Manuscript ***https://doi.org/10.5281/zenodo.12627197***

Response to Reviewers

Dear Dr. Kristan Reed,

Thank you for giving us the opportunity to submit a revised version of our manuscript titled " *A pipeline with pre-processing options to detect behaviour from accelerometer data using Machine Learning tested on dairy goats."* We sincerely appreciate the time and effort that you and the reviewers have put into providing detailed feedback, and we are grateful for the thoughtful comments and valuable suggestions that have contributed to improving this paper.

We have addressed all the reviewers' suggestions, and the corresponding changes are highlighted in the tracked changes document. Below, we provide a point-by-point response to each of the reviewers' comments and concerns, with our replies shown in red. The corresponding changes are also reported in green in each response point. All page references correspond to the revised manuscript.

Thank you again for your consideration and the opportunity to improve our work.

Kind regards,

Sarah Mauny

Reviewers' Comments to the Authors:

Reviewer 1

1. *Although I believe the title is acceptable as is, maybe it would match the objectives and results obtained by showcasing the pipeline a bit better. Something like "A pipeline for preprocessing and feature extraction of accelerometer data aimed at dairy goat behavior classification using machine learning" would maybe showcase that bit better.*

Author response: We appreciate the suggestion to highlight the pipeline in the title. We believe that this paper is primarily focused on applying the pipeline to a specific dataset involving dairy goats. The pipeline was designed to be generic and applicable to data from various animal species, so it is not aimed on goats, but in this work, we have demonstrated its use on dairy goat data.

Additionally, the data paper we submitted alongside this article specifically presents the development and functionality of the pipeline in greater detail.

That said, we have proposed a new title that better reflects the application of the pipeline on dairy goats in this study.

Corresponding change: "A pipeline with pre-processing options to detect behaviour from accelerometer data using Machine Learning tested on dairy goats."

Reviewer 2

1. ***Line 115****. Could you please list which performance score was considered in this study?*

Author response: We have added the performance scores used in this study in lines [115-116]. These are also mentioned in the Materials & Methods section, under [2.5. Behaviour Classification Model], on lines 293-295.

Corresponding changes (lines 115-116): "In this work, the main performance score was the AUC (Area Under the Curve) score. Accuracy, balanced accuracy, F1-score, sensitivity and specificity were also calculated."

2. ***Lines 116-119.*** *Did you mean that the model was not trained to detect behaviours in goats but can directly predict goat behaviours on the testing data? Or do you mean the model was not only trained but also tested?*

Author response: To clarify, we meant that first, the models were trained with the training dataset, which was composed of randomly chosen 80% of the data from eight goats, and tested on the other 20% of the data of the same eight goats.

Then, we also trained the models with all the data of only six of the goats and tested the models on the data of the two remaining goats. This methodology assesses the ability of the models to generalize across different animals.

We have revised the text accordingly to reflect this in lines [116-119]. This methodology is also described in the Materials & Methods section, under [2.5. Behaviour classification model], on lines 272-287.

Corresponding changes (lines 116-119): "Moreover, in this study, detection of behaviours in goats that were not part of model's training dataset was also tested to reflect the ability of the model to generalise detection of behaviours on data from new goats. The model was trained with a dataset of six goats and its performance was tested on the two remaining goats."

3. ***Lines 157-159.*** *Instead of relatively large amounts of data that can be collected for those four behaviours, could you please justify why those behaviours were selected from other perspectives, such as indicating the potential occurrence of a disease? Data availability is an important factor, but could those behaviours be specifically connected with any realistic applications?*

Author response: Thank you for your comment. Indeed, the chosen behaviours were selected not only due to their large representation in the dataset, but also because of their relevance to animal welfare and health, as they reflect the animals' general activity levels and feeding behaviour.

The explanation has been added in lines [157-162].

