.org/10.37496/rbz5520250201>

Editors:

Anderson Antonio Carvalho Alves
Ana Clara Baião Menezes

Copyright: This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



ABSTRACT - The use of digital technologies and wearable sensors has advanced in precision livestock farming. However, there is still no clear definition of the most suitable models for the use of sensors in grazing animals, especially under tropical systems. Thus, we aimed to compare the performance of six predictive models for grazing cattle behavior and to evaluate different testing strategies. The study was conducted in a pasture area, using nine Tabapuã heifers monitored by triaxial accelerometers attached to the nape, with data recorded every second. Ninety-six 12-h visual observations were performed and combined with sensor data to generate two datasets: grazing vs non-grazing (GNG) and grazing, ruminating, and idling (GRI). The models tested were generalized linear model (GLM), random forest (RF), k-nearest neighbors (KNN), gradient boosting (GB), light gradient boosting (LGB), and artificial neural network (ANN), with hyperparameter optimization performed via Bayesian search. Seven testing strategies were applied, including random holdout (RHO), leave-animal-out (LAO), leave-height-out (LHO), with four different sward heights, and an external test, obtained in a previous study. All models, except GLM, showed good performance (~80% accuracy for GNG); however, the inclusion of three classes reduced the average accuracy by 6.4%. LGB and KNN were the most computationally efficient models, in terms of data process time, while ANN was the most demanding. Testing strategies that reduced data dependence (LAO and LHO) decreased accuracy, highlighting the importance of real-world scenarios for the development of robust commercial applications. It is recommended to balance accuracy and computational cost, prioritizing LGB, GB, or KNN for practical applications. Moreover, testing strategies under different farm conditions or using external datasets should be performed.

Keywords: animal welfare, machine learning, precision livestock, sensor, smart farming

1. Introduction

In recent years, there has been notable progress in developing decision-making tools based on wearable technology and data science within animal farming (Kaur et al., 2023). Automation of data collection allows for individual animal monitoring even in large herds. This approach may not only enhance production and reproduction efficiency (Ribeiro et al., 2021; Romanzini et al., 2022), enable the early detection of health events, as proposed by Yigit et al. (2022), who sought to detect laminitis in horses using wearable sensors, or Abeni and Galli (2017), who used rumination and activity data for

the early detection of anaplasmosis in dairy calves, as well as even Teixeira et al. (2022), who monitored the activity and rumination time of dairy cows for the early detection of heat stress.

Despite the use of digital technologies being extensively studied in confined animals, the use of sensors in grazing animals is still incipient (Park and Park, 2021). Studies have been conducted in temperate pastures (Shafiullah et al., 2019; Wang et al., 2021) and there are few studies on tropical pastures (Watanabe et al., 2021; Romanzini et al., 2022). An aggravating factor for studies in tropical pastures is the large diversity of available forages and animal species, requiring robust testing strategies (Ribeiro et al., 2021). Additionally, grazing management practices change the sward structure and, therefore, affect animal behavior. This change in grazing behavior as sward structure and forage allowance vary can be used to support management decisions in grazing systems. Additionally, Romanzini et al. (2022) demonstrated that it is possible to forecast animal performance from animal behavior predicted with a 3-axis accelerometer in tropical grazing conditions. Therefore, there is a vast potential for sensor-supported tools in grazing systems.

However, a multitude of machine learning algorithms have been used in these studies. In addition, an even greater number of models are available in digital repositories, and they need to be evaluated according to the goals of the tool to be developed, as well as the characteristics of the dataset. Studies comparing different predictive models are not new (Sagi and Rokach, 2018; Warner et al., 2020; Li et al., 2022); however, few have done so in tropical pastures.

Additionally, to develop a commercially available tool, one needs to consider the testing strategies used when evaluating the models. The k-fold cross-validation strategy is the most commonly used to validate models that aim to predict animal behavior (Riaboff et al., 2019). However, Ribeiro et al. (2021) demonstrated that this strategy can inflate model performance, as its randomness does not take into account the biological interdependence animals, plants, and the environment. Therefore, the authors proposed the use of different k-fold strategies that mimic real-world situations on farms.

The objectives of this study were: 1) to compare the performance of six different models to predict behavior in grazing animals; 2) to evaluate the effect of the number of predicted behavioral classes on the quality of the predictions; and 3) to compare multiple testing strategies considering the biological relevance of the dataset and the goals of the tool.

2. Material and methods

Research on animals was conducted according to the ethics committee on animal use of the Universidade Federal de Lavras (protocol number 026/2019).

2.1. Data collection

The experiment was conducted at the experimental farm of the Universidade Federal de Lavras, from November 2019 to July 2020. It was part of another experiment evaluating grazing intensity in an intermittent grazing stocking system in a combined pasture of *Urochloa brizantha* (Hochst ex A. Rich) Stapf cv. Marandu and *Arachis pintoii* Krap. & Greg. cv. Mandobi (Rodrigues da Cruz et al., 2024). The animals had access to the paddocks when the sward reached 25 cm and were removed when it was grazed down to 20, 15, or 10 cm. Therefore, we were able to observe animal behavior in different sward structures. We conducted 96 12-h observation periods (approximately 6:00 to 18:00), on 48 non-consecutive days during three seasons (32 during spring, 34 during summer, and 30 during autumn). Each day of data collection involved two animals, ensuring two experimental units per sampling. Observations took place during the first day of grazing in a new paddock (25 cm, 47 observations) and during the last day before animals left the paddock, when sward height was approaching 20 cm (16 observations), 15 cm (13 observations) and 10 cm (20 observations).

