

METHODOLOGY

Open Access



# Comparing the accuracy of machine learning methods for classifying wild red deer behavior based on accelerometer data

Benjamin Bar-Gera<sup>1\*</sup>, Pia Anderwald<sup>2</sup>, Alina L. Evans<sup>3</sup>, Thomas Rempfler<sup>2</sup> and Claudio Signer<sup>1</sup>

## Abstract

**Background** Effective conservation requires understanding the behavior of the targeted species. However, some species can be difficult to observe in the wild, which is why GPS collars and other telemetry devices can be used to “observe” these animals remotely. Combined with classification models, data collected by accelerometers on a collar can be used to determine an animal’s behaviors. Previous ungulate behavioral classification studies have mostly trained their models using data from captive animals, which may not be representative of the behaviors displayed by wild individuals. To fill this gap, we trained classification models, using a supervised learning approach with data collected from wild red deer (*Cervus elaphus*) in the Swiss National Park. While the accelerometer data collected on multiple axes served as input variables, the simultaneously observed behavior was used as the output variable. Further, we used a variety of machine learning algorithms, as well as combinations and transformations of the accelerometer data to identify those that generated the most accurate classification models. To determine which models performed most accurately, we derived a new metric which considered the imbalance between different behaviors.

**Results** We found significant differences in the models’ performances depending on which algorithm, transformation method and combination of input variables was used. Discriminant analysis generated the most accurate classification models when trained with minmax-normalized acceleration data collected on multiple axes, as well as their ratio. This model was able to accurately differentiate between the behaviors lying, feeding, standing, walking, and running and can be used in future studies analyzing the behavior of wild red deer living in Alpine environments.

**Conclusion** We demonstrate the possibility of using acceleration data collected from wild red deer to train behavioral classification models. At the same time, we propose a new type of metric to compare the accuracy of classification models trained with imbalanced datasets. We share our most accurate model in the hope that managers and researchers can use it to classify the behavior of wild red deer in Alpine environments.

**Keywords** *Cervus elaphus*, Swiss National Park, Behavioral classification, Acceleration data, Data imbalance, Overall accuracy, Balanced accuracy, Specificity, Sensitivity, Neural networks, Discriminant analysis, Random forest, Support vector machines, Decision trees

## Background

In order to effectively protect and manage a species, it is important to understand its behavior [1]. Although visual observation is the most effective method to learn about an animal’s behavior, observing wild animals such as red deer (*Cervus elaphus*) can be difficult as they are often elusive, may live in habitats with tree cover, are

\*Correspondence:  
Benjamin Bar-Gera  
benibargera@gmail.com  
Full list of author information is available at the end of the article



nocturnal, move over large distances, and can easily be disturbed by the observer [2]. To overcome these challenges, GPS collars and other telemetry methods have been used to study the spatial movements of red deer, resource selection and seasonal migrations, as well as other factors affecting their movements, such as human activity [3–5]. While these methods can provide valuable knowledge about the spatiotemporal behavior of animals, they suffer from various limitations, most notably that it can be difficult to infer which behavior the animals are engaging in [6].

Accelerometers have become a frequent component of GPS collars [6]. They measure the collar's and thus the animal's intensity of movement as the difference in velocity between two consecutive measurements. Accelerometers usually record multiple measurements per second. The data is then either saved raw (i.e., high-resolution [7, 8]) or averaged over predefined time intervals, such as 1, 5 or 10 min (i.e., low-resolution [2, 9–11]). While resolution can also pertain to the device's sampling frequency or bit-resolution, we solely use this term regarding whether the acceleration data has been averaged over predefined time intervals or left in its raw state. Averaging and reducing the amount of data can be useful when animals wear the collars over long periods and the amount of data storage is limited [12]. Additionally, working with low-resolution data requires less computing power and tends to be more accessible from a technical point of view than working with high-resolution data.

Acceleration data can be used to infer an animal's relative level of activity as a result of time, seasonality, weather, sex or age [13, 14]. Combined with a classification model the acceleration data can provide knowledge about the animal's behavior and has been used for a variety of species in the wild, including pumas (*Puma concolor*) [15], Alpine ibex (*Capra ibex*) [16], polar bears (*Ursus maritimus*) [17], or various cervids [2, 7–9, 11, 18]. Previous behavioral classification models for cervids can be categorized by whether they were trained with captive [11] or wild animals [7], whether they use low- [9] or high-resolution [8] acceleration data and whether they are binary [18] or multiclass [2] models. A binary model classifies only two different modes, such as two behaviors (e.g., feeding vs. walking) or whether the animal is active or inactive, whereas a multiclass model has the potential to classify more than two behaviors (e.g., running, feeding, or standing).

To the best of our knowledge, no multiclass models have been trained using wild cervids and low-resolution acceleration data. While models trained on captive animals can be very useful in certain circumstances, previous studies have illustrated that such models may perform worse than models trained with wild animals

when classifying the behavior of wild animals, due to differences in behavior and/or habitat [17]. As there is always a tradeoff between the resolution of accelerometer data and memory capacity, long-term studies on wild animals frequently use collars that only save low-resolution acceleration data. For these two reasons, our first goal was to generate a multiclass behavioral model that is based both on low-resolution acceleration data and behavior of wild cervids.

With the adoption of sensors such as accelerometers, researchers are increasingly confronted with large datasets [19]. Machine learning (ML) algorithms can help find patterns in these datasets to generate new ecological insight (e.g., estimating animal populations with unmanned aerial vehicle footage [20]) or automate previously manual tasks (e.g., classifying trail camera images [21, 22]). As ML algorithms have become an increasingly popular tool in the field of ecology, they have also become easily accessible through R packages [23–31].

However, with so many different algorithms, it is difficult to know which ones to use. Previous studies have used discriminant function analysis [9, 11], recursive partitioning (i.e., classification and regression tree) [2, 10], or random forest [7, 8]. Nathan et al. [32] compared the efficacy of various ML algorithms for classifying the behavior of griffon vultures (*Gyps fulvus*) and Ladds et al. [33] did the same with fur seals (*Arctocephalus spp.*) and sea lions (*Neophoca cinerea*). However, to date no such comparison has been performed for cervids. Our second goal was therefore to fill this gap by using a variety of ML algorithms and analyzing which ones generate the most accurate classification models.

GPS collars usually include multiple accelerometers that can measure the movement on different axes (e.g., left–right, up–down, forward–backward). For our third goal, we generated models using different combinations of the axial acceleration values and their derived counterparts (sum, difference, and ratio). This allowed us to analyze not only which algorithms, but also which combination of input variables generate the most accurate models. We also applied various normalization methods to the acceleration data to identify which ones generated the most accurate models.

## Methods

### Study area

All field observations were conducted in and around the Swiss National Park (SNP), which covers an area of approximately 170 km<sup>2</sup> in eastern Switzerland [34]. The SNP is classified as an IUCN 1a conservation area (highest class of protection, wilderness area). Visitors must remain on the provided paths, plants may not be removed except for scientific reasons, meadows cannot

be mowed, and all hunting is forbidden. The SNP has a diverse topography typical for the Central Alps, with elevations ranging from 1380–3173 m a.s.l. The tree line varies between 2200–2300 m a.s.l. Most of the SNP's area is subalpine and composed of around 31% forest and 17% meadow, while the rest is free of vegetation. The mean annual temperature (1991–2020) at the weather station Buffalora was 1.1 °C, the coldest month being January (− 9.1 °C) and the warmest July (11 °C) [35]. Mean annual precipitation is 936 mm, with July and August being the wettest months with 118 and 130 mm of rain, respectively. Buffalora is located just outside the SNP and at a similar altitude (1971 m a.s.l.) as the observational locations.