Corresponding changes (lines 157-162): "The behaviours « ruminating », « lying », « standing » and « head in the feeder » were selected to develop the classification model because of their large representation in the observation period to maximise available data for model development, and for their relevance regarding welfare and health status of the animals. In this context, it has been shown that general activity levels and time spent feeding are reduced under bad health conditions such as lameness (Thorup et al., 2016), while time spent lying can increase."

4. ***Lines 179-180.*** *Could you please also add more descriptions about how those additional time series features were added for the modelling analysis as shown in Figure 4?*

Author response: The additional time-series are derived from the raw acceleration data using the formulas provided in Table 1.

Concretely, each additional time-series is computed for every row of acceleration data, creating new columns in the dataset, as illustrated in Figure 4. This clarification has been added in lines [207-209].

Corresponding changes (lines 207-209): "These additional time-series were calculated from the formulas indicated in **Table 1**. Pitch and roll angles, along with a transformation of the acceleration data, were calculated from each acceleration value and added as new variables to the dataset."

5. ***Lines 224-225****. What is the time length or size of each time window applied in this study? I saw time lengths for different time windows in the result section, but I think it would be more proper to clarify that before the result section as it was already mentioned several times. Please provide more detailed descriptions about the "time-windows".*

Author response: There are two points to clarify:

- First, one model was trained per behaviour. This clarification has been added at the beginning of section [2.4. Data Pre-processing], lines [173-175]. When sensitivity analysis was performed, it was done for each behaviour, resulting in different preferred pre-processing treatments for each behaviour, including different preferred time-window sizes; one per behaviour.

Corresponding changes (lines 173-175): "In this section, the different steps of the pipeline that prepare data for use in the ML algorithm are detailed. The data was pre-processed separately for each behaviour, resulting in four binary classification models, one for each behaviour."

- Second, when the sensitivity analysis evaluated the optimal time-window size, several time-window sizes were tested, and the model's performance was assessed for each size, ranging from 10 seconds to 120 seconds. This has been clarified in lines [230-232].

Corresponding changes (lines 230-232): "In this study, in order to find which size of time-window gives the best performance score for each behaviour, several sizes of time-windows were tested for each behaviour, ranging from 10 seconds to 120 seconds."

In conclusion, several time-window sizes were tested for each behaviour, resulting in one preferred time-window size per behaviour.

6. ***Lines 236-240****. If I understand correctly, I think both previous studies and this manuscript used predefined set of input features but the total number of features in this study might be more than others. Did this study include any new features that were never considered by previous studies? Or did you find that most of the previous studies did not apply feature selection via feature importance which thereby is one of the novelties of this study?*

Author response: Yes, both previous studies and this manuscript used a predefined set of features. However, the key difference is that previous studies selected data features considered as relevant in advance, whereas in this study, we did not make such assumptions. We calculated thousands of features for each time series, including some that had never been considered before, to explore potentially relevant features. This clarification has been added to lines [236-240].

Corresponding changes (lines 236-240): "In the field of ruminant behaviour prediction, previous studies have relied on a predefined set of features in both the time and frequency domains. In this study, an extensive feature extraction on each time-serie of the dataset was chosen to explore features that were never considered in previous studies. This was followed by a feature selection based on the importance of each feature during the model development rather than limiting the number of features before the prediction."

7. ***Line 241****. What does "(777)" mean and what's the purpose of showing this number here?*

Author response: This number provides an indication of the number of features calculated.

The Python package tsfresh automatically calculates 777 features for each time window and for each time-series. For example, if the dataset contains only acceleration values on three axes (x, y, and z), a total of 777 × 3 features is generated. For any given feature, for instance the "mean value", it is automatically calculated for the x-axis, y-axis, and the z-axis acceleration values on each time-window. The sentence in lines [241-242] has been reformulated.

Corresponding changes (lines 241-242): "A wide range of 777 different features was automatically calculated (Python (version 3.10) package tsfresh (Christ et al., 2018)) on each time-window and on each extracted time-series (**Figure 5**)."