Nine Tabapuã heifers with a mean initial body weight of 185 ± 17 kg were used in the experiment. However, in each observation day, two of them, randomly chosen, were used to graze the paddock.

Animals were painted with a non-toxic ink in the thoracic region to facilitate visual observation. Each 3-h shift of visual observation was carried out by one of ten trained observers. The animals were observed continuously and, with the aid of a wristwatch, the start time (hour, minute, and second) of each activity was registered. Animal behavior was recorded on a printed spreadsheet, always noting the exact moment when the animal changed its behavior. The spreadsheet contained columns with information on the date, animal identification, time of the behavioral change, and the new behavior observed. The description of each behavior is presented in Table 1. Water drinking activities and atypical behaviors (disputes for example) were removed from the final dataset.

Table 1 - Description of the animal behaviors observed in the study

Behavior	Description
Grazing	Animals apprehending food with the head down and chewing either with the head down or up in between apprehension bites.
Ruminating	Animals chewing or swallowing a ruminal bolus, standing or lying down.
Idle	Animals at rest, standing or lying down, without chewing movement.
Non-grazing	Sum of ruminating and idle behaviors.

A 3-axes (X, Y, and Z) wireless accelerometer sensor (Sensor Spotlight, Accelerometers, Monnit Corporation, South Salt Lake, UT, USA) set to measure from -2 to $+2$ g-force and powered by a coin cell battery (model CR2032 of 3.0 voltage) was attached to the halter on the back of each animal's neck, which is the most common location (Chelotti et al., 2024). The X, Y, and Z axes indicate longitudinal (front-to-back), horizontal (side-to-side), and vertical (up-to-down) head movements, respectively. Devices were set to send the raw data (g-force for X, Y, and Z) to a local storage system for each animal every second, equivalent to 1 Hz. The data capture rate was similar to the data transmission rate. Prior to the beginning of the observations, we synchronized the clocks of all the sensors with the computer clock and the wristwatch used by the observers. The iMonnit software (MONNIT, Salt Lake City, Utah, United States of America), which is compatible with the sensors, was used to receive and export the raw data in CSV format.

2.2. Predictive models and dataset

Data from visual observations and sensor recordings were combined and cleaned (i.e., missing or mismatched data were removed) using Python (Python Software Foundation, Delaware, United States of America) and the Jupyter IDE. The data were considered missing or inconsistent in two different scenarios: 1) when the data capture moment showed values obtained by the sensors but did not include the corresponding visual observation of behavior; 2) when the data capture moment did not show sensor values but included visual observation records of behavior. The data received from the sensors included date, animal ID, time, and acceleration along the X-, Y-, and Z-axes. In contrast, the behavioral data included date, animal ID, time, and observed behavior. The date, animal ID, and time were used to combine both datasets, resulting in a single dataset with information on date, time, acceleration along the X-, Y-, and Z-axes, and observed behavior.

Next, we constructed two different datasets: The first, called GNG, had only two behavior classes: "Grazing" and "Non grazing". The "Non grazing" class resulted from combining "Ruminating" and "Idle". The second, called GRI, included the three observed behaviors: "Grazing", "Ruminating" and "Idle".

Both datasets were analyzed by using six predictive models, aiming to evaluate a wide range of tools. The GLM was chosen as a simpler option, whereas Random Forest (RF), Gradient Boosting (GB), and k-Nearest Neighbors (KNN) are the three algorithms most commonly used to predict animal behavior (Valletta et al., 2017; Sagi and Rokach, 2018; Chelotti et al., 2024). Light Gradient Boosting (LGB) is an enhancement of GB, developed by Microsoft Research (Microsoft, Redmond, Washington, United States of America), and Artificial Neural Network (ANN) was chosen as a more complex model. The GLM is a

statistical technique for generalizing linear models, whereas predictive models RF, GB, LGB, KNN and ANN are classified as machine learning algorithms.

Furthermore, all predictive models were subjected to different testing strategies. Thus, the GRI dataset was subjected to six different predictive models, and each predictive model was subjected to six different testing strategies. On the other hand, the GNG dataset was subjected to six different predictive models, and each one was subjected to seven different testing strategies. All model training was implemented using the Python programming language.

Both datasets were divided into training, validation, and test sets. The amount of data reserved for training and testing varied according to the testing strategy adopted. For hyperparameter tuning, 20% of the data was always set aside for validation, using data previously allocated for training.

For the GLM model, we used the equation called multinomial logistic regression (Böhning, 1992), which generalizes logistic regression to multiclass problems. We used as parameters the Newton-Raphson method, the maximum number of interactions of 35 and the full output "True". In this study, we used a logit link function. The penalties were associated with ridge regression (ℓ_2). The lambda parameter and the C parameter, i.e., regularization factor, were selected through random search, being defined as 11.99 and 0.08, respectively. These metrics were used for GRI and GNG.

For the other models, the best hyperparameters for each dataset were identified. The search for the best hyperparameter was performed using the Bayesian optimization technique (Mockus, 1975). The parameters provided for the search for RF were: number of trees (50, 250), maximum number of features RF ("auto" or "sqrt"), maximum depth (0, 150), function to measure the quality of a split ("gini" or "entropy"), minimum number of samples in an internal node (2, 50), minimum number of samples in a leaf node (1 or 2) and bootstrapping ("True" or "False").