### Study animals and telemetry collars

In the SNP and surrounding areas, wild red deer have been equipped with GPS collars since 1998 as part of various research and management efforts. Wildlife officials of the canton of Grisons and the SNP immobilized and anesthetized the animals with dart guns and 3 ml Hella-brunn mixture containing 125 mg xylazine and 100 mg ketamine per ml. Capturing and collaring was conducted according to Swiss animal welfare law (permit GR2015-09). The animals wear the collars for a maximum of 2 years, after which they are released via a remote drop-off mechanism. Every individual receives a unique combination of colored ear tags, allowing visual identification. During the fieldwork period, we observed four individuals in the wild. These included two stags (No. 779 and No. 783) and two hinds (No. 761 and No. 762). At the time, they were estimated to be 7, 9, 13 and 9 years old, respectively. Both hinds were rearing a calf and were additionally accompanied by a yearling.

The observed red deer were equipped with two different types of GPS collars from VECTRONIC Aerospace GmbH, Berlin, Germany: PRO LIGHT and VERTEX PLUS [36, 37]. Besides the location, these collars measure intensity of movement using multiple accelerometers. The antenna and electronic housing, including the accelerometer, are located on top of the collar and

thus on the back of the animal's neck. The accelerometers measure acceleration continuously at 4 Hz on each axis as the difference in velocity between two consecutive measurements. Acceleration is averaged over 5-min intervals per axis and provided as a unit-free number ranging from 0–255, with 0 representing no movement and 255 maximum movement. Henceforth, these values will be referred to as acceleration values or acceleration data. The different types of collars are equipped with accelerometers on either two (x, y) or three axes (x, y, z), where the x-axis measures forward–backward motion, the y-axis sideways (i.e., left–right) motion, and the z-axis up–down movements. However, as two (762 and 779) of the four observed individuals wore collars which only measure x- and y-acceleration, we only used these two axes to generate the models. Acceleration data can be downloaded from the collars via UHF and VHF in the field or directly from the device after drop-off.

### Behavioral observations

Animal observations took place during July and August 2022. As hunting is prohibited and human activity inside the SNP is restricted to hiking trails, red deer are often active and visible in open habitats during the day. We observed the animals from a distance between 250–1200 m using a spotting scope, focusing on one collared individual at a time, as long as it was visible [38]. The behavior was logged simultaneously in the ethological app “Behayve”, which generated time-stamped behavioral logs for every observational session and individual [39]. While using Behayve proved to be very effective in collecting observational data, we additionally filmed most behavior through the spotting scope using a digiscoping adapter and a smartphone. The Android app “Timestamp Camera Pro” [40] was used for filming, as it displays the current time as a watermark. The filmed behavior served as a point of reference in case of any logging errors and was also used to distinguish specific behaviors more clearly.

Similar to previous studies [2, 7, 9–11], we distinguished between the behaviors lying, feeding, walking, running, standing, and fighting (Table 1). We also

**Table 1** Observed behavior of wild red deer in the Swiss National Park

Behavior	Description
Lying	Lying on the ground either resting or ruminating
Feeding	Grazing with or without moving from one spot to another; we observed no instances of browsing
Standing	Standing in one spot, either vigilant and/or ruminating
Walking	Moving slowly from one spot to another without grazing at the same time
Running	Moving quickly from one place to another, either trotting or galloping
Fighting	Stags only: clashing antlers with another stag

**Table 2** Formulae used to train the behavioral classification models

---


$$\text{behavior} \sim x_{\text{minmax}} + y_{\text{minmax}}$$

$$\text{behavior} \sim x_{\text{scale}} + y_{\text{scale}}$$

$$\text{behavior} \sim x_{\text{log}} + y_{\text{log}}$$

$$\text{behavior} \sim x_{\text{minmax}}$$

$$\text{behavior} \sim y_{\text{minmax}}$$

$$\text{behavior} \sim \text{sum}(x, y)_{\text{minmax}}$$

$$\text{behavior} \sim \text{diff}(x, y)_{\text{minmax}}$$

$$\text{behavior} \sim \text{ratio}(x, y)_{\text{minmax}}$$

$$\text{behavior} \sim x_{\text{minmax}} + y_{\text{minmax}} + \text{ratio}(x, y)_{\text{minmax}}$$


---

recorded whether the animals were ruminating or not while lying, as well as their vertical head position while walking and feeding. However, our preliminary analyses suggested that our models would not be able to distinguish between these modes, which is why we did not include them further in our study.

**Data analysis**

After completing the behavioral observations, the acceleration data was downloaded remotely from the GPS collars via VHF/UHF and the behavioral data was exported from the Behave app on the smartphone. The workflow of generating the classification models consisted of pre-processing the acceleration and behavioral data, model-training, and model-testing (see Additional file 1: Fig. S1 for a schematic overview of the data analysis process).

Preprocessing involved checking the behavioral and acceleration data visually for errors, normalizing, transforming, and labeling the acceleration data with the simultaneous behaviors, and splitting the data into a training (75%) and testing (25%) subset. To train the classification models, various algorithms were employed, with the acceleration data serving as input variables and the behavior as the output variable (Table 2). Model-testing was conducted with the testing subset to assess and compare the different models and their accuracy [8,

9, 41]. Data analysis was conducted with the R programming language [42] and using RStudio [43].

**Normalizing and transforming the acceleration data**

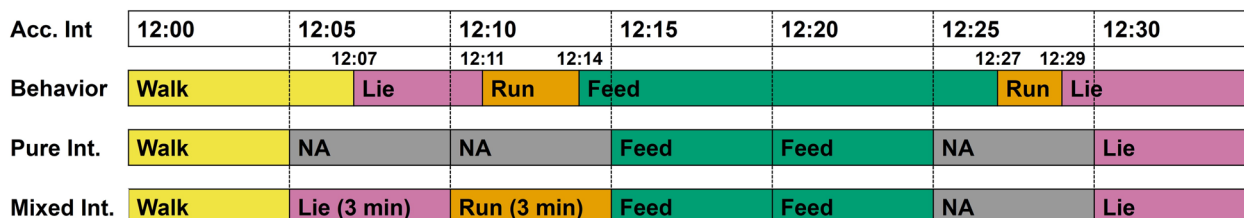
As the tightness of a collar can significantly affect the acceleration data, and the individuals wore different collar types, we tested whether inter-individual differences in the acceleration data existed [44]. We hypothesized that such differences might negatively affect the models’ ability to classify behaviors across all individuals [45]. Because we detected significant inter-individual differences (Kruskal–Wallis  $\chi^2=7745.7$ ,  $df=3$ ,  $p<0.001$  for x-acceleration and  $\chi^2=4979.8$ ,  $df=3$ ,  $p<0.001$  for y-acceleration), we separately applied scale-transformation to each individual and axis, thereby reducing these inter-individual differences [45].

Additionally, we applied minmax-normalization, which retains the original distribution of the values but projects them onto a 0–1 scale, improving the speed at which models can be trained. We also applied log-transformation to test whether this might have a positive effect on the models’ accuracy (see Additional file 2: Table S1 for a detailed description of the normalization methods and Additional file 3: Fig. S2 for a visualization of their effects).

Similarly, we generated derived acceleration values, including the sum ( $acc_x + acc_y$ ), difference ( $acc_x - acc_y$ ) and ratio ( $\frac{acc_x}{acc_y}$ ) of both axes. Having access to the variously transformed and derived acceleration values allowed us to compare their efficacy in generating accurate classification models.

**Combining acceleration and behavioral data**

The challenge in linking behavioral and acceleration data is that the acceleration intervals always last 5 min (12:00–12:05, 12:05–12:10, ...), but behaviors of red deer are not consistent with these intervals. As a solution, two types of labeled acceleration intervals were generated: pure and mixed intervals (Fig. 1). During a pure interval, the animal engages continuously in a single behavior. During a mixed interval, the animal may engage in multiple behaviors, but, more



**Fig. 1** Time-dependent linkage of acceleration intervals (constant time interval of 5 min) and behavioral data (variable duration). Every pure acceleration interval starts and ends within the same continuous behavior. Mixed acceleration intervals include all intervals during which a single behavior was engaged in for at least half the duration (> 2.5 min) and hence also include all pure intervals

importantly, engages in one behavior for at least have the acceleration interval (>2.5 min). As such, by this definition, mixed intervals also include pure intervals.

