Thank the authors for submitting the manuscript titled “Dairy goat behaviour classification from accelerometer data using a Machine Learning pipeline with extensive pre-processing and feature creation options”. This paper aims to utilize a machine learning pipeline with a supervised classification algorithm to predict goat behaviours. Although there are some novelties, I still have some comments for the authors to improve this manuscript.

1. Line 113: Could you please list which performance score was considered in this study?

2. Lines 115-116: 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?

3. Lines 154-155: 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?

4. Lines 201-204: Could you please also add more descriptions about how those additional time series features were added for the modeling analysis as shown in Figure 4?

5. Line 217: 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”.

6. Lines 227-230: 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?

7. Line 232: what does “(777)” mean and what’s the purpose of showing this number here?

8. Lines 239-244: 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.

9. Line 291: 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?

10. In Figure 7 box 3, it seems you did hyperparameter tuning after the testing set gets involved into this modeling analysis but in Line 256 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?