The search for the best hyperparameter for GB was performed using the parameters learning rate (log from 0.001 to log from 0.1), maximum depth (2 to 16), minimum child weight (1 to 16), gamma (0.1, 0.5) and col sample by tree (0.1, 1.0). The search for best hyperparameter for LGB used as parameters the learning rate (log from 0.001 to log 0.1), the number of leaves (2 to 512), the minimum weight of the child (1 to 500), the subsample (0.05, 1.0) and sample per tree (0.1, 1.0).

The search for the best hyperparameter for KNN was performed with the parameters number of neighbors (5 to 50), weights ("Uniform" or "Distance"), leaf size (20, 100) and p (1 or 2). Finally, the search for the best hyperparameter for ANN was performed using the parameter hidden layer sizes (20, 30 and 30, 20). We defined the activation function as "tanh" (hyperbolic tangent), set verbose to False, and random state to 0. The model included only one hidden layer.

2.3. Test

All predictive models, for the GNG dataset, were evaluated with seven different testing strategies. The RHO, using 20% of the data were randomly excluded and used for test and the remaining 80% of the data were used for training. The second test strategy was the LAO, in which all data from one animal were removed and used for testing, while data from other animals were used for training. This strategy aimed to evaluate the predictive performance of the models when applied to new animals, that did not contribute data to model training. The third, fourth, fifth, and sixth strategy were LHO, to simulate the performance of the models when a new sward structure is presented, different from the ones used to train them. For LHO, all data from one sward height (25, 20, 15, or 10 cm) were removed and used for test, while data from the other heights were used for model training. Finally, the seventh strategy was an external test (EV), using a dataset from a previous study by our group (Ribeiro et al., 2021) for testing, while the full dataset generated in the present study was used for model training. The EV strategy was intended to challenge predictive models to a completely new situation, without any biological interdependence between animal, plant and the environment with the training dataset. For the GRI dataset, the EV strategy was not used, because the external dataset had only two classes of animal behaviors.

2.4. Model performance evaluation

To evaluate the performance of the predictive models, a confusion matrix was generated, with the values of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). From these values we calculated accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predicted value (NPV), using the following equations:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Furthermore, the data process time (DPT) was also measured, which is an estimate of the real time elapsed for training the predictive models. Tests and generation of metrics used to evaluate predictive models were performed using the Python language.

3. Results

3.1. Model performance

The 96 periods of 12 hours of observations, collecting sensor data every second, should have generated 2,073,600 data points. However, there were data transmission problems, maybe due to the wireless network available, that caused failures in data reception from the sensors, and the final dataset contained 607,150 data points, equivalent to 29.28% of the data. Despite the substantial loss of data, our final dataset is still greater than what is reported in the literature for similar studies (Decandia et al., 2018; Ribeiro et al., 2021; Riaboff et al., 2022).

The GNG dataset had 54.8% of the datapoint as grazing behavior and 45.2% as non-grazing and, therefore, was well balanced (Table 2). On the other hand, splitting the data in three categories created some imbalance. The GRI dataset had 54.8% of the data points classified as grazing, 24.9% as ruminating and 20.3% as idle.

Table 2 - Distribution of data points within the evaluated animal behaviors and the percentage of total

Dataset	Grazing	Rumination	Idle	Not grazing
GRI	332,512 (54.8%)	151,381 (24.9%)	123,257 (20.3%)	-
GNG	332,512 (54.8%)	-	-	274,638 (45.2%)

GRI - grazing, ruminating, idle; GNG - grazing, non-grazing.

Table 3 - Predictive performance and data processing times of models used with the GNG dataset, and validated with a random holdout strategy

Predictive model	Category	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	DPT (s)
GLM	Grazing	78.2	44.8	63.3	62.7	63.1	001
	Non-grazing	44.8	78.2	62.7	63.3		
RF	Grazing	85.1	74.0	80.0	80.3	80.1	069
	Non-grazing	74.0	85.1	80.4	80.0		
KNN	Grazing	84.2	76.8	81.6	80.1	80.9	005
	Non-grazing	76.8	84.2	80.0	81.5		
GB	Grazing	85.0	73.9	79.9	80.2	80.0	059
	Non-grazing	73.9	85.0	80.2	79.8		
LGB	Grazing	84.6	73.9	79.8	79.8	79.9	003
	Non-grazing	73.9	84.6	79.8	79.8		
ANN	Grazing	83.3	74.2	79.7	78.5	79.2	300
	Non-grazing	74.2	83.3	78.5	79.7		

GNG - grazing vs non-grazing dataset; PPV - positive predictive values; NPV - negative predictive values; DPT - data process time; GLM - generalized linear model; RF - random forest; KNN, k-nearest neighbors; GB - gradient boosting; LGB - light gradient boosting; ANN - artificial neural network.

The worst accuracy was observed for GLM, which also had the highest difficulty to predict the non-grazing behavior, with a very low sensitivity for this behavior (Table 3). All the other models presented similar performance, with a mean ($\pm$ SD) of 80.0% ($\pm$ 0.61) for accuracy, 84.4% ($\pm$ 0.73) for sensitivity, 74.6% ($\pm$ 1.26) for specificity, 80.2% ($\pm$ 0.79) for PPV and 79.8% ($\pm$ 0.74) for NPV.