Acceleration data from wild animals will always include mixed intervals, especially when intervals are as long as 5 min. For the rare and short behaviors, standing and walking, the number of mixed intervals was significantly higher than the number of pure intervals (Table 4). Using mixed intervals allowed us to include these behaviors and generate multiclass models. Additionally, by training and testing models with pure intervals, for behaviors that are likely to include a high proportion of mixed intervals, we risk generating artificial and inflated estimates of the models' accuracies [9, 46]. For these reasons, we decided to use mixed intervals to train and test all our models. Due to the rarity of fighting, we were able to generate only two mixed intervals for this behavior and did not include this behavior in our subsequent analysis.

#### Train–test split

To train and test the classification models, we split the data into mutually exclusive datasets: 75% of the labeled intervals were used for training, and 25% for testing the models [41]. This was done separately for each behavior, ensuring that they roughly reflected the proportion of behaviors in the overall dataset.

#### Model-training

The models were trained using the labeled training intervals and a supervised learning approach [41]. For every interval, the model was provided with the acceleration values and the corresponding behavior. By providing the model with a large dataset of labeled intervals, it “learns” to predict which behavior an animal engaged in based purely on acceleration values. Behavior always served as the output variable, whereas the input variables consisted of different combinations

of the acceleration values and their derivatives (Table 2).

In our initial analysis, we found that models trained with minmax-normalized acceleration values classified behaviors more accurately than the models trained with log- or scale-transformed data. Therefore, all subsequent models were based on minmax-normalized data (Table 2).

For each of the mentioned formulae (Table 2), we trained models with various ML algorithms. Similarly to using different combinations of input variables, the purpose of using different algorithms was to find out which ones generated the most accurate classification models. The used algorithms, relevant literature and employed R packages are described in Table 3. Some of the algorithms allow the use of class weights to mitigate the sample size imbalance between the different classes (i.e., behaviors). In our preliminary analyses, we found that using custom weights did not improve the models' accuracy, which is why we decided against employing them.

#### Model-testing

After training the models, their accuracy was evaluated using the testing subset (25%) [9, 41]. Each model was used to predict the behavior of the testing intervals based on the acceleration values. The predicted behavior of each interval was then compared to the actual observed behavior of that interval (Additional file 1: Fig. S1) and a confusion matrix was generated.

To efficiently compare the different models, however, it is useful to have a single descriptive value. Previous studies have frequently used the correct classification rate (CCR), also known as overall accuracy [2, 9–11, 32]:

$$\text{Correct classification rate} = \frac{\text{Number of correctly classified intervals}}{\text{Total number of intervals}}$$

**Table 3** Algorithms used to train the classification models, respective literature and employed R packages

Algorithm and useful literature	Employed R packages
K-nearest neighbor (KNN) [47]	“Caret” [27]
Multinomial logistic regression [41, 48]	“nnet” [31]
Support vector machines (SVM): linear, polynomial, radial or sigmoidal [32, 41, 49]	“e1071” [29]
Discriminant analysis: linear (LDA) and flexible (FDA) [32, 41, 50]	LDA: “MASS” [31]; FDA: “mda” [25]
Artificial neural network (ANN) [32, 51]	“neuralnet” [24]
Naïve Bayes [52, 53]	“e1071” [29]
Gaussian process [54]	“kernlab” [26, 55]
Classification and regression tree (CART; simple and pruned) [41]	“tree” [30]
Ensemble decision tree models: Boosted regression trees (BRT; simple and tuned) and random forest (RF) [56]	BRT: “xgboost” [23]; RF: “randomForest” [28]

Unfortunately, CCR does not consider a dataset's imbalance. In our case, the most frequently labeled intervals were either of type lying or feeding (Table 4). A model that is capable of accurately classifying these two behaviors might therefore receive a high CCR, even though it poorly classifies rare behaviors (Fig. 5).

Recent studies [7, 8] have included precision (i.e., positive predictive value) and sensitivity (i.e., recall or true positive rate), which are calculated for each behavior separately and provide a fuller picture for imbalanced datasets [57]:

$$Precision = \frac{True\ positives}{False\ positives + True\ positives}$$

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives}$$

As some of our models predicted 0 intervals for some behaviors (0 false positives and 0 true positives), precision would divide by 0 and could therefore not be used. Specificity proved to be a useful alternative to precision [57]:

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives}$$

Whereas sensitivity measures a model's ability to detect a positive case, specificity measures a model's ability to detect a negative case. For the behavior of running, sensitivity evaluates a model's ability to detect running intervals, whereas specificity would evaluate the model's ability to classify that interval as not running (i.e., any other behavior such as feeding). A model can receive a high sensitivity or a high specificity for running by either classifying all intervals as running or 0 intervals as running, respectively. Balanced accuracy counteracts this possibility by calculating the mean of sensitivity and specificity for each behavior [57, 58]:

$$Balanced\ accuracy = \frac{sensitivity + specificity}{2}$$

After evaluating balanced accuracy for each model and behavior, the unweighted mean of all these balanced accuracies per model was calculated. We termed this average the macro-balanced accuracy (MBA). Because the balanced accuracy for each behavior is weighted equally in this metric, a model can only receive a high MBA if it can predict each behavior sufficiently well, regardless of how rare or frequent it is (Fig. 5). The MBA allowed us to compare the different models with each other and draw conclusions about which combinations

**Table 4** Number of observations, durations, and number of mixed/pure intervals per behaviors

Behavior	Lying	Standing	Feeding	Walking	Running
Observations (N)	142	148	312	156	18
Mean (min)	34.77	1.49	14.11	1.16	2.39
Median (min)	23.51	0.66	7.19	0.54	1.42
Minimum (min)	0.03	0.03	0.03	0.03	0.18
Maximum (min)	211.18	20.12	119.70	11.67	10.43
Pure intervals (N)	752	5	601	2	4
Mixed intervals (N)	884	25	771	26	5

The number (N) of observational instances per behavior with their mean, median, minimum, and maximum duration in minutes (min), and the respective number (N) of pure and mixed intervals. Pure intervals are defined as intervals during which the animal engaged in only one behavior. Mixed intervals are defined as intervals during which a single behavior was engaged in for at least half the duration (> 2.5 min), and hence also include pure intervals.

of input variables and ML algorithms generate the most accurate models.

## Results

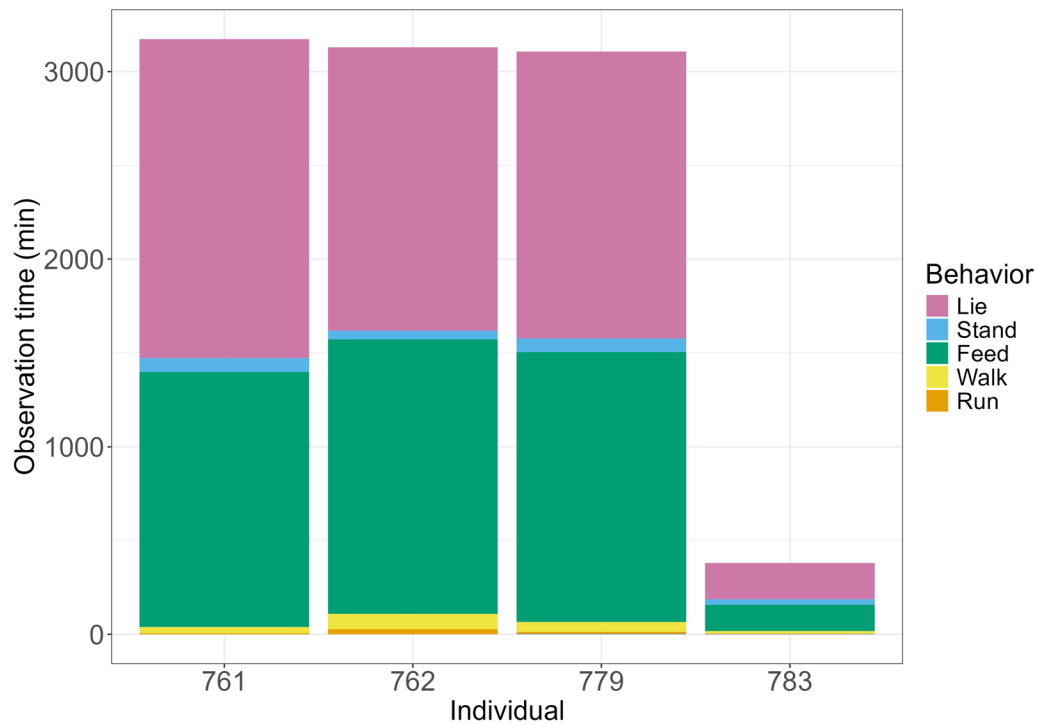
### Behavioral observations

We were able to observe the four collared red deer on 35 out of the 57 field days, resulting in a behavioral data set of 160 h. However, the frequency at which we observed individuals and behaviors was strongly imbalanced (Fig. 2). While we frequently observed the animals lying or feeding, we rarely observed them running, walking, or standing. Further, we were rarely able to observe individual 783.