Details have also been added in lines [302-307], in the [2.6. Features selection] selection (explained in the response to comment n°9 below).

Corresponding changes (lines 302-307): "Without features selection, the models were trained with 777 features multiplied by n (the number of time-series in the dataset). Note that the high dimensionality of the data was not addressed in this work, as Gradient Boosting can handle high-dimensional datasets and train models with thousands of features. But by selecting the most important features, it is possible to improve the performance of the models by focusing on the most relevant ones, and reduce the preprocessing workload, resulting in smaller data volumes and faster processing times.

8. ***Lines 241-248****. Is the purpose of this paragraph to introduce this "tsfresh" package or did this study calculate any of these feature characteristics using this package? I suggest focusing on things that are highly correlated with this study.*

Author response: The purpose of this paragraph is to present the features that we calculated, using the tsfresh package, as it is the tool we used to calculate all the features from our data. All feature characteristics in this study were derived using this package.

We have reformulated the paragraph to clarify this point in lines [241-248].

Corresponding changes (lines 241-248): "A wide range of 777 different features was automatically calculated (Python (version 3.10) package tsfresh (Christ et al., 2018)) on each time-window and on each extracted time-series (**Figure 5**). The calculated features are descriptive statistics features […], temporal characteristics […], value distribution features[…], spectral domain features[…], […]."

9. ***Line 299-301****. Does that mean there are at most 2000 input features for modeling analysis? Did you run into issues caused by high dimensionality and how did you solve it?*

Author response: Yes, there were even more for the models trained without features selection. As mentioned in comment n°7, 777 features were calculated per time series and used to train the models. Gradient Boosting effectively handles thousands of features, avoiding issues related to high dimensionality. This explanation has been added to lines [302-307].

Corresponding changes (lines 302-307): "Without features selection, the models were trained with 777 features multiplied by n (the number of time-series in the dataset). Note that the high dimensionality of the data was not addressed in this work, as Gradient Boosting can handle high-dimensional datasets and train models with thousands of features. But by selecting the most important features, it is possible to improve the performance score of the models by focusing on the most relevant ones, and reduce the preprocessing workload, resulting in smaller data volumes and faster processing times."

Additionally, we applied features selection to try improving the model's prediction performance by focusing on the most important features. This selection process is based on the weight of each feature, which indicates the impact of that feature on the prediction. Details on the selection process has been added lines [295-299].

Corresponding changes (lines 295-299): "The features_importances_ function from the scikit-learn package (Pedregosa et al., 2011) was used to evaluate the importance of each feature. It looks at all the decisions made by each tree in the model. It then figures out which features were most helpful on average for making those decisions and calculates a weight, which indicates the impact of that feature on the prediction. The features are then ranked by their weight, allowing us to keep only a limited number of the features with the highest importance."

10. ***Line 335****. In Figure 7 box 3, it seems you did hyperparameter tuning after the testing set gets involved into this modeling analysis but in Line 267 this study mentions that "A thorough hyperparameter tuning is performed using the validation set". Could you please make it clearer about what you did regarding the hyperparameter tuning?*

Author response: Hyperparameter tuning was done using the validation set, referred to as the "validation set" in **Figure 6** and abbreviated as "val" in **Figure 7**. We have added this clarification in lines [264-265] and line [268].

Corresponding changes (lines 264-265): "A thorough hyperparameter tuning was performed using the validation set ("validation set" in **Figure 6**; "val" set in **Figure 7**) and based on the highest AUC Area Under the Curve) score obtained."

Line 268: "Once the model was trained and tuned, the test set ("test set" in **Figure 6;** "X_test", "y_test" in **Figure 7**) was used to evaluate the final performance of the model […]"

In **Figure 7**, box 3, the dashed arrows indicate which set is involved in each step. The first performance evaluation, as well as the hyperparameter tuning, was conducted on the "val" set. The test set is only used for the final performance evaluation, as indicated by the second dashed arrow.