As all predictive models, except the GLM, had similar behavior, the choice should also consider the computational power requirement, measured through the DPT. The LGB and KNN models were the lightest, with processing being performed in 3 and 5 seconds respectively. The model that required the most computational power was the ANN, with a processing time of 300 seconds.

Similar results was observed for the GRI dataset (Table 4), also using the holdout test. The LGB and KNN models required the least computational power, while the ANN model required the most computational power to perform the predictions. The GLM predictive model was the most inefficient, whereas the other models performed similarly. For example, the GLM model achieved an accuracy of 56.9%, whereas the other models presented a mean accuracy ($\pm$ SD) of $73.6 \pm 1.24\%$.

The GRI dataset revealed a pronounced reduction in predictive accuracy for idle behavior using the holdout test. Excluding the GLM model, the mean sensitivity reached 88.9% ($\pm$ 0.57) for grazing, 66.0% ($\pm$ 1.97) for ruminating, and only 44.0% ($\pm$ 4.29) for idle, indicating a marked difficulty in capturing this behavior. Similarly, the mean PPV for grazing, excluding GLM, was 77.3% ($\pm$ 1.20), for ruminating 69.5% ($\pm$ 1.54), and for idle 63.5% ($\pm$ 2.70), underscoring the reduced precision in identifying idle events. NPV values were relatively consistent across all behaviors, with a slight advantage observed for grazing, suggesting a balanced capacity of the models to correctly classify negative instances.

Our second objective was to evaluate the effect of the number of predicted classes on the accuracy of the models. Table 5 presents the accuracy of all models for the two datasets (GNG vs GRI). Predicting three behavior classes instead of two behaviors decreased the accuracy of prediction by an average of 6.40% units. This difference between the two datasets was similar across predictive models (varied from 5.70 to 7.70% units), indicating that no model better handled the increase in the number of classes.

Table 4 - Predictive performance and data processing times of models used with the GRI dataset, and validated with a random holdout strategy

Predictive model	Category	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	DPT (s)
GLM	Grazing	91.6	23.8	59.3	70.0	56.9	003
	Ruminating	18.7	93.4	48.4	77.6		
	Idle	10.1	95.4	35.9	80.6		
RF	Grazing	88.9	68.9	77.6	83.7	74.1	075
	Ruminating	66.3	90.8	70.4	89.0		
	Idle	44.0	93.9	64.8	86.8		
KNN	Grazing	87.9	71.6	78.9	83.1	74.6	010
	Ruminating	68.3	90.3	70.1	89.6		
	Idle	46.5	93.4	64.3	87.3		
GB	Grazing	89.4	68.6	77.5	84.2	74.3	174
	Ruminating	66.7	90.7	70.5	89.1		
	Idle	42.9	94.3	65.7	86.6		
LGB	Grazing	89.1	67.2	76.9	80.0	73.5	009
	Ruminating	65.7	90.2	69.6	88.8		
	Idle	41.0	94.4	63.8	86.0		
ANN	Grazing	89.0	65.3	75.6	83.0	71.5	392
	Ruminating	62.9	89.6	66.8	87.9		
	Idle	35.1	93.7	58.8	85.0		

GRI - grazing vs ruminating vs idle dataset; PPV - positive predictive values; NPV - negative predictive values; DPT - data process time; GLM - generalized linear model; RF - random forest; KNN - k-nearest neighbors; GB - gradient boosting; LGB - light gradient boosting; ANN - artificial neural network.

Table 5 - Accuracy of models when predicting two (GNG) or three (GRI) classes

Dataset	GLM (%)	KNN (%)	RF (%)	GB (%)	LGB (%)	ANN (%)
GNG	63.1	80.9	80.1	80.0	79.9	79.2
GRI	56.9	74.6	74.1	74.3	73.5	71.5
Diference	6.20	6.30	6.00	5.70	6.40	7.70

GNG - grazing vs not grazing dataset; GRI - grazing vs rumination vs idle dataset; GLM - generalized linear model; RF - random forest; KNN - k-nearest neighbors; GB - gradient boosting; LGB - light gradient boosting; ANN - artificial neural network.

3.2. Testing strategies

The accuracies of predictive models with the different testing strategies are presented in Tables 6 (GNG dataset) and 7 (GRI dataset). The RHO strategy resulted in the greatest accuracy for all predictive models. The strategies that reduce data interdependence and the carry-over effects from animals and sward structure (LAO and LHO) resulted in intermediate values, with LHO presenting, on average, greater accuracy than LAO. The difference in accuracy among these three strategies was much smaller for GLM than for the other models. On the other hand, the difference between RHO and LAO or LHO was greater for the GRI dataset than for GNG. As expected, using an external data set to validate the models resulted in the lowest accuracies for all models and in both datasets. The drop in accuracy from the RHO strategy to the EV strategy varied from 2.8% unit for GLM in GRI to 28.9% units for KNN in GNG.