Additionally, some behaviors occurred for a much shorter duration than others. While the animals lay down, on average, for 34.77 min at a time, they walked, on average, for only 1.16 min at a time (Table 4). As a result, there were little to no pure intervals for the short duration behaviors walk, stand, and run (Table 4). Including mixed intervals provided a significant increase in the number of intervals for these behaviors and allowed us to generate a multiclass model, which would have been impossible with pure intervals only.

### Model performance

In total, we generated 144 classification models (16 algorithms \* 9 formulae). The performance of each model is listed in Additional file 4: Table S2. The most accurate model had an MBA of 81%, and balanced accuracies of 90% (lie), 57% (stand), 88% (feed), 71% (walk), and 100% (run). The model was trained using linear discriminant analysis. The models trained with flexible discriminant analysis performed almost always equally well as the



**Fig. 2** Duration of observed behaviors per individual (females 761 and 762, males 779 and 783)

linear model. The model was trained using the following formula:

$$behavior \sim x_{minmax} + y_{minmax} + ratio(x, y)_{minmax}$$

The model generated the confusion matrix detailed in Table 5.

There were significant differences in the performance of the models according to the input variables used to train them (Kruskal–Wallis  $\chi^2=79.855$ ,  $df=8$ ,  $p<0.001$ ; Fig. 3). The six models with the highest MBA were all trained using either the input variables  $x_{minmax} + y_{minmax} + ratio(x, y)_{minmax}$  or  $x_{minmax} + y_{minmax}$ . These two groups of models also had a higher median MBA than all other formula

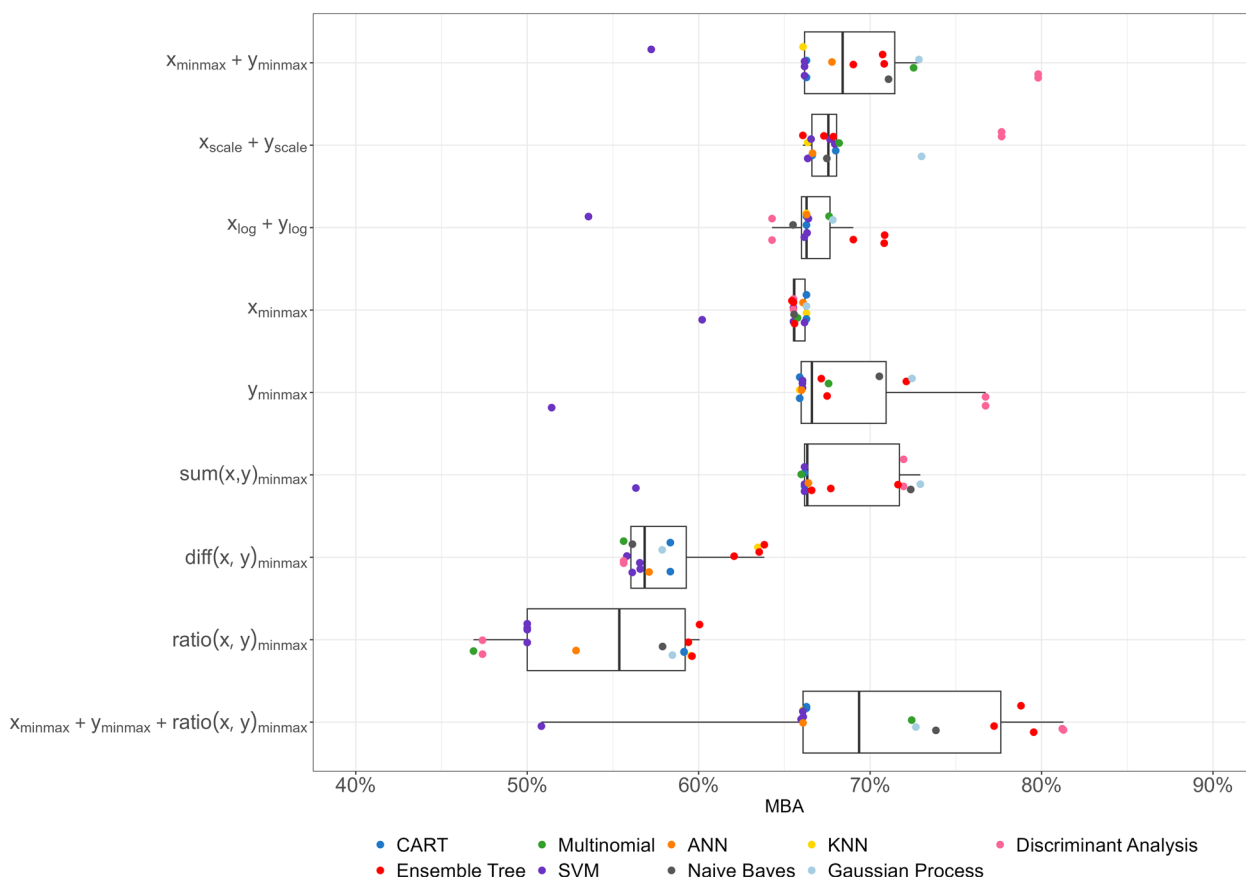
groups. The models trained with the input variables  $diff(x, y)_{minmax}$  and  $ratio(x, y)_{minmax}$  on their own, had the lowest median accuracies. However, when combined with  $x_{minmax} + y_{minmax}$ , using  $ratio(x, y)_{minmax}$  generally improved the MBA. The models trained only with  $x_{minmax}$ ,  $y_{minmax}$  or  $sum(x, y)_{minmax}$  were intermediate with regard to their median classification accuracy. In terms of normalization methods, the minmax-normalized models seem to outperform, on average, the scale-normalized and log-transformed models.

The type of algorithm also had a significant impact on the models’ MBA (Kruskal–Wallis  $\chi^2=21.043$ ,  $df=8$ ,  $p=0.007$ ; Fig. 4). The models with the highest median MBA were trained using the Gaussian process

**Table 5** Confusion matrix for the most accurate classification model

		Observations				
		Lying	Standing	Feeding	Walking	Running
Prediction	Lying	204	1	22	1	0
	Standing	0	1	0	0	0
	Feeding	16	4	166	2	0
	Walking	1	0	5	3	0
	Running	0	1	0	1	2

Confusion matrix of the model with the highest macro-balanced accuracy (MBA). The model was trained using discriminant analysis and the input variables  $x_{minmax} + y_{minmax} + ratio(x, y)_{minmax}$



**Fig. 3** Boxplot visualizing each model’s macro-balanced accuracy (MBA) (Additional file 4: Table S2). Each model (point) was trained with a different combination of input variables and algorithm. In this figure, the models are vertically grouped by the combination of input variables that was used to train them

algorithm. However, the 11 models with the highest MBA were all trained using discriminant analysis or ensemble decision tree algorithms (Additional file 4: Table S2). Interestingly, discriminant analysis generated the models with the highest and some of the lowest MBA. The models with the lowest median MBA were trained using KNN, ANN, SVM, and CART. In general, there was no algorithm that always outperformed all others. However, discriminant analysis and ensemble decision tree models generated the most accurate models when combined with the right set of input variables and performed relatively similar at their upper end.

Figure 5 visualizes the benefits of using the MBA as a metric. Each point represents a model. While the vertical axis denotes that model’s MBA and CCR, respectively, the horizontal axis denotes its balanced accuracy for the behavior running. Running serves as an example for any of the rare behaviors, including standing and walking. Plot a demonstrates that a model’s ability to accurately classify running has as significant effect on its MBA ( $R=0.7$ ). If the model is unable to accurately predict

running, it will not receive a high MBA. Plot b demonstrates that a model’s ability to accurately classify running has little effect on the CCR ( $R=0.24$ ). The model can still have a high CCR even though it predicts the rare behavior poorly. The MBA therefore provides a more balanced perspective on a model’s ability to classify all behaviors, regardless of how frequent or rare they are.

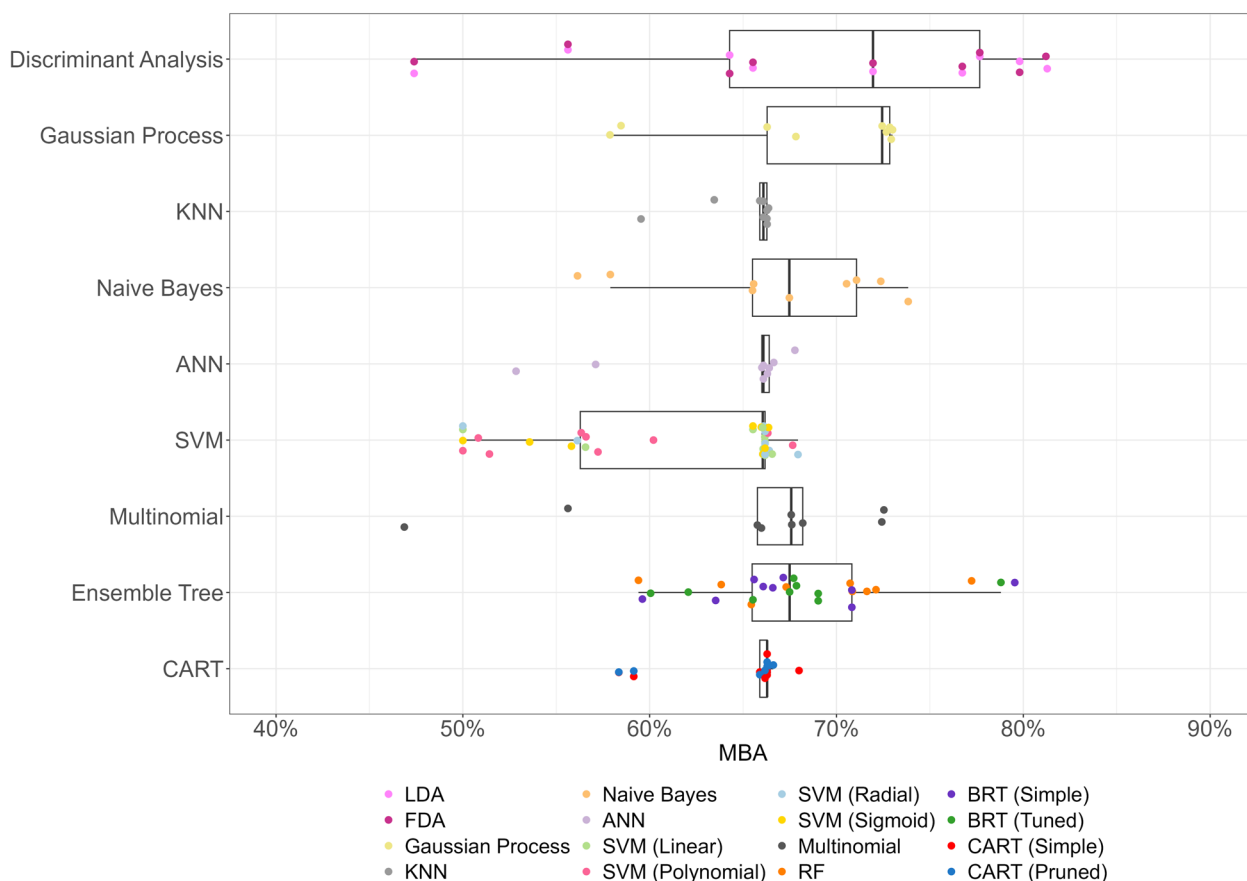
**Discussion**

**Generalizability of the models**

There have been a number of studies that have generated multiclass classification models for cervids in captivity [2, 8–11] or binary models for wild cervids [18, 59–63]. However, so far there have been fewer studies to have generated multiclass models for wild cervids [7] and, to our knowledge, no such studies for wild cervids living in an alpine environment or using low-resolution (5-min) acceleration values.

The most obvious reason for a lack of models trained with wild cervids is the significant increase in the effort





**Fig. 4** Boxplot visualizing each model’s macro-balanced accuracy (MBA) (Additional file 4: Table S2). Each model (point) was trained with a different combination of input variables and algorithm. In this figure, the models are vertically grouped by the category of algorithm that was used to train them

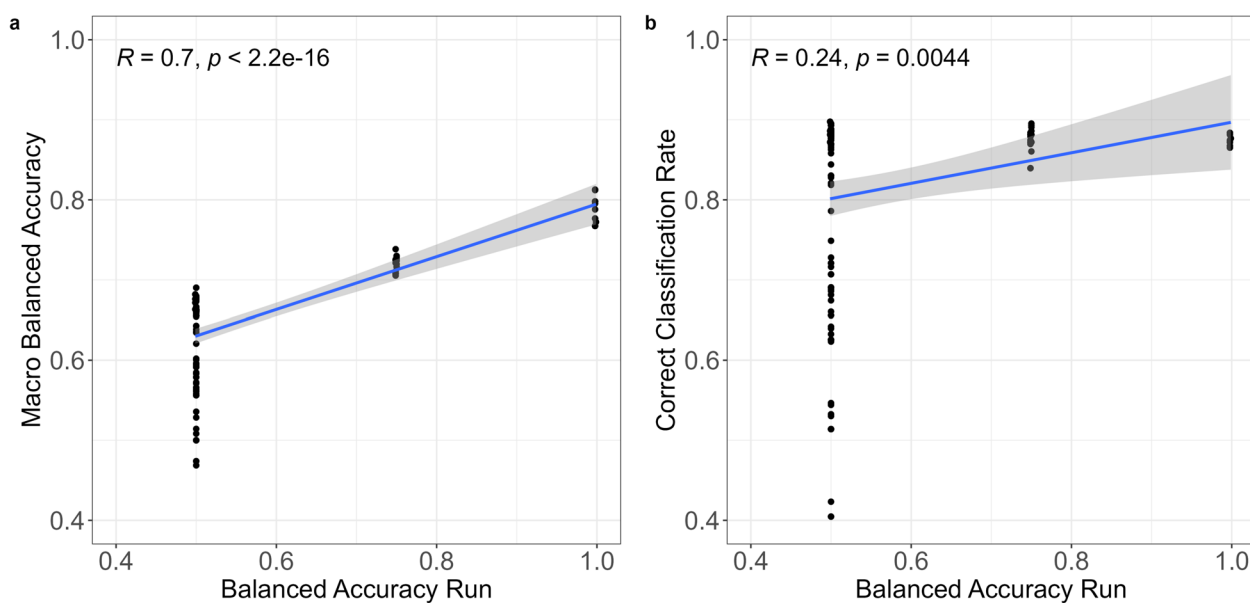
it takes to collect sufficient behavioral data. For the 160 h of behaviors we observed, we spent roughly 570 h in the field. Collecting this amount of behavioral data would, most likely, be much more time-efficient with captive animals. However, this raises the question whether a model trained on captive red deer would be generalizable to wild animals [8, 17, 33, 64]. Red deer in captivity might show different behaviors than wild red deer or move differently, depending on the landscape they live in. Although observing captive animals might be much more time-efficient than observing wild animals, we argue that it is worthwhile collecting observational data from wild animals, even if only used for testing the models.