Table 6 - Effect of the test strategy on the accuracy of predictive models with a two classes data set (GNG)

Test	GLM (%)	RF (%)	KNN (%)	GB (%)	LGB (%)	ANN (%)
Random holdout (20%)	61.1	80.1	80.9	80.0	76.7	79.2
Leave animals out	60.3	59.4	60.3	59.8	60.0	60.8
Leave height out 10 cm	54.6	60.0	59.5	60.3	60.1	59.7
Leave height out 15 cm	69.4	64.5	64.3	64.6	64.6	65.0
Leave height out 20 cm	59.5	64.1	64.1	64.3	64.6	62.2
Leave height out 25 cm	59.5	64.1	59.5	64.3	64.6	65.2
External test	43.3	56.8	52.0	55.9	56.7	53.9

GLM - generalized linear model; RF - random forest; KNN - k-nearest neighbors; GB - gradient boosting; LGB - light gradient boosting; ANN - artificial neural network.

Table 7 - Effect of the test strategy on the accuracy of predictive models with a three-class data set (GRI)

Test	GLM (%)	RF (%)	KNN (%)	GB (%)	LGB (%)	ANN (%)
Random holdout (20%)	56.9	74.1	74.6	74.3	73.5	71.5
Leave animals out	54.8	52.1	51.4	52.7	52.6	52.2
Leave height out 10 cm	46.6	53.8	53.1	53.9	54.1	53.4
Leave height out 15 cm	65.1	60.3	59.2	60.6	60.8	59.9
Leave height out 20 cm	52.5	52.9	53.1	52.9	54.1	53.1
Leave height out 25 cm	54.1	53.7	52.9	54.0	54.3	53.6

GLM - generalized linear model; RF - random forest; KNN - k-nearest neighbors; GB - gradient boosting; LGB - light gradient boosting; ANN - artificial neural network.

4. Discussion

The study described herein was part of a larger goal of developing tools to support decision-making for grazing management through automation and data science. It is well recognized that grazing behavior changes as animals graze down rotationally managed swards, to the point that forage intake is negatively affected (Carvalho et al., 2009; Rodrigues da Cruz et al., 2024). Therefore, if these changes could be detected with wearable sensors, the ideal moment to remove animals from a paddock could be predicted.

Studies designed to develop predictive models of animal behavior have mostly focused on confined animals and have used data from a specific feeding situation and for short periods (Abeni and Galli, 2017; Cairo et al., 2020). However, grazing animals face challenges related to the constantly changing sward structure and the time budget required to harvest enough forage. Predicting ingestive behavior under these circumstances might be substantially more complex, requiring different machine learning techniques (Chelotti et al., 2024). Thus, we monitored nine grazing animals for eight months and throughout the reduction of canopy height in an intermittent grazing stock system to capture a wide range of feeding scenarios and evaluate predictive models varying in complexity.

Our results showed that a simpler model such as GLM was not capable of adequately classifying either two or three classes of behavior. Usually, a linear model is included to evaluate the simplest possible solution to a problem. However, the inadequacy of this tool to predict animal behavior has often been demonstrated (Cairo et al., 2020; Ribeiro et al., 2021; Warner et al., 2020). On the other hand, the most complex model evaluated, ANN, did not outperform the other algorithms and required more computational power to make predictions. Even though this algorithm was considered an appropriate model for predicting animal behavior based on three-axis accelerometer data (Nadimi et al., 2012), others have reported findings similar to ours, with no superiority of ANN to in handling animal data such as ingestive behavior (Ribeiro et al., 2021).

Surprisingly, despite the greater complexity of our dataset (varying seasons, sward structures and animals), the machine learning algorithms evaluated performed very similarly. Both RF and GB are ensemble learning methods and presented very similar results, including DPT. The LGB is an improvement of the GB that requires less computational power. Finally, KNN uses an approximation method, that despite being simpler than RF, GB and ANN, performed similarly. Even though many models perform adequately, RF often presents slightly superior predictive metrics in animal studies, such as salt-licking behavior (Simanungkalit et al., 2021), lameness (Warner et al., 2020) and grazing (Ribeiro et al., 2021). Romanzini et al. (2022) also used a three-axis accelerometer to predict grazing behavior in beef cattle in tropical pasture and evaluated three algorithms (RF, convolutional neural network, and linear discriminant analysis). The authors reported a distinct superiority in accuracy for RF (82.1% vs an average of 61.1% for the other two models).

On average, the RHO strategy resulted in an accuracy of 80% in the GNG dataset. This is only slightly greater than the values reported by Ribeiro et al. (2021), even though our dataset contained more than 607,000 records, while the authors worked with approximately 80,000 records. Additionally, the values are lower than those reported in some studies in which dairy cow behavior was predicted with accelerometers (Dutta et al., 2015; Riaboff et al., 2019). However, these results came from confined animals, which may present more consistent feeding behavior. On the other hand, Wang et al. (2025) evaluated the prediction of four classes of animal behavior in grazing beef cattle with three-axis accelerometers and machine learning models, reporting accuracies between 91 and 95%, with a substantially smaller dataset (22,987 records collected over 197 min from six beef cows). It is possible that a smaller dataset, collected over a short period of animal observations, resulted in less complexity than that captured in the present study. However, there might also be an effect of data processing. Additionally, they applied downsampling using a sliding window method and computed the mean, minimum, maximum, and variance for each instantaneous feature, generating 28 features. After removing some of the features due to collinearity, the authors ended up with 20 features to develop the model, whereas we only had three.