Spreading the fieldwork phase over a longer period might allow for models that are generalizable to different seasons and might also provide observational data of additional individuals and behaviors. Most of the red deer in the SNP move from their alpine summer habitats to their winter habitats at lower elevations around October and November [4, 65]. Due to differences in their habitat and possibly in their behavior, the animals might

move differently in the winter than in the summer. Additionally, as red deer undergo seasonal changes in their body weight, it would be valuable to test whether the “summer models” are still generalizable to red deer in the winter [44, 66].

**Data imbalance**

Similarly to previous studies, we were able to observe the behaviors lying and feeding much more frequently and over longer durations than other behaviors such as running, fighting, standing, and walking (Fig. 2; [2, 7–9, 11]). Additionally, some behaviors tend to occur for less than the duration of the 5-min acceleration intervals (Table 4). For these behaviors, we only had access to a very small number of pure intervals and would have been unable to generate a multiclass model (Table 4). While having access to shorter acceleration intervals (e.g., 1 min duration) or even the acceleration data in its raw state (i.e., high resolution) would be ideal, this is not always possible. This might be due to working with older data, or, as in our case, due to memory storage constraints and the



**Fig. 5** Scatterplots visualizing the correlation between each models' balanced accuracy for running and its **a** macro-balanced accuracy (MBA), as well as its **b** correct classification rate (CCR). Each point denotes one model. Running correlates much stronger with MBA than it does with CCR, illustrated by the regression line (blue; 95% confidence interval), Pearson correlation coefficient ( $R$ ) and respective  $p$  value

long employment period of the GPS collars (> 2 years). We were able to mitigate this issue by using mixed intervals, which allowed us to label significantly more intervals for the rare and short behaviors than if we had relied only on pure intervals. Nevertheless, we suggest that authors of future behavioral classification studies carefully consider which type of acceleration data fits their research goals, yet still complies with the technical constraints imposed by the storage capacity and employment duration of the GPS collars.

Of the studies which generated multiclass classification models for cervids with low-resolution acceleration data, only Gaylord & Sanchez [9] used mixed intervals. Loettker et al. [2], Heurich et al. [10] and Naylor & Kie [11] worked exclusively with pure intervals. The viability of using mixed intervals can be evaluated from different angles. On the one hand, every classified mixed interval is, per definition, also misclassified. To alleviate this issue, we only labeled mixed intervals where the animal engaged in one behavior for at least half of that interval's duration. On the other hand, Gaylord & Sanchez [9] argue that "datasets from free-ranging animals inherently include mixed intervals... reliance on pure-interval models to classify behaviors of free-ranging animals should be avoided" (p. 64). They argue that relying on pure intervals models can lead to an inflated sense of behavior classification accuracy.

Another issue stemming from the behavioral imbalance pertains to the evaluation of the models' accuracy. Previous studies have primarily used the CCR

which evaluates a model's overall accuracy at classifying behaviors [2, 9–11]. However, when faced with a strong data imbalance, as in our case, the CCR can provide an inflated sense of a model's accuracy because it is strongly biased towards the behaviors that are most commonly represented in the dataset (Fig. 5). More recent studies have used alternatives to the CCR when evaluating the performance of behavioral classification models. Kröschel et al. [7] used CCR, sensitivity and the positive predictive value for their roe deer models. Kirchner et al. [8] used recall and precision for their moose models.

In our case, we decided to use the MBA. The MBA is the mean of each behavior's balanced accuracy and thereby weighs a model's ability to classify each behavior equally, regardless of how rarely or frequently it has been observed. As such, the MBA avoids the CCR's bias towards the more commonly represented behaviors. However, the MBA is not without its own limitations. When working with very small classes, such as running or walking, a small number of misclassifications in these behaviors can have an inflated effect on the final MBA. Whichever metric one might use, we suggest that authors of future behavioral classification models try out various metrics to test which ones fit their dataset situation and research questions [57].

### Model performance

By using 16 different ML algorithms, as well as 9 different combinations of input variables, we were able to train and test 144 different models. This allowed us to draw conclusions about what generates accurate models with respect to the used input variables, normalization methods and algorithms.

While there was no combination of input variables that always outperformed all other input variable combinations, using  $x_{\min\max} + y_{\min\max}$  or  $x_{\min\max} + y_{\min\max} + \text{ratio}(x, y)_{\min\max}$  generally generated the most accurate models (Fig. 3). Interestingly, models trained with only  $\text{ratio}(x, y)_{\min\max}$  as an input variable had the lowest median MBA. However, when combined with  $x_{\min\max} + y_{\min\max}$ , using  $\text{ratio}(x, y)_{\min\max}$  generally improved the MBA. Visually, the behaviors running and walking appear to have a higher x-to-y acceleration ratio, which might explain why using ratio as an additional input variable improves these models (Additional file 3: Fig. S2).

As there were significant differences in the acceleration values from the four individuals, we expected that decreasing these inter-individual differences might improve the models' ability to classify the behavior of all individuals. We were surprised that the models using minmax-normalized acceleration data (thereby retaining inter-individual differences) had, on average, a higher MBA than models trained with scale-normalized acceleration data (Fig. 3). Retaining inter-individual differences and the original distribution of acceleration values seems to be vital to generating accurate classification models.

Similar to the combination of input variables, there was no type of algorithm that outperformed all other algorithms for every model (Fig. 4). However, the best-performing models were all trained with discriminant analysis or ensemble decision tree algorithms. In their classification models for griffon vultures, Nathan et al. [32] found that RF outperformed SVM which outperformed discriminant analysis. While we were also able to generate accurate models with RF, we found that discriminant analysis had the highest and SVM the lowest median MBA. Similar to Nathan et al. [32], Ladds et al. [33] also found that SVM generally performed well for classifying the behaviors of fur seals and sea lions. Similar to our study, Ladds et al. [33] were also able to generate accurate models with RF and BRT. The differences in these studies' findings should underline the importance of trying out and comparing various algorithms for each new dataset and classification process.

Regarding the differences within algorithm groups, we did not find that pruning CART or tuning BRT resulted in a significant improvement in their accuracy. In fact, tuning and pruning seemed to have had a slightly negative

effect on the models' MBA (Fig. 4). Similarly, whether we used flexible or linear discriminant analysis seemed to have had little to no effect on the models' MBA. However, it is possible that for other datasets, these variations improve the models' accuracy and should therefore not be dismissed [41].

The variation of the MBA was much greater within the algorithm groups (Fig. 4) than within the input variable groups (Fig. 3). This wider variation might be caused by the specific algorithms within each algorithm group performing very differently, or by the strong effect of the used input variables. Whichever the case, the different combinations of input variables should be tested just as rigorously as the algorithms, when deciding on which ones to use.

While we purposefully did not use the CCR to determine the best model, it still provides an interesting point of comparison in relation to previous similar studies. In contrast to the model with the highest MBA, the model with the highest CCR, was trained using  $x_{\text{scale}} + y_{\text{scale}}$  in combination with multinomial logistic regression (Additional file 4: Table S2). The model had a low MBA of 68%, but a high CCR of 90%, comparable to the results of previous studies that used low-resolution acceleration data to generate multiclass behavioral models for cervids [2, 9–11].

### Conclusion

In conclusion, this study found that while it is possible to train classification models based on the behavior of wild red deer, one is faced with a relatively small dataset, especially for rare or short-lived behaviors, such as standing, walking, running, and fighting. We suggest the use of mixed intervals to deal with this difficulty and argue that mixed intervals provide a more realistic depiction of a model's accuracy. Finally, we recommended the use of alternative metrics in addition to the CCR when evaluating the accuracy of behavioral classification models. While we decided to use MBA, there are other metrics that could be used in this scenario [57].

The behavioral classification models for wild red deer living in an alpine environment, generated as part of this study, have various potential applications. For example, such a model could be used to generate activity budgets for unobserved but collared wild red deer and analyze how human activity, seasonality, weather, or climatic changes affect their behavior. Specifically, we could evaluate how a red deer's daily activity budget changes during the hunting season or during unusually warm or cold periods. In a future project, it would be interesting to generate a web-based user interface to allow people to easily generate behavior sequences based on acceleration data, without expertise in R. For

now, anyone familiar with programming in R and working with acceleration data, can use the attached R-script and model to turn their acceleration data into a timed behavioral sequence, provided the acceleration data ranges between 0–255, is averaged over 5-min intervals and includes x- and y-acceleration values (Additional file 5: Script S1). Finally, we hope that our comparative analysis of using different ML algorithms and input variables to generate classification models, our approach to labeling mixed intervals, and the suggested usage of the MBA as an alternative to the CCR can prove useful for future studies working with wild cervids and acceleration data.

#### Abbreviations

ML	Machine learning
SNP	Swiss National Park
IUCN	International Union for Conservation of Nature
KNN	K-nearest neighbor
SVM	Support vector machines
LDA	Linear discriminant analysis
FDA	Flexible discriminant analysis
ANN	Artificial neural network
CART	Classification and regression tree
BRT	Boosted regression trees
RF	Random forest

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40317-025-00401-9>.

**Additional file 1: Fig. S1.** Simplified visualization of the data analysis process to train and test behavioral classification models based on data collected by accelerometers included in GPS collars.

**Additional File 2: Table S2.** Description of the normalization methods applied to the acceleration data. The methods were applied separately to each acceleration axes and individual.

**Additional File 3: Fig. S2.** Visualization of the effects of the various normalization methods applied to the acceleration data. Each point signifies the x- and y-acceleration value of a single 5-min acceleration interval. The plots display the untransformed (a), minmax- (b), scale- (c) and log-normalized (d) acceleration data.

**Additional File 4: Table S2.** The results and properties of all behavioral classification models, including the macro balanced accuracy (MBA), the correct classification rate (CCR), and the balanced accuracies (BA) for the respective behaviors.

**Additional File 5: Script S1.** ZIP-folder containing everything necessary to run the model with the highest macro balanced accuracy that were generated as part of this study. We included an R-script, the model file, and an example dataset (acceleration values and simultaneous behavior of a red deer individual) to run this model. The most accessible approach would be to unzip the folder and open the Behavioral\_classification.Rproj file in RStudio directly.

#### Acknowledgements

The authors would like to thank everyone at the Swiss National Park, Zurich University of Applied Sciences and Inland Norway University of Applied Sciences who made this study possible. We would also like to thank the game keepers of the Canton of Grisons, rangers of the Swiss National Park and veterinarians who were responsible for the capture and handling of the red deer. We thank Bill Fulton (Behayve App) for all his technical support and Adam J. Gaylord for his insights in the initial stages of this study. We thank the Swiss

National Park, Zurich University of Applied Sciences, and the Swiss Academy of Sciences for their materials and financial support.

#### Author contributions

All co-authors conceived the study. BB designed the study with valuable inputs from all co-authors. BB planned and conducted the fieldwork and collected the data with support from TR and CS. BB performed the analyses with support from all co-authors, primarily PA. BB wrote the manuscript and visualized the results with valuable inputs from all co-authors.

#### Funding

Open access funding provided by ZHAW Zurich University of Applied Sciences. This paper is based on the Master's thesis written by the corresponding author and was funded by the Institute of Natural Resource Sciences (Zurich University of Applied Sciences), the Department of Forestry and Wildlife Management (Inland Norway University of Applied Sciences), and the Swiss National Park.

#### Availability of data and materials

The most accurate classification model trained as part of this study is provided in the additional files. Additionally, we provide an R-Script and a sample dataset to demonstrate how one can use this model with their own acceleration data. Behavioral and acceleration data, as well as additional R-Scripts, are available upon reasonable request.

#### Declarations

##### Ethics approval and consent to participate

Capturing and collaring was conducted according to Swiss animal welfare law (permit GR2015-09).

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare no competing interests.

##### Author details

<sup>1</sup>Wildlife Management Unit, Institute of Natural Resource Sciences, Zurich University of Applied Sciences, PO Box, CH-8820 Wädenswil, Switzerland. <sup>2</sup>Swiss National Park, Zerne, Switzerland. <sup>3</sup>Department of Forestry and Wildlife Management, Faculty of Applied Ecology, Inland Norway University of Applied Sciences, Campus Evenstad, Koppang, Norway.

Received: 24 September 2024 Accepted: 18 January 2025

Published online: 08 March 2025

#### References

- Sutherland WJ. The importance of behavioural studies in conservation biology. *Anim Behav.* 1998;56:801–9.
- Loettker P, Rummel A, Traube M, Stache A, Sustr P, Mueller J, et al. New possibilities of observing animal behaviour from a distance using activity sensors in GPS-collars: an attempt to calibrate remotely collected activity data with direct behavioural observations in red deer (*Cervus elaphus*). *Wildl Biol.* 2009;15:425–34.
- Coppes J, Burghardt F, Hagen R, Suchant R, Braunisch V. Human recreation affects spatio-temporal habitat use patterns in red deer (*Cervus elaphus*). *PLoS ONE.* 2017. <https://doi.org/10.1371/journal.pone.0175134>.
- Georgii B, Schröder W. Home range and activity patterns of male red deer (*Cervus elaphus* L.) in the alps. *Oecologia.* 1983;58:238–48.
- Sigrist B, Signer C, Wellig SD, Ozgul A, Filli F, Jenny H, et al. Green-up selection by red deer in heterogeneous, human-dominated landscapes of Central Europe. *Ecol Evol.* 2022. <https://doi.org/10.1002/ece3.9048>.
- Laube P. Representation: Trajectories. In: Richardson D, Castree N, Goodchild MF, Kobayashi A, Liu W, Marston RA, editors. *International Encyclopedia of Geography: People, the Earth, Environment and Technology.* Oxford, UK: John Wiley & Sons, Ltd; 2017. p. 1–11.