We also evaluated the capacity of the models to predict three classes, splitting the non-grazing behavior into rumination and idle time. Model performance was reduced when predicting 3 classes rather than two, and the size of the reduction in accuracy was similar across models (5.7 to 7.7% units). When more than two classes are predicted, multiple confusion matrices are generated to compare two classes at a time (one vs all the other classes) and the average accuracy of all matrices is reported (Andrew, 2018). Therefore, if one class is poorly predicted, it impacts the average and decreases overall model performance, as observed in our results, in which performance metrics were worse for rumination and idle behaviors compared to grazing. Hasan et al. (2016), investigating predictive models for automatic annotation of clinical text fragments, observed a marked decrease in accuracy when using RF to 41 classes (49.5%), compared to 20 (56.3%) and 17 (67%) classes.

The imbalance of data among classes may harm the prediction of underrepresented classes. In our GNG dataset, classes were more evenly distributed, with 54.8% for grazing and 45.2% for non-grazing. Conversely, in the GRI dataset, while grazing accounted for 54.8% of the records, rumination accounted for 24.9% and idle for only 20.3% of the records. The main reason for the imbalance in our dataset is that our visual observations occurred during the daytime (06:00 to 18:00 h), when animals perform most of the grazing. Rumination and idle behaviors are more frequent at night, and 24-hour observation periods would be more adequate to predict the other classes. There are strategies that can be employed to balance the dataset, such as undersampling, which randomly removes data from the majority class to balance the proportions, or oversampling, which generates synthetic data to increase the representation of minority classes. Since grazing was our most important behavior, the undersampling strategy was not a good approach. On the other hand, data generated through oversampling strategies may not accurately reflect reality and may increase the risk of overfitting, posing a significant limitation (Gosain and Sardana, 2017; Mohammed et al., 2020; Joloudari et al., 2023).

The model comparison was performed using the cross-validation strategy RHO. This approach randomly removes a portion of the data to use as a test set, while the remaining data are used for

training the model. However, as discussed by Ribeiro et al. (2021) and Wang et al. (2025), there are biological dependencies between the data in both datasets, such as the same animal or the same pasture conditions, that can inflate the accuracy of the model. Therefore, cross-validation strategies that avoid data interdependence were evaluated, also known as block cross-validation, such as removing one animal (LAO) or one pasture height (LHO) at a time to use as a test set. The results indeed demonstrated a reduction in accuracy when LAO and LHO were utilized for validation. The drop of approximately 20 percentage points (from 80% to an average of 60%) in accuracy for the GNG dataset is almost the same as reported by Ribeiro et al. (2021). These authors reported an accuracy of 76.5% when applying the RHO strategy with the RF model to predict grazing behavior, while for the LAO strategy the accuracy was 56.6%. Similarly, Wang et al. (2025) observed a decrease from 90–96% accuracy using RHO to 66–82% using LAO.

Even though more studies are adopting block cross-validation to avoid overfitting and artificially inflated accuracy, very few studies validate models using an external dataset. The EV strategy imposes an even more dramatic change in the scenario, as animals, time, location, and forage species were all different, simulating what would happen if the tool was commercially available and used in a real-world setting. For instance, while the training dataset was collected from Tabapuã heifers in an intermittent grazing system with a combined pasture of *Urochloa brizantha* and *Arachis pintoii*, the test data were collected from Nellore bulls maintained in a continuous grazing system with Marandu grass. An additional reduction of approximately 10 percentage points in accuracy for GNG demonstrates that applying the models in a new scenario remains challenging.

In addition to the external test, our animal observations were collected during a grazing management experiment evaluating the effect of three grazing intensities (20, 15, and 10 cm of post-grazing height) on sward structure, composition, nutritive value, and animal behavior (Rodrigues da Cruz et al., 2024). An interesting finding of their study is that intake rate (g of forage per min) was significantly decreased with grazing intensity, but grazing time did not differ. Thus, even if our models were better at predicting grazing behavior, allowing the calculation of grazing time, this would not have been helpful to assess the moment at which sward height starts harming intake. The behavior changes observed were in biting rate (bites/min) and grazing events. Therefore, future sensor prediction research may be more successful if focused on this type of behavior.

5. Conclusions

Based on the results, it is concluded that, although the machine learning models can adequately predict grazing behavior, this occurs primarily under random cross-validation. When data interdependence is removed through block cross-validation or external testing, model performance is considered inaccurate. Data preprocessing aimed at increasing the number of features may represent an alternative approach to improve model performance.

Data availability

The data used in the preparation of this manuscript is not available in public repositories. But it can be provided upon request by email.

Author contributions

Conceptualization: Danes, M. A. C. **Data curation:** Silva, L. H.; Silva, C. M. L. and Maziero, E. G. **Formal analysis:** Silva, L. H.; Silva, C. M. L. and Maziero, E. G. **Funding acquisition:** Casagrande, D. R. and Danes, M. A. C. **Investigation:** Silva, L. H. and Danes, M. A. C. **Methodology:** Silva, L. H.; Silva, C. M. L.; Casagrande, D. R. and Danes, M. A. C. **Project administration:** Silva, L. H. and Danes, M. A. C. **Resources:** Maziero, E. G. and Danes, M. A. C. **Supervision:** Danes, M. A. C. **Validation:** Silva, L. H.; Silva, C. M. L.; Maziero, E. G. and Danes, M. A. C. **Visualization:** Silva, L. H. and Danes, M. A. C. **Writing – original draft:** Silva, L. H. and Danes, M. A. C. **Writing – review & editing:** Danes, M. A. C.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

The authors acknowledges support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG/APQ-01869-22). We also acknowledge support from the research groups INPPAR and NEFOR.

Financial support

Financial support was provided by FAPEMIG.

Declaration of generative AI in scientific writing

The authors declare that they use generative AI tools to assist in textual transcription, accompanied by human review.

References

- Abeni, F. and Galli, A. 2017. Monitoring cow activity and rumination time for an early detection of heat stress in dairy cow. *International Journal of Biometeorology* 61:417-425. <https://doi.org/10.1007/s00484-016-1222-z>
- Andrew, N. G. 2018. Machine learning yearning. DeepLearning.ai. Palo Alto - CA.
- Böhning, D. 1992. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics* 44:197-200. <https://doi.org/10.1007/BF00048682>
- Cairo, F. C.; Pereira, L. G. R.; Campos, M. M.; Tomich, T. R.; Coelho, S. G.; Lage, C. F. A.; Fonseca, A. P.; Borges, A. M.; Alves, B. R. C. and Dorea, J. R. R. 2020. Applying machine learning techniques on feeding behavior data for early estrus detection in dairy heifers. *Computers and Electronics in Agriculture* 179:105855. <https://doi.org/10.1016/j.compag.2020.105855>
- Carvalho, P. C. F.; Trindade, J. K.; Mezzalana, J. C.; Poli, C. H. E. C.; Nabinger, C.; Genro, T. C. M. and Gonda, H. L. 2009. Do bocado ao pastoreio de precisão: compreendendo a interface planta-animal para explorar a multi-funcionalidade das pastagens. *Revista Brasileira de Zootecnia* 38:109-122. <https://doi.org/10.1590/S1516-35982009001300013>
- Chelotti, J. O.; Martinez-Rau, L. S.; Ferrero, M.; Vignolo, L. D.; Galli, J. R.; Planisich, A. M.; Rufiner, H. L. and Giovanini, L. L. 2024. Livestock feeding behaviour: A review on automated systems for ruminant monitoring. *Biosystems Engineering* 246:150-177. <https://doi.org/10.1016/j.biosystemseng.2024.08.003>
- Decandia, M.; Giovanetti, V.; Molle, G.; Acciaro, M.; Mameli, M.; Cabiddu, A.; Cossu, R.; Serra, M. G.; Manca, C.; Rassu, S. P. G. and Dimauro, C. 2018. The effect of different time epoch settings on the classification of sheep behaviour using tri-axial accelerometry. *Computers and Electronics in Agriculture* 154:112-119. <https://doi.org/10.1016/j.compag.2018.09.002>
- Dutta, R.; Smith, D.; Rawnsley, R.; Bishop-Hurley, G.; Hills, J.; Timms, T. and Henry, D. 2015. Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and Electronics in Agriculture* 111:18-28. <https://doi.org/10.1016/j.compag.2014.12.002>
- Gosain, A. and Sardana, S. 2017. Handling class imbalance problem using oversampling techniques: A review. p.79-85. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICICI). Udipi, India. <https://doi.org/10.1109/ICACCI.2017.8125820>
- Hasan, M.; Kotov, A.; Carcone, A. I.; Dong, M.; Naar, S. and Hartlieb, K. B. 2016. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of Biomedical Informatics* 62:21-31. <https://doi.org/10.1016/j.jbi.2016.05.004>
- Joloudari, J. H.; Marefat, A.; Nematollahi, M. A.; Oyelere, S. S. and Hussain, S. 2023. Effective class-imbalance learning based on SMOTE and convolutional neural networks. *Applied Sciences* 13:4006. <https://doi.org/10.3390/app13064006>
- Kaur, U.; Malacco, V. M. R.; Bai, H.; Price, T. P.; Datta, A.; Xin, L.; Sen, S.; Nawrocki, R. A.; Chiu, G.; Sundaram, S.; Min, B. C.; Daniels, K. M.; White, R. R.; Donkin, S. S.; Brito, L. F. and Voyles, R. M. 2023. Invited review: integration of technologies and systems for precision animal agriculture-a case study on precision dairy farming. *Journal of Animal Science* 101:skad206. <https://doi.org/10.1093/jas/skad206>
- Li, Y.; Shu, H.; Bindelle, J.; Xu, B.; Zhang, W.; Jin, Z.; Guo, L. and Wang, W. 2022. Classification and analysis of multiple cattle unitary behaviors and movements based on machine learning methods. *Animals* 12:1060. <https://doi.org/10.3390/ani12091060>