7. Kröschel M, Reineking B, Werwie F, Wildi F, Storch I. Remote monitoring of vigilance behavior in large herbivores using acceleration data. *Anim Biotelemetry*. 2017;5:10.
8. Kirchner TM, Devineau O, Chimienti M, Thompson DP, Crouse J, Evans AL, et al. Predicting moose behaviors from tri-axial accelerometer data using a supervised classification algorithm. *Anim Biotelemetry*. 2023;11:32.
9. Gaylord AJ, Sanchez DM. Ungulate activity classification: calibrating activity monitor GPS collars for Rocky Mountain elk, mule deer, and cattle. Master's Thesis. Oregon State University; 2013.
10. Heurich M, Traube M, Stache A, Loettker P. Calibration of remotely collected acceleration data with behavioral observations of roe deer (*Capreolus capreolus* L.). *Acta Theriol (Warsz)*. 2016;57:251–5.
11. Naylor L, Kie J. Monitoring activity of Rocky Mountain elk using recording accelerometers. *Wildl Soc Bull*. 2004;32:1108–13.
12. Nuijten RJM, Gerrits T, Shamoun-Baranes J, Nolet BA. Less is more: on-board lossy compression of accelerometer data increases biologging capacity. *J Anim Ecol*. 2020;89:237–47.
13. Brivio F, Bertolucci C, Tettamanti F, Filli F, Apollonio M, Grignolio S. The weather dictates the rhythms: Alpine chamois activity is well adapted to ecological conditions. *Behav Ecol Sociobiol*. 2016;70:1291–304.
14. Stache A, Heller E, Hothorn T, Heurich M. Activity patterns of European roe deer (*Capreolus capreolus*) are strongly influenced by individual behaviour. *Folia Zool*. 2013;62:67–75.
15. Wang Y, Nickel B, Rutishauser M, Bryce CM, Williams TM, Elkaim G, et al. Movement, resting, and attack behaviors of wild pumas are revealed by tri-axial accelerometer measurements. *Mov Ecol*. 2015. <https://doi.org/10.1186/s40462-015-0030-0>.
16. Signer C, Ruf T, Schober F, Fluch G, Paumann T, Arnold W. A versatile telemetry system for continuous measurement of heart rate, body temperature and locomotor activity in free-ranging ruminants. *Methods Ecol Evol*. 2010;1:75–85.
17. Pagano AM, Rode KD, Cutting A, Owen MA, Jensen S, Ware JV, et al. Using tri-axial accelerometers to identify wild polar bear behaviors. *Endanger Species Res*. 2017;32:19–33.
18. Roberts CP, Cain JW III, Cox RD. Application of activity sensors for estimating behavioral patterns. *Wildl Soc Bull*. 2016;40:764–71.
19. Tuia D, Kellenberger B, Beery S, Costelloe BR, Zuffi S, Risse B, et al. Perspectives in machine learning for wildlife conservation. *Nat Commun*. 2022;13:792.
20. Eikelboom JAJ, Wind J, Van De Ven E, Kenana LM, Schroder B, De Kneegt HJ, et al. Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods Ecol Evol*. 2019;10:1875–87.
21. Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, et al. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc Natl Acad Sci*. 2018;115(25):25.
22. Schneider S, Taylor GW, Linquist S, Kremer SC. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods Ecol Evol*. 2019;10:461–70.
23. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. xgboost: extreme gradient boosting. R package version 1.7.3.1. 2023. <https://CRAN.R-project.org/package=xgboost>.
24. Fritsch S, Guenther F, Wright MN. neuralnet: training of neural networks. R package version 1.44.2. 2019. <https://CRAN.R-project.org/package=neuralnet>.
25. Hastie T, Tibshirani R, Leisch F, Hornik K, Ripley BD, Narasimhan B. mda: mixture and flexible discriminant analysis. R package version 0.5–3. 2022. <https://CRAN.R-project.org/package=mda>.
26. Karatzoglou A, Smola A, Hornik K. kernlab: kernel-based machine learning lab. R package version 0.9–32. 2023. <https://CRAN.R-project.org/package=kernlab>.
27. Kuhn M. caret: classification and regression training. R package version 6.0–93. 2022. <https://CRAN.R-project.org/package=caret>.
28. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2:18–22.
29. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7–12. 2022. <https://CRAN.R-project.org/package=e1071>.
30. Ripley B. Tree: classification and regression trees. R package version 1.0–42. 2022. <https://CRAN.R-project.org/package=tree>.
31. Venables WN, Ripley BD. Modern applied statistics with s. New York: Springer; 2002.
32. Nathan R, Spiegel O, Fortmann-Roe S, Harel R, Wikelski M, Getz WM. Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures. *J Exp Biol*. 2012;215:986–96.
33. Ladds MA, Thompson AP, Slip DJ, Hocking DP, Harcourt RG. Seeing it all: evaluating supervised machine learning methods for the classification of diverse otariid behaviours. *PLoS ONE*. 2016;11: e0166898.
34. Haller H, Eisenhut A, Haller R. Atlas des Schweizerischen Nationalparks: Die ersten 100 Jahre. 1st ed. Bern: Haupt; 2013.
35. MeteoSchweiz. Klimanormwerte Buffalora: normperiode 1991–2020. Bundesamt für Meteorologie und Klimatologie MeteoSchweiz; 2021.
36. VECTRONIC Aerospace GmbH. GPS Plus Collar. 2015.
37. VECTRONIC Aerospace GmbH. VERTEX Plus Collar. 2016.
38. Altmann J. Observational study of behavior: sampling methods. *Behaviour*. 1974;49:227–67.
39. Fulton B. Behayve. Android app version 0.6. 2022. <https://behayve.com>.
40. Di Bian. Timestamp camera pro. Android app. 2023. [https://play.google.com/store/apps/details?id=com.jeyluta.timestampcamera&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.jeyluta.timestampcamera&hl=en_US&gl=US)
41. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York NY: Springer, US; 2021.
42. R Core Team. R: a language and environment for statistical computing. Programming language version 4.2.3. 2023. <https://R-project.org/>.
43. RStudio Team. RStudio: integrated development environment for R. Software version 2023.03.0. 2023. <https://rstudio.com/>.
44. Dickinson ER, Stephens PA, Marks NJ, Wilson RP, Scantlebury DM. Best practice for collar deployment of tri-axial accelerometers on a terrestrial quadruped to provide accurate measurement of body acceleration. *Anim Biotelemetry*. 2020;8:9.
45. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*. 2016;17:29–39.
46. Resheff YS, Bensch HM, Zöttl M, Harel R, Matsumoto-Oda A, Crofoot MC, et al. How to treat mixed behavior segments in supervised machine learning of behavioural modes from inertial measurement data. *Mov Ecol*. 2024;12:44.
47. Raschka S. 2018 STAT 479: machine learning lecture notes. University of Wisconsin–Madison: UK.
48. UCLA: Statistical Consulting Group. Multinomial logistic regression | R data analysis examples. 2021. <https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/>. Accessed 20 Mar 2023.
49. Brownlee J. Support vector machines for machine learning. Machine Learning Mastery. 2016. <https://machinelearningmastery.com/support-vector-machines-for-machine-learning/>. Accessed 7 Mar 2023.
50. Starmer J. Linear discriminant analysis. In: The StatQuest illustrated guide to machine learning. 2020.
51. Amazon Web Services, Inc. What is a Neural Network? - Artificial Neural Network Explained - AWS. 2023. <https://aws.amazon.com/what-is/neural-network/>. Accessed 10 Mar 2023.
52. Awan AA, Navlani A. Naive bayes classifier tutorial: with python scikit-learn. DataCamp. 2023. <https://datacamp.com/tutorial/naive-bayes-scikit-learn>. Accessed 26 Apr 2023.
53. Starmer J. Naive bayes. In: The StatQuest illustrated guide to machine learning. 2020.
54. Sağlam C, Çetin N. Machine learning algorithms to estimate drying characteristics of apples slices dried with different methods. *J Food Process Preserv*. 2022;46(10): e16496.
55. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab - An S4 Package for Kernel Methods in R. *J Stat Softw*. 2004;11:1–20.
56. Wilber J, Santamaría L. Decision trees. MLU-EXPLAIN. 2023. <https://mlu-explain.github.io/decision-tree/>. Accessed 10 Mar 2023.
57. Tharwat A. Classification assessment methods. *Appl Comput Inform*. 2021;17:168–92.
58. Akosa J. Predictive accuracy: A misleading performance measure for highly imbalanced data. 2017.
59. Beier P, McCullough DR. Motion-sensitive radio collars for estimating white-tailed deer activity. *J Wildl Manag*. 1988;52:11–3.

60. Brivio F, Bertolucci C, Marcon A, Cotza A, Apollonio M, Grignolio S. Dealing with intra-individual variability in the analysis of activity patterns from accelerometer data. *Hystrix-Ital J Mammal*. 2021;32:41–7.
61. Gottardi E, Tua F, Cargnelutti B, Maublanc M-L, Angibault J-M, Said S, et al. Use of GPS activity sensors to measure active and inactive behaviours of European roe deer (*Capreolus capreolus*). *Mammalia*. 2010;74:355–62.
62. Krop-Benesch A, Berger A, Hofer H, Heurich M. Long-term measurement of roe deer (*Capreolus capreolus*) activity using two-axis accelerometers in GPS-collars. *Ital J Zool*. 2013;80:69–81.
63. Moen R, Pastor J, Cohen Y. Interpreting behavior from activity counters in GPS collars on moose. *Alces*. 1996;32:101–8.
64. Campbell H, Gao L, Bidder O, Hunter J, Franklin C. Creating a behavioural classification module for acceleration data: using a captive surrogate for difficult to observe species. *J Exp Biol*. 2013;216(4):4501.
65. Blankenhorn HJ, Buchli C, Voser P. Wanderungen und jahreszeitliches Verteilungsmuster der Rothirschpopulationen (*Cervus elaphus* L.) im Engadin, Münstertal und Schweizerischen Nationalpark. *Rev Suisse Zool*. 1978;85:779–89.
66. Mitchell B, McCowan D, Nicholson IA. Annual cycles of body weight and condition in Scottish Red deer, *Cervus elaphus*. *J Zool*. 1976;180:107–27.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.