- Mockus, J. 1975. On bayesian methods for seeking the extremum. p.400-404. In: Marchuk, G. I. (ed.). Optimization Techniques IFIP Technical Conference Novosibirsk. Optimization Techniques 1974. Lecture Notes in Computer Science, vol. 27. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-07165-2_55
- Mohammed, R.; Rawashdeh, J. and Abdullah, M. 2020. Machine learning with oversampling and undersampling techniques: overview study and experimental results. p.243-248. In: 2020 11th International Conference on Information and Communication Systems (ICICS). Irbid, Jordan. <https://doi.org/10.1109/ICICS49469.2020.239556>
- Nadimi, E. S.; Jørgensen, R. N.; Blanes-Vidal, V. and Christensen, S. 2012. Monitoring and classifying animal behavior using ZigBee-based mobile ad hoc wireless sensor networks and artificial neural networks. Computers and Electronics in Agriculture 82:44-54. <https://doi.org/10.1016/j.compag.2011.12.008>
- Park, J. K. and Park, E. Y. 2021. Monitoring method of movement of grazing cows using cloud-based system. ECTI Transactions on Computer and Information Technology 15:24-33. <https://doi.org/10.37936/ecti-cit.2021151.240087>
- Riaboff, L.; Aubin, S.; Bédère, N.; Couvreur, S.; Madouasse, A.; Goumand, E.; Chauvin, A. and Plantier, G. 2019. Evaluation of pre-processing methods for the prediction of cattle behaviour from accelerometer data. Computers and Electronics in Agriculture 165:104961. <https://doi.org/10.1016/j.compag.2019.104961>
- Riaboff, L.; Shalloo, L.; Smeaton, A.F.; Couvreur, S.; Madouasse, A. and Keane, M. T. 2022. Predicting livestock behaviour using accelerometers: A systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data. Computers and Electronics in Agriculture 192:106610. <https://doi.org/10.1016/j.compag.2021.106610>
- Ribeiro, L. A. C.; Bresolin, T.; Rosa, G. J. M.; Casagrande, D. R.; Danes, M. A. C. and Dórea, J. R. R. 2021. Disentangling data dependency using cross-validation strategies to evaluate prediction quality of cattle grazing activities using machine learning algorithms and wearable sensor data. Journal of Animal Science 99:skab206. <https://doi.org/10.1093/jas/skab206>
- Rodrigues da Cruz, P. J.; Silva, D.; Lima, I. B. G.; Alves, G. C.; Homem, B. G. C.; Alves, B. J. R.; Boddey, R. M.; Sbrissia, A. F. and Casagrande, D. R. 2024. Marandu palisade grass-forage peanut mixed pastures: Forage intake, animal behaviour, and canopy structure as affected by grazing intensities. Grass and Forage Science 79:666-677. <https://doi.org/10.1111/gfs.12688>
- Romanzini, E. P.; Watanabe, R. N.; Fonseca, N. V. B.; Berça, A. S.; Brito, T. R.; Bernardes, P. A.; Munari, D. P. and Reis, R. A. 2022. Modern livestock farming under tropical conditions using sensors in grazing systems. Scientific Reports 12:2654. <https://doi.org/10.1038/s41598-022-06650-5>
- Sagi, O. and Rokach, L. 2018. Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery 8:e1249. <https://doi.org/10.1002/widm.1249>
- Shafiullah, A. Z.; Werner, J.; Kennedy, E.; Leso, L.; O'Brien, B. and Umstätter, C. 2019. Machine learning based prediction of insufficient herbage allowance with automated feeding behaviour and activity data. Sensors 19:4479. <https://doi.org/10.3390/s19204479>
- Simanungkalit, G.; Barwick, J.; Cowley, F.; Dobos, R. and Hegarty, R. 2021. A pilot study using accelerometers to characterise the licking behaviour of penned cattle at a mineral block supplement. Animals 11:1153. <https://doi.org/10.3390/ani11041153>
- Teixeira, V. A.; Lana, A. M. Q.; Bresolin, T.; Tomich, T. R.; Souza, G. M.; Furlong, J.; Rodrigues, J. P. P.; Coelho, S. G.; Gonçalves, L. C.; Silveira, J. A. G.; Ferreira, L. D.; Facury Filho, E. J.; Campos, M. M.; Dorea, J. R. R. and Pereira, L. G. R. 2022. Using rumination and activity data for early detection of anaplasmosis disease in dairy heifer calves. Journal of Dairy Science 105:4421-4433. <https://doi.org/10.3168/jds.2021-20952>
- Valletta, J. J.; Torney, C.; Kings, M.; Thornton, A. and Madden, J. 2017. Applications of machine learning in animal behaviour studies. Animal Behaviour 124:203-220. <https://doi.org/10.1016/j.anbehav.2016.12.005>
- Wang, L.; Arablouei, R.; Alvarenga, F. A. P. and Bishop-Hurley, G. J. 2021. Animal behavior classification via accelerometry data and recurrent neural networks. arXiv:2111.12843. <https://doi.org/10.48550/arXiv.2111.12843>
- Wang, J.; Yu, Z.; Chebel, R. C. and Yu, H. 2025. Impact of cross-validation designs on cattle behavior prediction using machine learning and deep learning models with tri-axial accelerometer data. bioRxiv 2025.01.22.634181. <https://doi.org/10.1101/2025.01.22.634181>
- Warner, D.; Vasseur, E.; Lefebvre, D. M. and Lacroix, R. 2020. A machine learning based decision aid for lameness in dairy herds using farm-based records. Computers and Electronics in Agriculture 169:105193. <https://doi.org/10.1016/j.compag.2019.105193>
- Watanabe, R. N.; Bernardes, P. A.; Romanzini, E. P.; Braga, L. G.; Brito, T. R.; Teobaldo, R. W.; Reis, R. A. and Munari, D. P. 2021. Strategy to predict high and low frequency behaviors using triaxial accelerometers in grazing of beef cattle. Animals 11:3438. <https://doi.org/10.3390/ani11123438>
- Yigit, T.; Han, F.; Rankins, E.; Yi, J.; McKeever, K. H. and Malinowski, K. 2022. Wearable inertial sensor-based limb lameness detection and pose estimation for horses. IEEE Transactions on Automation Science and Engineering 19:1365-1379. <https://doi.org/10.1109/TASE.2022.3157793>